Low-rank Interaction Contingency Tables

Geneviève Robin, Julie Josse, Éric Moulines, Sylvain Sardy

École Polytechnique

genevieve.robin@polytechnique.edu

July 11, 2017

High dimensional count data

- *Ecological data* (abundance of species across environments)

	Alop.alpi	Alch.pent	Geum.mont	Pote.aure	Sali.herb
AR26	0	0	2	2	0
AR08	1	0	2	1	0
AR05	0	0	3	3	0
AR06	0	0	3	0	0
AR69	1	0	2	2	2
AR32	2	0	3	3	1
AR40	2	3	3	4	0

Table: Excerpt of Aravo dataset. 82 species of plants across 75 environments in the French Alps (Dray and Dufour, 2007).

How do species interact with environments ?

- $Y \in \mathbb{N}^{m_1 \times m_2}$, Y_{ij} independent Poisson; estimate $X_{ij} = \log\left(\mathbb{E}[Y_{ij}]\right)$;

$\Rightarrow$ Log-linear model:

$$X_{ij} = \alpha_i + \beta_j + \Theta_{ij},\ \mathrm{rk}(\Theta) = K$$

Log-linear model with known covariates

Environment characteristics, species traits are known.

	Aspect	Slope	Form	PhysD	ZoogD	Snow
AR26	5	0	3	20	no	140
AR08	8	20	3	60	some	160
AR05	9	10	4	20	high	150
AR06	8	20	3	40	high	160
AR69	8	30	2	30	high	160
AR32	8	10	5	20	some	160
AR40	8	15	4	10	some	180

	Height	Spread	Angle	Area	Thick	SLA	N_mass	Seed
Alop.alpi	5.00	20	20	190.90	0.20	15.10	203.85	0.21
Poa.alpi	8.00	15	45	160.00	0.18	10.70	204.37	0.32
Alch.pent	2.00	20	15	218.10	0.16	23.70	364.98	0.31
Geum.mont	5.00	10	15	852.60	0.20	11.30	223.74	1.67
Plan.alpi	0.50	10	20	40.00	0.22	11.90	242.76	0.33
Pote.aure	3.00	20	15	264.50	0.10	17.50	253.75	0.24
Sali.herb	1.00	50	60	82.50	0.18	14.70	367.50	0.05

Figure: Environment (left) an species (right) covariates for Aravo data (excerpt)

$$X_{ij} = (R\alpha)_{ij} + (\beta C)_{ij} + \Theta_{ij}$$

- $X \in \mathbb{R}^{m_1 \times m_2}$. Column covariates $C \in \mathbb{R}^{K_2 \times m_2}$, row covariates $R \in \mathbb{R}^{m_1 \times K_1}$, $\alpha \in \mathbb{R}^{K_1 \times m_2}$, $\beta \in \mathbb{R}^{m_1 \times K_2}$, Θ_{ij}
- α_{ij} effect of i-th row covariate on j-th species
- β_{ij} effect of j-th column covariate on i-th environment

Low-rank interaction log-linear model

Penalized negative Poisson log-likelihood for $\lambda > 0$ (relaxation of the rank constraint)

$$\Phi_Y^\lambda(X,\Theta) = -(m_1 m_2)^{-1}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\left(Y_{ij}X_{ij} - \exp(X_{ij})\right) + \lambda \left\|\Theta\right\|_*$$

$$\begin{aligned} \hat{X}^\lambda, \hat{\Theta}^\lambda & = \operatorname*{argmin}_{X\in\mathcal{K}} \quad \Phi_Y^\lambda(X) \\ \text{s.t.} \quad & \mathcal{T}(X) = \Theta \end{aligned},$$

$\mathcal{T} : X \in \mathbb{R} \mapsto \Pi_R^\perp X \Pi_C^\perp$ projection on matrix subspace orthogonal to R and C.

Parameter λ tuned with cross-validation or Quantile Universal Threshold (QUT) Diaz Rodriguez and Sardy (2014).

Alternating direction method of multipliers (ADMM) Boyd et al. (2011)

Augmented Lagrangian indexed by τ, Γ dual variable:

$$\mathcal{L}_\tau(X, \Theta, \Gamma) = \Phi_Y(X) + \lambda \|\Theta\|_{\sigma,1} + \langle \Gamma, \mathcal{T}(X) - \Theta \rangle + \frac{\tau}{2} \|\mathcal{T}(X) - \Theta\|_2^2 .$$

At iteration $k+1$ ADMM update rules are given by

$$\begin{aligned}
X^{k+1} &= \mathrm{argmin}_{X \in \mathcal{K}} \quad \mathcal{L}_\tau\left(X, \Theta^k, \Gamma^k\right) \\
\Theta^{k+1} &= \mathrm{argmin}_{\Theta \in \mathcal{K}_\mathcal{T}} \quad \mathcal{L}_\tau\left(X^{k+1}, \Theta, \Gamma^k\right) \\
\Gamma^{k+1} &= \Gamma^k + \tau\left(\mathcal{T}(X^{k+1}) - \Theta^{k+1}\right).
\end{aligned}$$

Statistical guarantees

Theorem (Risk bound)

Assume

- Y_{ij} *have bounded means and variance;*
- Y_{ij} *are subexponential variables;*
- $m_1 + m_2 \geq C_1$, $\lambda = C_2\sqrt{2(m_1 \vee m_2)\log(m_1 + m_2)}/(m_1 m_2)$.

Then

$$\frac{\left\| X - \hat{X}_\lambda \right\|_{\sigma,2}^2}{m_1 m_2} \lesssim \frac{(m_1 + m_2)\,(\mathrm{rk}(\Theta) + K_1 + K_2)}{m_1 m_2},$$

with probability at least $1 - (m_1 + m_2)^{-1}$.

C_1, C_2 are constants and $\lesssim$ denotes inequality up to constants and log factors.

Aravo data

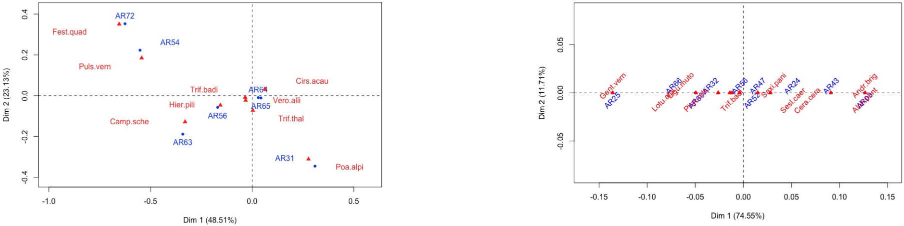

Figure: Visualization of the 10 largest interactions between environments (blue) and species (red) in the two first dimensions of interaction with GAMMIT without covariates (left) and with explanatory covariates (right).

Conclusion

- Use GAMMIT to impute contingency tables
- Extend the model to analyze mixed data
- Application to analysis of healthcare data

- *Low-rank Interaction Contingency Tables* on `https://arxiv.org` for more details
- Implementation of the method available at `https://github.com/genevievelrobin/GAMMIT`

References

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning 3*(1), 1–22.

Diaz Rodriguez, J. and S. Sardy (2014). Quantile universal threshold: model selection at the detection edge for high-dimensional regression. *ArXiv e-prints*.

Dray, S. and A. Dufour (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software 22*(4), 1–20.