Low-rank Interaction Contingency Tables

Julie Josse

École Polytechnique, INRIA

Joint work with: Geneviève Robin, Éric Moulines & Sylvain Sardy

March 17, 2017

Research activities

- Dimensionality reduction methods to visualize complex data (PCA based): multi-sources data, textual data, arrays
- Missing values - matrix completion
- Low rank estimation, selection of regularization parameters
- Fields of application: bio-sciences (agronomy, sensory analysis), health data (hospital data)
- R community: book R for Statistics, R foundation, R taskforce, R packages and JSS papers:

  FactoMineR explore continuous, categorical, multiple contingency tables (correspondence analysis), combine clustering and PC, ..
  MissMDA for single and multiple imputation, PCA with missing
  denoiseR to denoise data

Overview

1. Motivations
2. Generalized additive main effects & multiplicative interaction thresholded (GAMMIT)
   - model
   - optimization algorithm
3. Automatic selection of the regularization parameter
   - cross validation
   - quantile universal threshold
4. Experiments
5. Data analyses

Motivations

High dimensional count data

- Single-cell RNA sequencing (counts of genes in cells)
- Image processing (number of photons on a grid)
- *Ecological data* (abundance of 82 species across 75 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
...	...	...	...	...	...

Table: Aravo data. Plants in France Alpes. (Dray and Dufour, 2007).

$\Rightarrow$ How do species interact with environments?

$\Rightarrow$ **Denoise and visualize data**

Log-linear model

Observation matrix $Y \in \mathbb{N}^{m_1 \times m_2}$, Y_{ij} counts occurrences of (i, j)
Y_{ij} independent, $Y_{ij} \sim \mathcal{P}(\mu_{ij})$). Estimate $\mathbb{E}[Y_{ij}] = \mu_{ij}$; $X_{ij} := \log(\mu_{ij})$

$$X_{ij} = \alpha_i + \beta_j + \Theta_{ij} \text{ (Christensen, 1990; Agresti, 2013)}$$

- α_i effect of i-th environment
- β_j effect of j-th species
- Θ_{ij} interaction between i-th environment and j-th species

Log-linear model

Observation matrix $Y \in \mathbb{N}^{m_1 \times m_2}$, Y_{ij} counts occurrences of (i, j)
Y_{ij} independent, $Y_{ij} \sim \mathcal{P}(\mu_{ij})$). Estimate $\mathbb{E}[Y_{ij}] = \mu_{ij}$; $X_{ij} := \log(\mu_{ij})$

$$X_{ij} = \alpha_i + \beta_j + \Theta_{ij} \text{ (Christensen, 1990; Agresti, 2013)}$$

- α_i effect of i-th environment
- β_j effect of j-th species
- Θ_{ij} interaction between i-th environment and j-th species
- Θ has rank $K < \min(m_1 - 1, m_2 - 1)$

$$X_{ij} = \alpha_i + \beta_j + (UDV^\top)_{ij},$$

$UDV^\top$, the truncated SVD of Θ at K.

(RC model, Goodman, 1985; log-bilinear model, Falguerolle, 1998; GAMMI, Gower, 2011)

$\Rightarrow$ requires K; overfitting issues

Generalized additive main effects and multiplicative interaction thresholded (GAMMIT)

- Adding covariates
- Improving on MLE by regularization

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)

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

- $C \in \mathbb{C}^{K_2 \times m_2}$ matrix of column covariates, $R \in \mathbb{R}^{m_1 \times K_1}$ matrix of row covariates, $\alpha \in \mathbb{R}^{K_1}$, $\beta \in \mathbb{R}^{K_2}$, Θ_{ij} interaction matrix
- α_i effect of i-th row covariate
- β_j effect of j-th column covariate

$\Rightarrow$ Estimate the interaction Θ not explained by covariates

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)

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

- $C \in \mathbb{R}^{K_2 \times m_2}$ matrix of column covariates, $R \in \mathbb{R}^{m_1 \times K_1}$ matrix of row covariates, $\alpha \in \mathbb{R}^{K_1 \times m_2}$, $\beta \in \mathbb{R}^{m_1 \times K_2}$, 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

$\Rightarrow$ Estimate the interaction Θ not explained by covariates

Model

We can re-write model $X = R\alpha + \beta C + \Theta$

$$X = X_0 \overset{\perp}{+} \Theta,\ X_0 \in \mathcal{V},\ \Theta \in \mathcal{V}^{\perp},$$

Π_1 orthogonal project on subspace span by columns of C, Π_2 span by R;
$\mathcal{V}$ subspace span by columns of $\tilde{X} = \Pi_1 X + X\Pi_2 - \Pi_1 X \Pi_2$;
$\mathcal{T} : X \mapsto \Theta$ orthogonal projection operator on $\mathcal{V}^{\perp}$;
$\Rightarrow$ **Covariates effects**

$$\begin{aligned} X_0 &= \Pi_1 X + X\Pi_2 - \Pi_1 X \Pi_2 \\ X_0 &= X - \mathcal{T}(X) \end{aligned}$$

$\Rightarrow$ **Remaining interaction**

$$\begin{aligned} \Theta &= (I - \Pi_2) X (I - \Pi_1) \\ \Theta &= \mathcal{T}(X) \end{aligned}$$

$\Rightarrow$ Rq: RC model $\Rightarrow$ double centering (classical identifiability constraints)

Penalized log-bilinear model

$\Rightarrow$ Penalized Poisson log-likelihood for $\lambda > 0$ (convex relaxation of rank)

$$\hat{X}^{\lambda} = \underset{X}{\text{argmin}} \quad \Phi_Y(X) + \lambda \|\mathcal{T}(X)\|_{\sigma,1}$$

$$\Phi_Y(X) = -(m_1 m_2)^{-1} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (Y_{ij} X_{ij} - \exp(X_{ij}))$$

Trade-off between data fitting and low rank interaction.

Penalized log-bilinear model

$\Rightarrow$ Penalized Poisson log-likelihood for $\lambda > 0$ (convex relaxation of rank)

$$\hat{X}^{\lambda} = \underset{X}{\operatorname{argmin}} \quad \Phi_Y(X) + \lambda \|\mathcal{T}(X)\|_{\sigma,1}$$

$$\Phi_Y(X) = -(m_1 m_2)^{-1} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (Y_{ij} X_{ij} - \exp(X_{ij}))$$

Trade-off between data fitting and low rank interaction.

Bounded entries: parameter set $\mathcal{K} = [\underline{\gamma}, \bar{\gamma}]^{m_1 \times m_2}, \underline{\gamma} > 0$, $\bar{\gamma} < \infty$ compact.

$$\underset{X \in \mathcal{K}, \quad \Theta \in \mathcal{K}_{\mathcal{T}}}{\operatorname{argmin}} \quad \Phi_Y(X) + \lambda \|\Theta\|_{\sigma,1} \text{ s.t. } \mathcal{T}(X) = \Theta, \tag{1}$$

where $\mathcal{K}_{\mathcal{T}}$ image of $\mathcal{K}$ by $\mathcal{T}$ is compact.

(1) is a separable, linearly constrained, strongly convex program on a compact set $\Rightarrow$ unique solution

Alternating direction method of multipliers (ADMM)

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

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

Rq: τ has an influence on speed of convergence and not on final result

Update rules

- X update: gradient descent

$$X^{k+1} = \text{argmin}_{X \in \mathcal{K}} \quad \Phi_Y(X) + \lambda \left\| \Theta^k \right\|_{\sigma,1} + \langle \Gamma^k, \mathcal{T}(X) - \Theta^k \rangle + \frac{\tau}{2} \left\| \mathcal{T}(X) - \Theta^k \right\|_2^2$$

$$\nabla_X \mathcal{L}_\tau \left(X, \Theta^k, \Gamma^k \right) = \nabla \Phi_Y(X) + \Gamma^k + \tau \left(\mathcal{T}(X) - \Theta^k \right).$$

Update rules

- X update: gradient descent

$$X^{k+1} = \text{argmin}_{X \in \mathcal{K}} \quad \Phi_Y(X) + \lambda \left\| \Theta^k \right\|_{\sigma,1} + \langle \Gamma^k, \mathcal{T}(X) - \Theta^k \rangle + \frac{\tau}{2} \left\| \mathcal{T}(X) - \Theta^k \right\|_2^2$$

$$\nabla_X \mathcal{L}_\tau \left(X, \Theta^k, \Gamma^k \right) = \nabla \Phi_Y(X) + \Gamma^k + \tau \left(\mathcal{T}(X) - \Theta^k \right).$$

- Θ update: $\text{argmin}_{\mathcal{K}_\mathcal{T}} \quad \lambda \left\| \Theta^k \right\|_{\sigma,1} + \frac{\tau}{2} \left\| \mathcal{T}(X^{k+1}) + \Gamma^k/\tau - \Theta^k \right\|_2^2$
  $\Rightarrow$ closed form (**rank selection**)

$$\Theta^{k+1} = \mathcal{D}_{\lambda/\tau} \left(\mathcal{T}(X^{k+1}) + \Gamma^k/\tau \right),$$

$\mathcal{D}_{\lambda/\tau}$ operator for soft-thresholding of singular values at level λ/τ.

Automatic selection of λ

Cross-validation

- Remove a fraction of the entries of Y
- Compute $\hat{X}^{\lambda}$ for all λ
- Compute $\left\| \exp(\hat{X}^{\lambda}_{\text{mis}}) - Y_{\text{mis}} \right\|_2^2$ for each λ
- Repeat N times

Select $\lambda_{\text{CV}} = \underset{\lambda}{\text{argmin}} \quad 1/N \sum_{i=1}^{N} \left\| \exp(\hat{X}^{\lambda}_{\text{mis}}{}^{(i)}) - Y_{\text{mis}} \right\|_2^2$.

$\Rightarrow$ requires EM algorithm to estimate the parameters $\hat{X}^{\lambda}$ from an incomplete data set

Cross validation

$$\Rightarrow \text{EM algorithm}$$

$Y = (Y_{\text{obs}}, Y_{\text{mis}})$. At iteration k

E step: $\mathbb{E}_{Y_{\text{mis}}}[\Phi^{\lambda}_{(Y_{\text{obs}}, Y_{\text{mis}})}(X)|\, Y_{\text{obs}}; \hat{X}^{\lambda}{}^{k}]$

- $Y^{k+1}_{\text{mis}} = \mathbb{E}[Y_{\text{mis}}|\hat{X}^{\lambda}{}^{k}] = \exp(\hat{X}^{\lambda}_{\text{mis}}{}^{k})$
- $Y^{k+1}_{\text{obs}} = Y_{\text{obs}}$

M step: $\hat{X}^{\lambda}{}^{k+1} = \text{argmax} \quad \mathbb{E}_{Y_{\text{mis}}}[\Phi^{\lambda}_{(Y_{\text{obs}}, Y_{\text{mis}})}(X)|\, Y_{\text{obs}}; \hat{X}^{\lambda}{}^{k}]$

- Run the ADMM algorithm

$\Rightarrow$ Iterative imputation (common in Gaussian case)

$\Rightarrow$ Time consuming

Quantile universal threshold (QUT)

CV designed to minimize prediction error. What about selecting the rank ? (number of non-zero singular values)

$\Rightarrow$ Extend the work of Giacobino et al. (2016) on zero-thresholding function

Theorem (Zero-thresholding function)

The interaction estimator $\mathcal{T}(\hat{X}_\lambda)$ associated to regularization parameter λ is null if and only if $\lambda \geq \lambda_0(Y)$, where λ_0 is the zero-thresholding function given by

$$\lambda_0(Y) = (m_1 m_2)^{-1} \left\| \mathcal{T}(Y - \exp(\hat{X}_0)) \right\|_{\sigma,\infty},$$

where $\hat{X}_0 = \underset{X \in \mathcal{K},\quad \mathcal{T}(X)=0}{\operatorname{argmin}} \Phi_Y(X).$

Quantile universal threshold (QUT)

Ex: $X_{ij} = \alpha_i + \beta_j + \Theta_{ij}$, MLE $\hat{X}_0$ can be computed in closed form

$$(\hat{X}_0)_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j,$$

$$\begin{aligned}
\hat{\mu} &= \frac{1}{m_1}\sum_{i=1}^{m_1}\log(\sum_{j=1}^{m_2} Y_{ij}) + \frac{1}{m_2}\sum_{j=1}^{m_2}\log(\sum_{i=1}^{m_1} Y_{ij}) - \log(\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} Y_{ij}) \\
\hat{\alpha}_i &= \log(\sum_{j=1}^{m_2} Y_{ij}) - \frac{1}{m_1}\sum_{i=1}^{m_1}\log(\sum_{j=1}^{m_2} Y_{ij}) \\
\hat{\beta}_j &= \log(\sum_{i=1}^{m_1} Y_{ij}) - \frac{1}{m_2}\sum_{j=1}^{m_2}\log(\sum_{i=1}^{m_1} Y_{ij}).
\end{aligned}$$

$$\lambda_0(Y) = (m_1 m_2)^{-1} \left\| \mathcal{T}(Y - \exp(\hat{X}_0)) \right\|_{\sigma,\infty}.$$

Thresholding test

$$\mathcal{T}(\hat{X}^{\lambda}) = 0 \Leftrightarrow \lambda_0(Y) \leq \lambda$$

$\Rightarrow$ Definition: the **null thresholding statistic** $\Lambda = \lambda_0(Y)$, where Y comes from the null model $\mathcal{T}(X) = 0$

$\Rightarrow$ Test null hypothesis $\mathbf{H}_0 : \mathcal{T}(X) = 0$ against $\mathbf{H}_1 : \mathcal{T}(X) \neq 0$

$$\phi(Y) = \begin{cases} 1 & \text{if } \mathcal{T}(\hat{X}^{\lambda}) = 0 \\ 0 & \text{otherwise,} \end{cases}$$

defines a test of level $1 - \varepsilon$ for $\mathbf{H}_0$ if λ is a $1 - \varepsilon$ quantile of Λ.

$\Rightarrow$ Alternative to the Chi-square test.

$\Rightarrow$ Rank recovery property. Heuristic: large λ kills noise interaction and leaves real ones untouched

Quantile universal threshold

In practice distribution of Λ is unknown.

$\Rightarrow$ Parametric Bootstrap

- Compute $\hat{X}_0$ under H_0
- Generate M_1 Poisson matrices $Y_\ell \sim \mathcal{P}(\exp(\hat{X}_0))$, $1 \leq \ell \leq M_1$
- For all ℓ compute $\lambda_0(Y_\ell)$
- Set λ_{QUT} to the $1-\varepsilon$ quantile of the $\lambda_0(Y_\ell)$.

$\Rightarrow$ Monte Carlo simulation of the distribution of the largest singular value

$\Rightarrow$ Not computationally costly

Experiments

Existing method

$\Rightarrow$ Simulation under RC model

$$X_{ij} = \alpha_i + \beta_j + (UDV^\top)_{ij}$$

$\Rightarrow$ Vary size $m1, m2$, the rank K, the SNR $\frac{\|\Theta\|_{\sigma,1}}{\|\hat{X}_0\|_{\sigma,1}}$

- Estimation through maximization of a Poisson log-likelihood

$$\begin{aligned} \Phi_Y(X) &= (m_1 m_2)^{-1} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (Y_{ij} X_{ij} - \exp(X_{ij})) \\ \hat{X}^{\text{MLE}} &= \operatorname{argmax} \Phi_Y(X) \text{ s.t. } \operatorname{rk}(\Theta) = K \end{aligned}$$

- Implemented in the R package gnm (Turner and Firth, 2015)
- Requires to know K - Fails with large values of K, m_1, m_2

Choice of λ

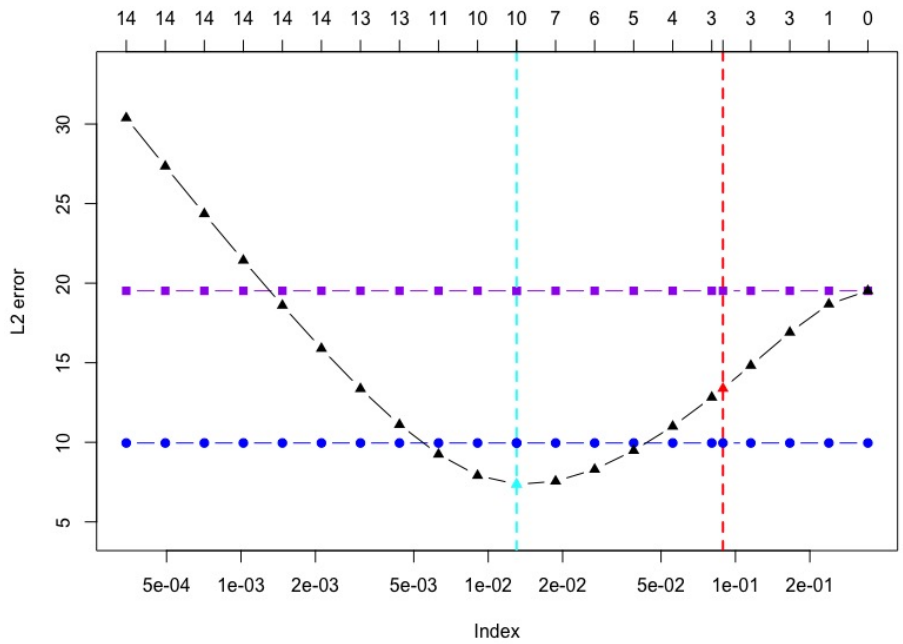


Figure: L_2 loss (black triangles) of ADMM estimator for $\lambda \in [1\mathrm{e}-4, 0.2]$ $m_1 = 20$, $m_2 = 15$, $K = 3$. Comparison of λ_{CV} (cyan dashed line) and λ_{QUT} (red dashed line) with the independence model RC(0) (purple squares) and the MLE with oracle rank RC(K) (blue points).

$\Rightarrow$ Two-step approach

Regularization grids

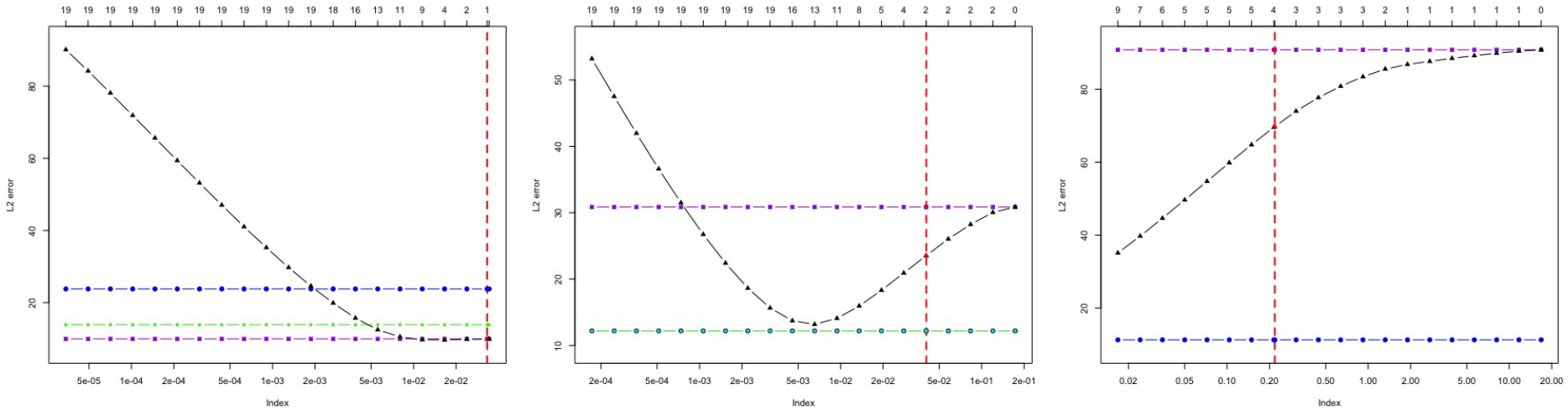


Figure: 50×20 matrices. Comparison of the L_2 error of GAMMIT (black triangles) with the independence model (purple squares), the rank oracle RC(K) model (blue points) and the RC(K_{QUT}) (green diamonds). Results are drawn for a grid of λ with λ_{QUT} (red dashed line). The rank of the interaction is written on the top axis for every λ. $K = 2$, $SNR = 0.2, 0.7, 1.7$ (left to right).

Thresholding test

N	chisq	thresh
13	1.00	1.00
673	0.95	0.96
4537	0.95	0.95
89556	0.95	0.94
990027	0.95	0.95

Table: Comparison of the levels of the thresholding and χ^2 tests.

$\Rightarrow$ needs further investigation (power, etc.)

Data analyses

Mortality data

Crosses 65 causes of death over 12 age categories in 2006 in France.

Use GAMMIT for biplot visualization

$$X_{ij} = \alpha_i + \beta_j + \tilde{U}_{i,.}\tilde{V}_{.,j} = \tilde{\alpha}_i + \tilde{\beta}_j - \frac{1}{2}\|\tilde{U}_{i,.}^T - \tilde{V}_{.,j}\|^2, \quad (2)$$

$\tilde{U} = UD^{1/2}$ and $\tilde{V} = VD^{1/2}$.

Represent data points on axis $(\tilde{U}, \tilde{V})$: two close points interact highly.

Mortality data

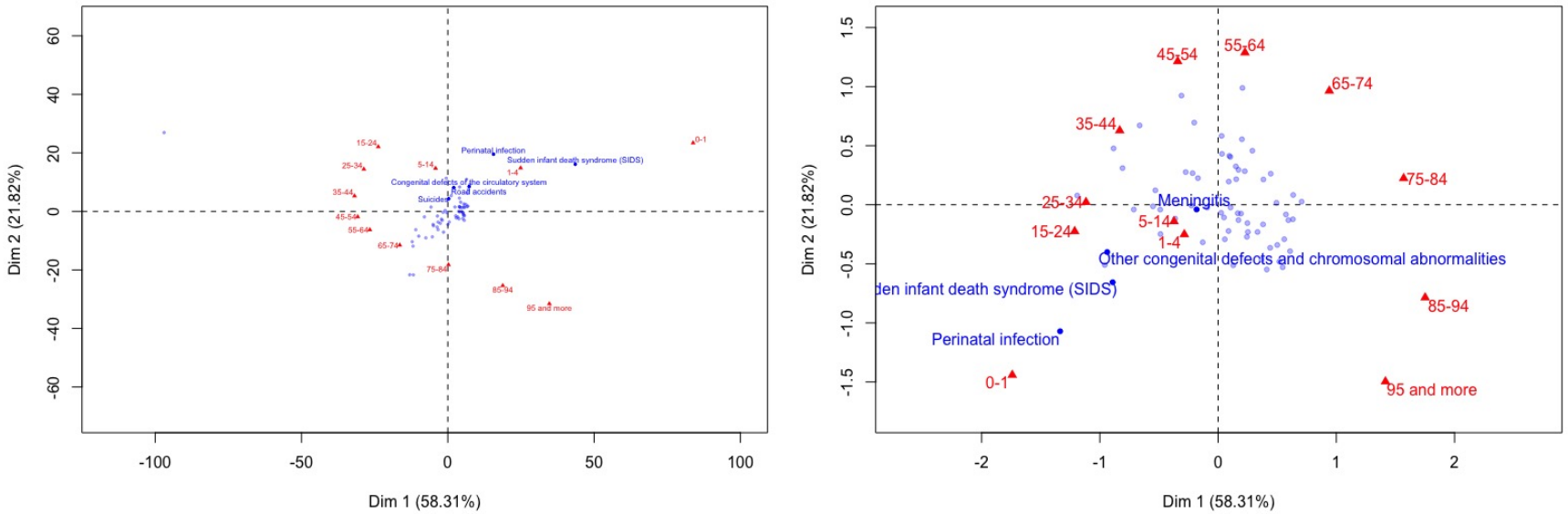


Figure: Visualization of the 10 largest interactions between age categories (red) and mortality causes (blue) in the two first dimensions of interaction with the RC(3) model (left) and GAMMIT (right).

Aravo data

- Crosses 82 species and 75 environments
- Environments and species covariates are known

$\Rightarrow$ Compare the results of GAMMIT with

- $X_{ij} = \alpha_i + \beta_j + \Theta_{ij}$ and
- $X_{ij} = (\alpha R)_{ij} + (C\beta)_{ij} + \Theta_{ij}$.

Aravo data

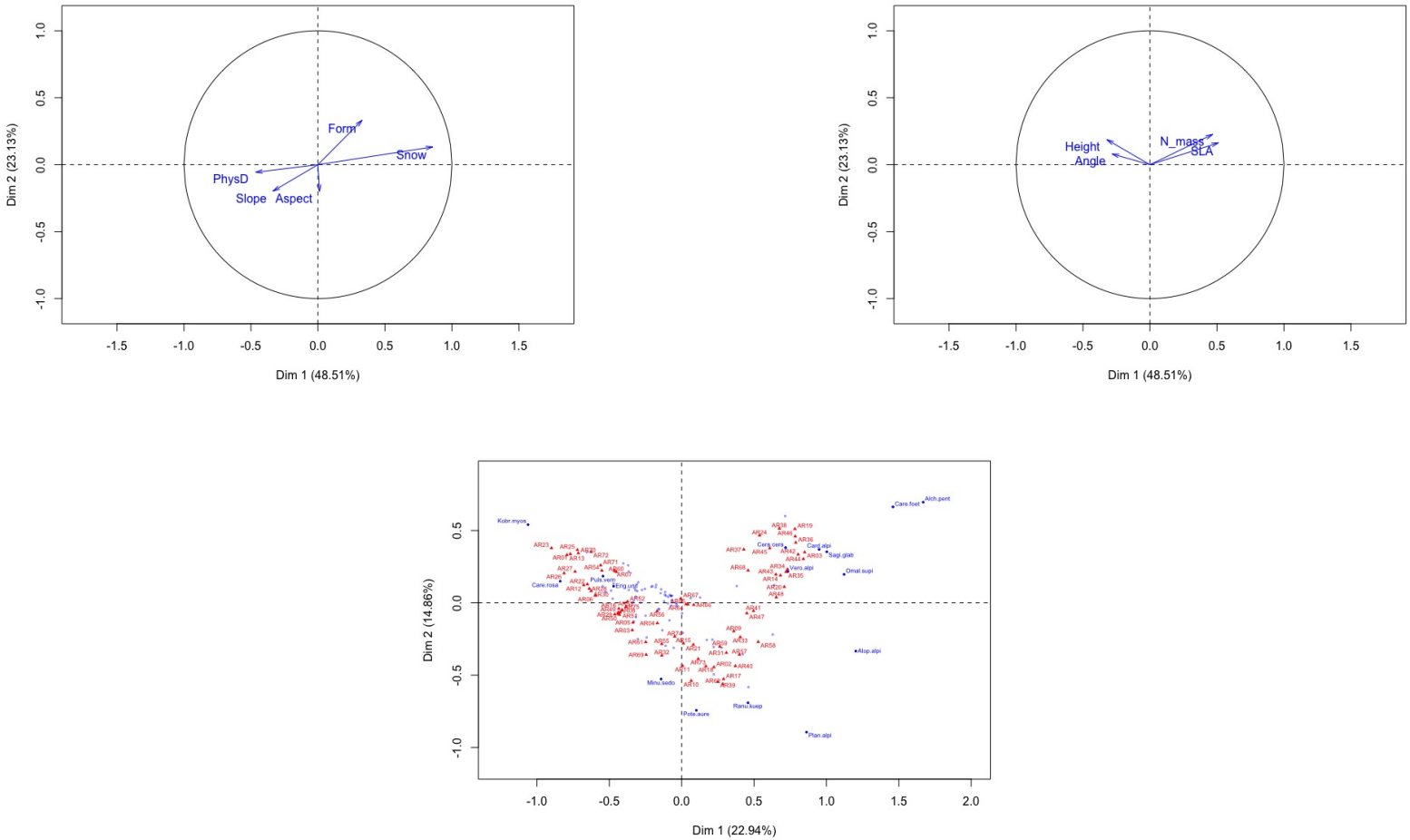
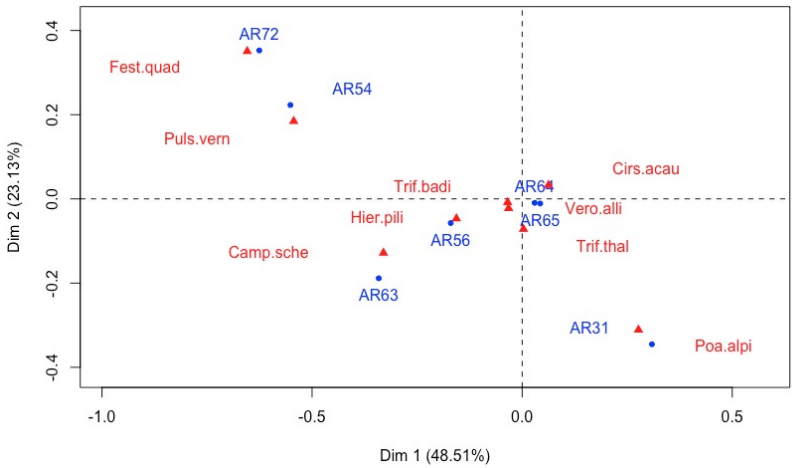


Figure: Correlation between environment (left) and species (right) covariates with the 2 first GAMMIT dim. biplot of the 2 first interaction dim. SLA (specific leaf area: ratio of the leaf surface to its dry mass)

Aravo data



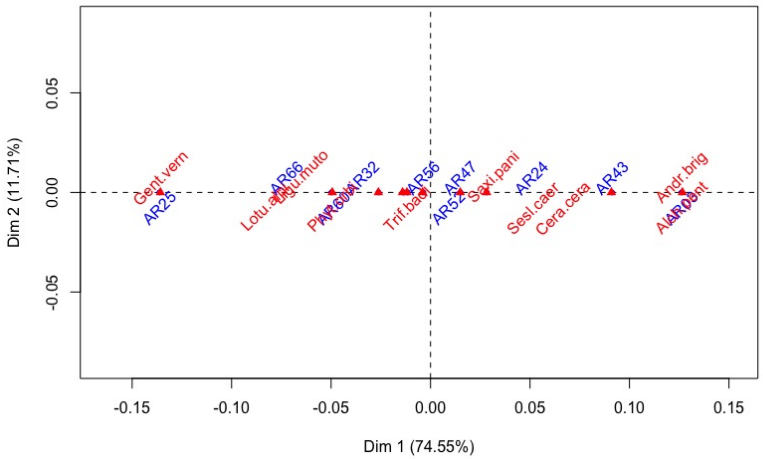


Figure: Visualization of the 10 largest interactions between environments (blue) and species (red) in the two first dimensions of interaction with GAMMIT for row-column indices (left) and explanatory covariates (right).

Conclusion

Summary

- Low-rank model for contingency table analysis with known covariates
- optimization algorithm, automatic choice of λ, rank recovery property
- Visualization and interpretation through biplots

Perspectives

- Adaptive regularization of singular values (Josse and Sardy, 2015)
- Add regularization of X_0
- Use GAMMIT to impute contingency tables
- Other sparsity inducing penalties
- Define a pivotal test statistic for QUT test

References

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

Giacobino, C., S. Sardy, J. Diaz Rodriguez, and N. Hengardner (2016). Quantile universal threshold for model selection. *arXiv:1511.05433v2*.

Josse, J. and S. Sardy (2015). Adaptive shrinkage of singular values. *Statistics and Computing*, 1–10.

Turner, H. and D. Firth (2015). *Generalized nonlinear models in R: An overview of the gnm package*. R package version 1.0-8.