Low-rank Interaction Contingency Tables

Geneviève Robin, Julie Josse, Éric Moulines, Sylvain Sardy

Low-rank Interaction Contingency Tables

Geneviève Robin[1], Julie Josse[1], Éric Moulines[1]
Sylvain Sardy[2]
[1]Center of Applied Mathematics, École Polytechnique
[2]Department of Mathematics, Université de Genève

September 18, 2017

Abstract

Contingency tables are collected in many scientific and engineering tasks including image processing, single-cell RNA sequencing and ecological studies. Low-rank methods have proved useful to analyze them, by facilitating visualization and interpretation. However, common methods do not take advantage of extra information which is often available, such as row and column covariates. We propose a method to denoise and visualize high-dimensional count data which directly incorporates the covariates at hand. Estimation is done by minimizing a Poisson log-likelihood and enforcing a low-rank structure on the interaction matrix with a nuclear norm penalty. We also derive theoretical upper and lower bounds on the Frobenius estimation risk. A complete methodology is proposed, including an algorithm based on the alternating direction method of multipliers, and automatic selection of the regularization parameter. The simulation study reveals that our estimator compares favorably to competitors. Then, analyzing environmental science data, we show the interpretability of the model using a biplot visualization. The method is available as an R package.

Keywords Count data; Dimensionality reduction; Ecological data; Low-rank matrix recovery, Quantile universal threshold

1 Introduction

Consider an $m_1 \times m_2$ observation matrix of counts Y with independent cells of expectations $E(Y_{ij}) = \exp(X^*_{ij})$. The log-linear model [Agresti, 2013, Christensen, 2010] with rank constrained interaction, often referred to as the *generalized additive main effects and multiplicative interaction* model [Goodman, 1985, de Falguerolles, 1998, Gower et al., 2011, Fithian and Josse, 2017] or the *row-column model* of rank K, is commonly used to describe the structure of the matrix X^*, and is written as follows:

$$X^*_{ij} = \alpha^*_i + \beta^*_j + \Theta^*_{ij}, \ \mathrm{rk}(\Theta^*) = K, \tag{1}$$

where $\mathrm{rk}(\Theta^*)$ denotes the rank of Θ^* and $K \leq \min(m_1 - 1, m_2 - 1)$. In these models, the terms which only depend on the index of the row or column (α^*_i and β^*_j) are called *main effects*, and terms which depend on both (here Θ^*_{ij}) are called *interactions* [Kateri, 2014, Section 4.1.2, p.87]. The estimation of the means $\exp(X^*_{ij})$ is done by minimizing a normalized negative Poisson

log-likelihood, defined for $X \in \mathbb{R}^{m_1 \times m_2}$ by

$$\Phi_Y(X) = -\frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \left(Y_{ij} X_{ij} - \exp(X_{ij}) \right). \tag{2}$$

Our first contribution is to introduce a two-fold extension of the *row-column model* (1). First, our method allows us to incorporate general covariates as well as interactions between them. Second, instead of assuming that the rank is fixed, we penalize the nuclear norm of the interaction matrix. More formally, let $R \in \mathbb{R}^{m_1 \times K_1}$ (resp. $C \in \mathbb{R}^{m_2 \times K_2}$) be matrices of known row (resp. column) covariates, and $\alpha^* \in \mathbb{R}^{K_1 \times m_2}$ (resp. $\beta^* \in \mathbb{R}^{K_2 \times m_1}$) matrices of unknown parameters. We model the matrix X^* as follows:

$$X^* = R\alpha^* + (C\beta^*)^T + \Theta^*, \tag{3}$$

with $(C\beta^*)^T$ denoting the transpose of matrix $(C\beta^*)$. In model (3), $R\alpha^*$ and $C\beta^*$ incorporate *interactions*, i.e., $(R\alpha^*)_{ij}$ and $(C\beta^*)^T_{ij}$ both depend on the indices i and j. Lastly, Θ^* corresponds to the interaction unexplained by the known covariates R and C. In the ecology example we present, columns of the contingency table represent species while rows represent environments, and cell Y_{ij} counts the abundance of species j in environment i. The row features R embed geographical information about the environments such as the slope and temperature, while C codes physical traits about species like height or mass. $R\alpha^*$ corresponds to effects, which can depend on the species, of the environmental covariates (temperature, etc.); a similar interpretation can be made for $C\beta^*$.

The paper is organized as follows. In Section 2, we define our estimator through the minimization of the negative log-likelihood term (2) penalized by the nuclear norm of the interaction matrix Θ^*. We also propose an optimization algorithm based on the *alternating descent method of multipliers* [Boyd et al., 2011]. Under mild assumptions on the true parameter matrix X^*, we derive in Section 3 upper and lower bounds for the estimation risk, that hold with high probability and for a number of generalized linear models. Another major contribution is to propose in Section 4 two methods to choose the regularization parameter automatically. Lastly, we show in Section 5 that on simulated data our procedure compares favorably to competitors, and highlight in Section 6 the interpretability of the method with an application in ecology. The methods and experiments presented in this article are available as an R [R Core Team, 2016] package at `https://github.com/genevievelrobin/gammit`.

Related approaches for count matrix recovery and dimensionality reduction can be embedded within the framework of low-rank exponential family estimation [Collins et al., 2001, de Leeuw, 2006, Li and Tao, 2013, Josse and Wager, 2016, Liu et al., 2016] as well as its Bayesian counterpart [Mohamed et al., 2009, Gopalan et al., 2014]. Existing models impose low ranks either to the parameter matrix X^* [Collins et al., 2001] or to the mean matrix with cells $\exp(x^*_{ij})$ [Liu et al., 2016, Josse and Wager, 2016]. Estimation approaches include iterative partial updates of the parameters [Salmon et al., 2014] and augmented Lagrangian methods [Figueiredo and Bioucas-Dias, 2010, Chambolle and Pock, 2011, Jeong et al., 2013]. The theoretical performance of nuclear norm penalized estimators for Poisson denoising has been studied in Cao and Xie [2016], where the authors prove uniform bounds on the empirical error risk by extending results from compressed sensing and 1-bit matrix completion [Raginsky et al., 2010, Davenport et al., 2012], and in Lafond [2015] where optimal bounds are proved for matrix completion in the exponential family. Poisson matrix estimation has also been considered via singular value shrinkage, extending the Gaussian setting [Shabalin and Nobel, 2013, Gavish and Donoho, 2014a,b, Josse

and Sardy, 2015]. Bigot et al. [2016] have studied optimal singular value shrinkage in the exponential family, while Liu et al. [2016] have suggested a new shrinkage for covariance matrix estimation.

None of these methods above can account for the effect of known covariates, and to the best of our knowledge, no algorithm or theory has been developped for model (3). Attempts to include row and column effects in matrix recovery and completion have nonetheless been made in the context of the Netflix challenge, but they do not use additional features as we do. Some are reviewed in Feuerverger et al. [2012], and Hastie et al. [2014] briefly addressed this issue through centering and scaling steps.

2 Low-rank Interaction Contingency Tables

2.1 Notation and model

For $A \in \mathbb{R}^{m_1 \times m_2}$, we denote $\|A\|_*$ the sum of the singular values of A (nuclear norm), $\|A\|_F$ the Frobenius norm, $\|A\|$ the the largest singular value (operator norm), and $\|A\|_\infty$ the largest entry in absolute value. We also denote $\mathcal{V}_R^\perp$ (resp. $\mathcal{V}_C^\perp$) the subspace of $\mathbb{R}^{m_1 \times m_2}$ of matrices whose columns (resp. rows) are orthogonal to the columns of R (resp. rows of C), and $\mathcal{V}^\perp = \mathcal{V}_R^\perp \bigcap \mathcal{V}_C^\perp$. We make the following assumption, common in the Poisson matrix denoising litterature.

Assumption 1. *There exist* $\gamma_{\text{MIN}} > -\infty$ *and* $\gamma_{\text{MAX}} < \infty$ *such that for all* $i = 1, \dots, m_1$ *and* $j = 1, \dots, m_2$,

$$\gamma_{\text{MIN}} \leq \log E(Y_{ij}) \leq \gamma_{\text{MAX}}.$$

Moreover there exist $\sigma_{\text{MIN}} > 0$ *and* $\sigma_{\text{MAX}} < \infty$ *such that for all* $i = 1, \dots, m_1$ *and* $j = 1, \dots, m_2$,

$$\sigma_{\text{MIN}}^2 \leq \text{var}(Y_{ij}) \leq \sigma_{\text{MAX}}^2.$$

Define the compact set $\mathcal{K} = [\gamma_{\text{MIN}}, \gamma_{\text{MAX}}]^{m_1 \times m_2}$. We can now define our estimator, for a given regularization parameter λ, as the minimizer of the penalized negative log-likelihood:

$$\tilde{\alpha}^\lambda, \tilde{\beta}^\lambda, \tilde{\Theta}^\lambda = \underset{\substack{R\alpha + (C\beta)^T + \Theta \in \mathcal{K} \\ \Theta \in \mathcal{V}^\perp}}{\text{argmin}} \quad \phi_Y(\alpha, \beta, \Theta) + \lambda \|\Theta\|_*, \tag{4}$$

$$\phi_Y(\alpha, \beta, \Theta) = \Phi_Y(R\alpha + (C\beta)^T + \Theta), \tag{5}$$

where $\Theta \in \mathcal{V}^\perp$ serves as an identifiability constraint and Φ_Y is defined in (2). This problem is not jointly convex or separable in α, β and Θ. Our contribution on the optimization side is to derive a reparametrization of (4) which simplifies computation.

2.2 Reparametrization

Define $\mathcal{T}$ the projection operator on the subspace $\mathcal{V}^\perp = \mathcal{V}_R^\perp \bigcap \mathcal{V}_C^\perp$. The reformulated problem

$$\hat{X}^\lambda, \hat{\Theta}^\lambda = \underset{\substack{X \in \mathcal{K} \\ \Theta = \mathcal{T}(X)}}{\text{argmin}} \quad \Phi_Y(X) + \lambda \|\Theta\|_*, \tag{6}$$

where α and β are not included explicitly in the minimization problem, yields the same solution in Θ and X as problem (4): $\hat{\Theta}^\lambda = \tilde{\Theta}^\lambda$ and $\hat{X}^\lambda = R\tilde{\alpha}^\lambda + \left(C\tilde{\beta}^\lambda\right)^T + \tilde{\Theta}^\lambda$. Moreover, the identifiability

constraint $\mathcal{T}(X) = \Theta$ ensures that we can compute $R\hat{\alpha}^\lambda + (C\hat{\beta}^\lambda)^T$ a posteriori based on $\hat{X}^\lambda$ and $\hat{\Theta}^\lambda$ only by applying simple projections. Problem (6) is now strongly convex on a compact set, linearly constrained and separable in X and Θ. The parameter set $\mathcal{K}$ is compact and Φ_Y^λ is strongly convex on $\mathcal{K}$, which guarantee existence and uniqueness of the solution of (6). We solve (6) by using the *alternating directions method of multipliers* [Glowinski and Marrocco, 1974], whose convergence stems from Boyd et al. [2011, Theorem 3.2.1].

2.3 Optimization algorithm

The alternating direction method of multipliers is a variant of the augmented Lagrangian method of multipliers which solves the dual problem through iterated partial updates. The augmented Lagrangian, indexed by a positive real parameter τ is

$$\mathcal{L}_\tau(X, \Theta, \Gamma) = \Phi_Y(X) + \lambda \|\Theta\|_* + \langle \Gamma, \mathcal{T}(X) - \Theta \rangle + \frac{\tau}{2} \|\mathcal{T}(X) - \Theta\|_F^2 , \tag{7}$$

where $\langle ., . \rangle$ denotes the trace scalar product on $\mathbb{R}^{m_1 \times m_2}$. The algorithm consists in updating separately the primal variables X, Θ, and the dual variable Γ, at iteration ℓ to maximize (7) according to the following equations:

$$\begin{aligned} X^{\ell+1} &= \operatorname{argmin}_{X \in \mathcal{K}} \quad \mathcal{L}_\tau(X, \Theta^\ell, \Gamma^\ell) \\ \Theta^{\ell+1} &= \operatorname{argmin}_{\Theta \in \mathcal{K}_\mathcal{T}} \quad \mathcal{L}_\tau(X^{\ell+1}, \Theta, \Gamma^\ell) \\ \Gamma^{\ell+1} &= \Gamma^\ell + \tau(\mathcal{T}(X^{\ell+1}) - \Theta^{\ell+1}). \end{aligned} \tag{8}$$

The function Φ_Y and $\|.\|_*$ are closed, proper and convex on $\mathbb{R}^{m_1 \times m_2}$. This guarantees the solvability of the minimization problems defined in update (8). Moreover Φ_Y is differentiable, so the optimization in X can be done using gradient descent. The update of Θ can itself be done in closed form and involves singular value decomposition and thresholding [Cai et al., 2010]:

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

Here, $\mathcal{D}_{\lambda/\tau}$ is the soft-thresholding operator of singular values at level λ/τ. To speed convergence, we implemented a warm-start strategy [Friedman et al., 2007, Hastie et al., 2015]. We start by running the algorithm with $\lambda = \lambda_0(Y)$, the smallest value of the regularization parameter that sets the interaction to 0 (see Section 4); we then solve the optimization problem for decreasing values of λ, each time using the previous estimator as an initial value. As for the tuning of parameter τ, we apply the method described in Boyd et al. [2011], Section 3.4.1, which consists in having it vary at every iteration, depending on the value of the residual.

2.4 Remarks

Estimator (6) is very similar to what can be found in the matrix completion literature where data-fitting losses penalized by the nuclear norm are optimized [Klopp, 2014, Lafond, 2015]. Problems are often written as $\hat{X}^\lambda = \operatorname{argmin}_X \quad L(X; Y) + \lambda \|X\|_*$, where L is a loss function, and the main difference with (6) is in the regularization. By penalizing $\Theta = \mathcal{T}(X)$, our method actually regularizes only the directions in X which are orthogonal to the covariates R and C.

A possible substitute to the alternating direction method of multipliers is alternating minimization, which has been used in low-rank estimation problems [Udell et al., 2014]. This consists

in partially minimizing the objective in (4) alternatively with respect to Θ, α and β, while keeping all other parameters fixed. In our case, the optimization in Θ with fixed α and β yields a constrained problem of the form

$$\Theta^{\ell+1} = \underset{\substack{R\alpha^{\ell}+\left(C\beta^{\ell}\right)^{T}+\Theta\in\mathcal{K} \\ \Theta\in\mathcal{V}^{\perp}}}{\operatorname{argmin}} \phi_Y(\alpha^{\ell}, \beta^{\ell}, \Theta) + \lambda \|\Theta\|_*,$$

which has itself to be solved with the alternating direction method of multipliers or, for example, projected gradient methods, and is therefore more computationally intensive.

Lastly, the estimation procedure does not depend on the explicit form of the Poisson likelihood, but only relies on the convexity of Φ_Y. Other generalized linear models can therefore be handled directly.

3 Statistical Guarantees

3.1 Upper bound

We now derive an upper bound on the Frobenius estimation error of estimator (6). Denote $M = \max(m_1, m_2)$.

Assumption 2 (Subexponentiality)**.** *There exists $\delta > 0$ such that for all $i = 1, \ldots, m_1$ and $j = 1, \ldots, m_2$,*

$$E\left(\exp(|Y_{ij}|/\delta)\right) < +\infty.$$

Theorem 1. *There exists a constant c such that the following statement holds. Set*

$$\lambda = 2c\sigma_{\text{MAX}} \frac{(2M\log(m_1+m_2))^{1/2}}{m_1 m_2}.$$

Under Assumptions 1 and 2, assume $m_1 + m_2 \geq \max\left\{\delta^2(2\sigma^2_{\text{MAX}}\sigma^2_{\text{MIN}})^{-1}, (4\delta^2/\sigma^2_{\text{MAX}})^4\right\}$. Then with probability at least $1-(m_1+m_2)^{-1}$,

$$\frac{\left\|X^* - \hat{X}^{\lambda}\right\|_F^2}{m_1 m_2} \leq \left(\frac{4\sigma^2_{\text{MAX}}}{\sigma^4_{\text{MIN}}}\right) \frac{M\left(\text{rk}(\Theta^*) + K_1 + K_2\right)\log(m_1+m_2)}{m_1 m_2}. \tag{9}$$

Proof. See Section 8.1. □

The constant that appears in bound (9) grows linearly with the upper bound σ^2_{MAX} and quadratically with the inverse of σ^2_{MIN}. This means that by relaxing Assumption 1 to allow $\text{var}(Y_{ij})$ to grow as fast as $\log(m_1+m_2)$ or decrease as fast as $1/\log(m_1+m_2)$, we only loose a log-polynomial factor in the bound. Furthermore, the explicit form of the data-fitting term Φ_Y does not appear in these results. The theoretical results therefore hold for a number of other generalized linear models, including multinomial and exponential ones, as does the inference procedure.

3.2 Lower bound

We now derive a lower bound on the Frobenius estimation error. Define $\gamma = \min(|\gamma_{\text{MIN}}|, |\gamma_{\text{MAX}}|)$, where γ_{MIN} and γ_{MAX} are defined in Assumption 1, and $\mathcal{F}(r, \gamma)$ the set of matrices

$$\mathcal{F}(r, \gamma) = \left\{ X \in \mathbb{R}^{m_1 \times m_2} : \text{rk}\left(\mathcal{T}(X)\right) \leq r, \|X\|_\infty \leq \gamma \right\}.$$

Denote $N = r(M - K_1) + K_1 m_2 + K_2 m_1 - K_1 K_2$ and for $X \in \mathbb{R}^{m_1 \times m_2}$ $\mathbb{P}_X$ the law of $m_1 \times m_2$ independant random Poisson variables with means $\exp(X_{ij})$.

Theorem 2. *For all $m_1, m_2 \geq 2$, $1 \leq r \leq m$,*

$$\inf_{\hat{X}} \sup_{X \in \mathcal{F}(r,\gamma)} \mathbb{P}_X \left(\frac{\left\| \hat{X} - X \right\|_F^2}{m_1 m_2} > \min\left\{ 2\gamma^2, \frac{N}{m_1 m_2 \sigma_{\text{MAX}}^2} \right\} \right) \geq d(\eta),$$

$$d(\eta) = \frac{1}{1 + 2^{-N/16}} \left(1 - 2\eta - \frac{1}{2} \frac{\eta^{1/2}}{(N \log(2))^{1/2}} \right),$$

where the infimum is computed over all estimators.

4 Automatic selection of λ

4.1 Cross-validation

The cross-validation procedure consists in erasing a fraction of the observed values in Y, estimating a complete parameter matrix $\hat{X}^\lambda$ for a range of λ values, and choosing the parameter λ that minimizes the prediction error. Because the entries Y_{ij} are independent the estimation can be done by skipping the missing entries. Let Ω denote the set of indices of the observed entries, and denote $\Phi_{\Omega(Y)}$ the negative log-likelihood taken at the observed entries only. The optimization problem becomes

$$\hat{X}^\lambda, \hat{\Theta}^\lambda = \underset{\substack{X \in \mathcal{K} \\ \Theta = \mathcal{T}(X)}}{\text{argmin}} \quad \Phi_{\Omega(Y)}(X) + \lambda \|\Theta\|_*,$$

and can also be solved using the alternating direction method of multipliers. Repeating this procedure N times for a grid of λ, we select the value λ_{CV} that minimizes the prediction squared error $\text{PSE}(\lambda) = N^{-1} \sum_{i=1}^N \left\| Y_{\text{mis}} - \hat{X}^{(i)}_{\lambda,\text{mis}} \right\|_F^2$. In the process, we have defined an algorithm to estimate X^* from incomplete observations, which can be seen as a single imputation method and still holds when entries are *missing at random* (Little and Rubin [1987, 2002], Section 1.3), and could be used to complete contingency tables with missing values.

4.2 Quantile universal threshold

We suggest an alternative method to cross-validation inspired by Donoho and Johnstone [1994], and by the work of Giacobino et al. [2016] on *quantile universal thresholds*. In Proposition 3 below, we define the so-called *zero-thresholding statistic* of estimator (3), a function of the data $\lambda_0(Y)$ for which the estimated interaction matrix $\hat{\Theta}^{\lambda_0(Y)}$ is null and the same estimate $\hat{\Theta}^\lambda = 0$ is obtained for any $\lambda \geq \lambda_0(Y)$. We prove Proposition 3 in Section 8.4.

Proposition 3 (Zero-thresholding statistics)**.** *The interaction estimator* $\hat{\Theta}^{\lambda}$ *associated with regularization parameter* λ *is null if and only if* $\lambda \geq \lambda_0(Y)$*, where* $\lambda_0(Y)$ *is the zero-thresholding statistic*

$$\lambda_0(Y) = \frac{1}{m_1 m_2} \left\| Y - \exp(\hat{X}_0) \right\|, \quad \hat{X}_0 = \underset{X \in \mathcal{K},\ \mathcal{T}(X)=0}{\operatorname{argmin}} \Phi_Y(X).$$

We propose a heuristic selection of λ based on this zero-thresholding statistic $\lambda_0(Y)$. To explain further the procedure, we first need to define the following test:

$$\mathbf{H}_0 : \Theta^* = 0 \quad \text{against the alternative} \quad \mathbf{H}_1 : \Theta^* \neq 0 \tag{10}$$

which actually boils down to testing if the measured covariates are sufficient to explain the interaction. For $0 < \varepsilon < 1$, consider a value λ_ε that satisfies $\mathbb{P}_{\mathbf{H}_0}(\lambda_0(Y) > \lambda_\varepsilon) < \varepsilon$. The test which consists in comparing the statistics $\lambda_0(Y)$ to λ_ε is of level $1 - \varepsilon$ for (10). This can be seen as an alternative to the χ^2 test for independence, with the additional advantage of handling covariates. In practice we do not have access to the distribution under the null $\mathbb{P}_{\mathbf{H}_0}(\lambda_0(Y) < \lambda)$, but perform parametric bootstrap [Efron, 1979] to compute a proxy $\check{\lambda}_\varepsilon$. We define $\lambda_{\mathrm{QUT}} := \check{\lambda}_{.05}$ the value we use in practice, and refer in what follows to this method of selecting λ as *quantile universal threshold*. In Section 5.1 we show that cross-validation achieves good prediction errors, while the quantile universal threshold method has better rank recovery properties.

5 Simulation study

5.1 Comparison of quantile universal threshold and cross-validation

Here, we generate a contingency table according to the model $Y \sim \mathcal{P}(\exp(X^*))$, where $\mathcal{P}$ denotes the Poisson distribution, with $X^* = X_0^* + \Theta^*$, $(X_0^*)_{ij} = \alpha_i^* + \beta_j^*$. We draw row and column effects α_i^* and β_j^* uniformly and generate $\Theta^* = UDV^T$ with random orthonormal matrices $U = (u_{ij})$ and $V = (v_{ij})$, where $D \in \mathbb{R}^{K \times K}$ is a diagonal matrix with the singular values of Θ^* on its diagonal. The parameters of our simulation are the size of X^* ($m_1 \times m_2$), the rank K of Θ^* and the ratio of the nuclear norm of the interaction Θ^* to the nuclear norm of the additive part X_0^*, denoted $\mathrm{SNR} = \|\Theta^*\|_* / \|X_0^*\|_*$.

We start the simulation study without additional covariates to compare our estimator in terms of L_2 error to a competitor, the estimator of the row-column model (1) with different ranks: the independence model with rank 0, the oracle rank K and the rank $\hat{K}_{\mathrm{QUT}}$ estimated with quantile universal threshold. We consider a representative setting with $m_1 = 20$, $m_2 = 15$ and $K = 3$. Figure 1 shows the L_2 error of recovery between the estimator $\hat{X}^{\lambda}$ and the true parameter X^* as a function of λ. The maximum likelihood estimation in the independence model ($\Theta^* = 0$) can be used as a benchmark. When λ is close to 0 we recover the saturated (unconstrained) model , while as λ increases, we tend to the independence model. The rank of the estimator $\hat{\Theta}^{\lambda}$, which we define here as the number of singular values above 10^{-6}, decreases with λ. The two proposed procedures for choosing λ prove useful: λ_{QUT} selects the correct rank ($K = 3$) for the interaction, and cross-validation achieves the best prediction error. An alternative procedure would be a two-step approach where we fit the maximum likelihood estimator with the rank found using quantile universal threshold. We observe the same results over 1000 replications.

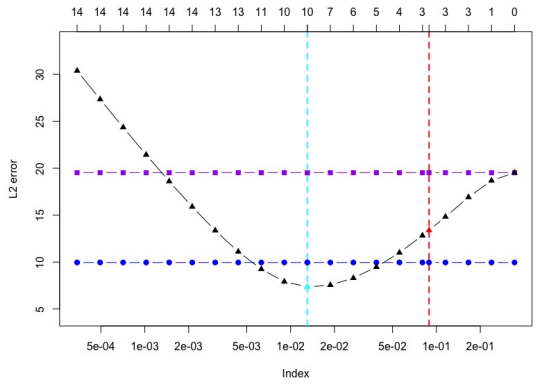


Figure 1: L_2 loss (black triangles) of the estimator as a function of λ; $m_1 = 20$, $m_2 = 15$, $K = 3$. Comparison of λ_{CV} (cyan dashed line) and λ_{QUT} (red dashed line) with the independence model (purple squares) and the *row-column model* with oracle rank (blue points). The rank of Θ is written along the top for each λ.

5.2 Estimation performance

With the same simulation scheme, we further investigate the performance of our method in different situations. Figure 2 highlights three interaction regimes. Over all values of the rank we observe similar behaviors. In the small interaction regime (Figure 2, top left, SNR = 0.2), the interaction is too small to be distinguished from the Poisson noise, so the independence model achieves the best performance. The rank selected by quantile universal threshold is of 1, and we see that the error of the row-column model with rank 1 is very close to that with rank 0. In the medium interaction regime (Figure 2, top center, SNR = 0.7) we recover the correct rank of 2 with quantile universal threshold but have a higher error than the oracle row-column model with rank 2. These two situations suggest to use a two-step procedure. In the high interaction setting (Figure 2, top right SNR = 1.7), quantile universal threshold overestimates the rank (here 6 instead of 2), and the row-column model fails to calculate the maximum likelihood estimation (possibly because of numerical issues that occur in available R libraries).

6 Analysis of the Aravo data

6.1 Description of the data

The Aravo dataset [Choler, 2005] measures the abundance of 82 species of alpine plants in 75 sites in France. Initially, ecologists aimed to understand how species interacted with different biological environments, trying to uncover whether certain species thrive or decay in specific environments. The data consist of a contingency table collecting the abundance of species across sampling sites. Covariates about the environments and species are also available, with 8 species traits, providing physical information about plants (height, spread, etc.), as well as 6 environmental variables about the geography and climate of the various sites. These covariates are considered as known factors of variability, and the question is whether they are sufficient to describe the interaction. In this study we will compare the simple model (1) where the covariates

are not taken into account with our model (3), to see how the incorporation of covariates impacts the interpretation.

6.2 Comparison of two models

We show in Figure 3 biplot visualizations of the data in the two first dimensions of interaction, defined by the first two singular vectors of $\hat{\Theta}^{\lambda}$. The plots are interpreted in terms of distance as follows: a species and an environment that are close interact highly [Fithian and Josse, 2017]. The first difference between the two models is in the rank of $\hat{\Theta}^{\lambda}$. In Figure 3a where we do not use the covariates, we find a rank of 3 for the interaction, while in Figure 3b after incorporating the covariates we find a rank 1. This suggests that an additional unknown variable summarizes the remaining interactions. We can also compare the distances between species and environments before and after discarding the variability due to the covariates. Figure 3 shows the species and environments that have the 10 highest interactions (smallest distances on the biplot), for both models. We see that the species and environments involved differ, and thus that our procedure could possibly lead to new interpretations. In particular, after incorporating covariates, we

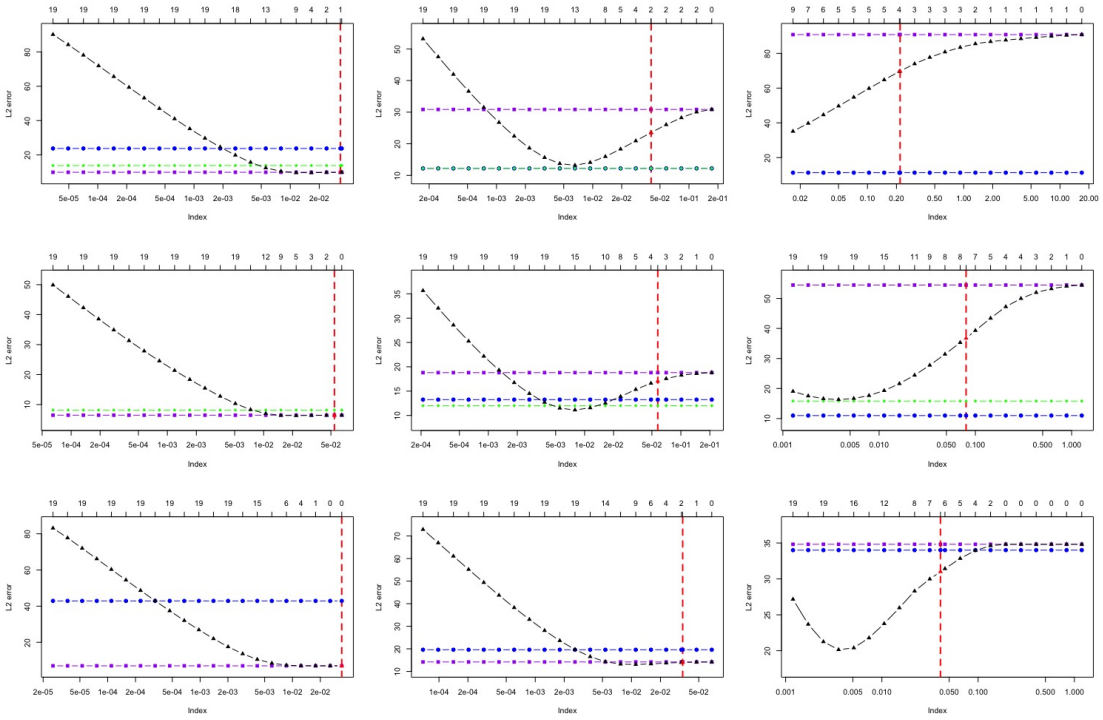


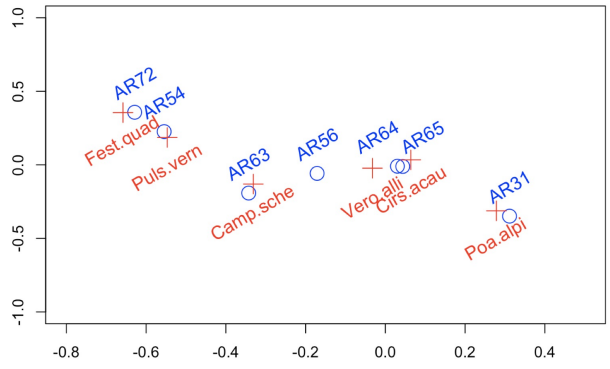
Figure 2: 50×20 matrices. Comparison of the L_2 error of our method (black triangles) with the independence model (purple squares), the row-column model with oracle rank (blue points) and with rank $\hat{K}_{\mathrm{QUT}}$ (green diamonds). Results are drawn for a grid of λ with λ_{QUT} (red dot). The rank of the interaction is written along the top for each value of λ. Top $K = 2$, middle $K = 5$, bottom $K = 10$. From left to right $\mathrm{SNR} = 0.2, 0.7, 1.7$.

extract species-environment couples more clearly.

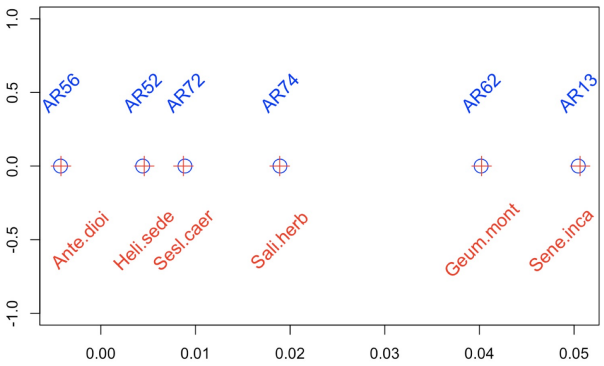
In the case of model (1) with no covariates, we can look at the correlations between the known traits and the interaction directions of $\hat{\Theta}^{\lambda}$. Figure 4a shows that environment covariates and the two first directions of interaction are correlated. The first direction is correlated with the amount of *Snow*, and the second with the *Aspect* variable (which denotes the compass direction, e.g., north, south, etc. that the site faced). On the biplot in Figure 3, the first direction therefore separates environments with respect to the amount of snow, while the second direction separates environments with respect to compass direction. Similarly, in Figure 4b, the species covariates are correlated with the estimated directions of interaction, therefore in Figure 3 the first direction separates the plants with respect to their *SLA* (specific leaf area, defined as the ratio of the leaf surface to its dry mass) and their mass-based nitrogen content. (*Nmass*). On the contrary, when we incorporate the covariates in the model, the correlations between the known traits and the interaction directions are reduced by a factor of between 3 and 10 (these are now too small to be represented on a plot).

7 Discussion

We conclude by discussing some opportunities for further research. One possible extension is to also penalize the main covariate effects, with an ℓ_1 penalty on α and β. Secondly, our algorithm directly handles missing values and could be used to impute contingency tables, but theoretical guarantees would have to be extended to the missing data framework. The properties of the thresholding test also merit further investigation, in particular the power could be assessed. Lastly, it may be of interest to consider other sparsity inducing penalties. In particular, penalizing the Poisson log-likelihood by the absolute values of the coefficients of the interaction matrix Θ could possibly lead to solutions where some interactions are driven to 0 and a small number of large interactions are selected.



(a) Without covariates (model (1))



(b) With covariates (model (3))

Figure 3: Comparison of biplot visualizations with models (1) and (3). Environments are represented in blue and species in red, in the Euclidean space defined by the two first principal directions of $\hat{\Theta}^{\lambda}$.

8 Proofs

8.1 Proof of Theorem 1

The proof of Theorem 1 derives from the strong convexity of $\Phi_Y$and tail bounds for the largest singular value of random matrices with subexponential entries. For the sake of clarity, we write in what follows $\hat{X}$ and $\hat{\Theta}$ instead of $\hat{X}^\lambda$ and $\hat{\Theta}^\lambda$. We first state the following result.

Proposition 4. *Under Assumption 1, assume* $\lambda \geq 2\left\|\nabla\Phi_Y(X^*)\right\|$. *Then*

$$\frac{\left\|X^* - \hat{X}_\lambda\right\|_F^2}{m_1 m_2} \leq \left(\frac{16\lambda^2 m_1 m_2}{\sigma_{\text{MIN}}^4}\right)\left(\text{rk}(\Theta^*) + K_1 + K_2\right). \tag{11}$$

We prove this result in Section 8.2.

Proposition 4 is deterministic but relies on the condition $\lambda \geq 2\left\|\nabla\Phi_Y(X^*)\right\|$ which is a random. Let us therefore compute a value of λ such that this condition holds with high probability. We define the random matrices

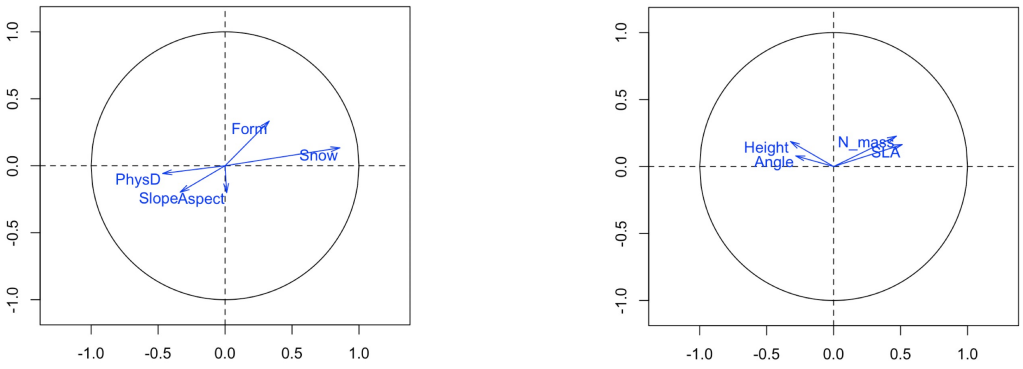
$$Z_{ij} = (Y_{ij} - \exp(X^*_{ij}))E_{ij},$$

with E_{ij} the (i,j)-th canonical matrix, and the quantity

$$\sigma_Z^2 = \max\left(\frac{1}{m_1 m_2}\left\|\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} E\left(Z_{ij}Z_{ij}^T\right)\right\|, \quad \frac{1}{m_1 m_2}\left\|\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} E\left(Z_{ij}^T Z_{ij}\right)\right\|\right). \tag{12}$$

We have

$$E\left(Z_{ij}\right) = 0 \text{ and } \sigma_{\text{MIN}} \leq E\left(\left\|Z_{ij}Z_{ij}^T\right\|\right), E\left(\left\|Z_{ij}^T Z_{ij}\right\|\right) \leq \sigma_{\text{MAX}} \text{ for all } i,j.$$



(a) Environment covariates

(b) Species covariates

Figure 4: Correlation between the two first dimensions of interaction and the covariates (the covariates are not used in the estimation).

Moreover, note that

$$\nabla\Phi_Y = \frac{1}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} Z_{ij}.$$

Using Assumption 2, Klopp [2014], Proposition 11, and

$$\frac{M}{m_1 m_2}\sigma_{\text{MIN}}^2 \leq \sigma_Z^2 \leq \frac{M}{m_1 m_2}\sigma_{\text{MAX}}^2,$$

ensure that there exists a constant c such that with probability at least $1-(m_1+m_2)^{-1}$,

$$\|\nabla\Phi_Y(X^*)\| \leq c\max\left\{\sigma_{\text{MAX}}\frac{(2M\log(m_1+m_2))^{1/2}}{m_1 m_2}, \delta\left(\log m^{1/2}\frac{\delta}{\sigma_{\text{MIN}}}\right)\frac{2\log(m_1+m_2)}{m_1 m_2}\right\}. \quad (13)$$

The condition

$$m_1+m_2 \geq \max\left\{\frac{\delta^2}{2\sigma_{\text{MAX}}\sigma_{\text{MIN}}^2}, (4\frac{\delta^2}{\sigma_{\text{MAX}}^2})^4\right\}$$

ensures that the left term dominates. Then, taking

$$\lambda = 2c_\delta\sigma_{\text{MAX}}\frac{(2M\log(m_1+m_2))^{1/2}}{m_1 m_2}$$

and plugging this value in (11) of Proposition 4 directly gives the result.

8.2 Proof of Proposition 4

We start with some notations and preparatory lemmas. Given a matrix $X \in \mathbb{R}^{m_1\times m_2}$, we denote $\mathcal{S}_1(X)$ (resp. $\mathcal{S}_2(X)$) the span of left (resp. right) singular vectors of X. Let $P_{\mathcal{S}_1(X)}^{\perp}$ (resp. $P_{\mathcal{S}_2(X)}^{\perp}$) be the orthogonal projector on $\mathcal{S}_1(X)^{\perp}$ (resp. $\mathcal{S}_2(X)^{\perp}$). We define the projection operator $\mathcal{P}_X^{\perp}: \tilde{X} \mapsto P_{\mathcal{S}_1(X)}^{\perp}\tilde{X}P_{\mathcal{S}_2(X)}^{\perp}$, and $\mathcal{P}_X : \tilde{X} \mapsto \tilde{X} - P_{\mathcal{S}_1(X)}^{\perp}\tilde{X}P_{\mathcal{S}_2(X)}^{\perp}$.

Lemma 5. *For $X \in \mathbb{R}^{m_1\times m_2}$ and $\Theta = \mathcal{T}(X) \in \mathcal{V}^{\perp}$,*

(i) $\left\|\Theta^* + \mathcal{P}_{\Theta^*}^{\perp}(\Theta^*)\right\|_* = \|\Theta^*\|_* + \left\|\mathcal{P}_{\Theta^*}^{\perp}(\Theta^*)\right\|_*,$

(ii) $\|\Theta^*\|_* - \|\Theta\|_* \leq \|\mathcal{P}_{\Theta^*}(\Theta-\Theta^*)\|_* - \left\|\mathcal{P}_{\Theta^*}^{\perp}(\Theta-\Theta^*)\right\|_*,$

(iii) $\|\mathcal{P}_{\Theta^*}(\Theta-\Theta^*)\|_* \leq \sqrt{2\text{rk}(\Theta^*)}\,\|X-X^*\|_F.$

Proof. By definition of $\mathcal{P}_{\Theta^*}$ the singular vector spaces of Θ^* and of $\mathcal{P}_{\Theta^*}^{\perp}(\Theta^*)$ are orthogonal:

$$\left\|\Theta^* + \mathcal{P}_{\Theta^*}^{\perp}(\Theta^*)\right\|_* = \|\Theta^*\|_* + \left\|\mathcal{P}_{\Theta^*}^{\perp}(\Theta^*)\right\|_*,$$

which proves (i). Writing $\Theta = \Theta^* + \mathcal{P}_{\Theta^*}^{\perp}(\Theta-\Theta^*) + \mathcal{P}_{\Theta^*}(\Theta-\Theta^*)$ we get

$$\|\Theta\|_* \quad \geq \|\Theta^*\|_* + \left\|\mathcal{P}_{\Theta^*}^{\perp}(\Theta-\Theta^*)\right\|_* - \|\mathcal{P}_{\Theta^*}(\Theta-\Theta^*)\|_*.$$

Then, the triangular inequality and the orthonormality of the left and right singular vector spaces of Θ^* and $\mathcal{P}_{\Theta^*}^{\perp}(\Theta-\Theta^*)$ yield

$$\|\Theta\|_* - \|\Theta^*\|_* \leq \|\mathcal{P}_{\Theta^*}(\Theta-\Theta^*)\|_* - \left\|\mathcal{P}_{\Theta^*}^{\perp}(\Theta-\Theta^*)\right\|_*,$$

which gives (ii). For all $X \in \mathbb{R}^{m_1 \times m_2}$, $\mathcal{P}_{\Theta^*}(\Theta) = P_{S_1(\Theta^*)}(\Theta - \Theta^*)P^{\perp}_{S_2(\Theta^*)} + (\Theta - \Theta^*)P_{S_2(\Theta^*)}$ implies that $\mathrm{rk}(\mathcal{P}_{\Theta^*}(\Theta - \Theta^*)) \leq 2\mathrm{rk}(\Theta^*)$. This and the Cauchy-Schwarz inequality give

$$\begin{aligned} \|\mathcal{P}_{\Theta^*}(\Theta - \Theta^*)\|_* &\leq \sqrt{2\mathrm{rk}(\Theta^*)}\,\|\Theta - \Theta^*\|_F \\ &\leq \sqrt{2\mathrm{rk}(\Theta^*)}\,\|X - X^*\|_F\,, \end{aligned}$$

which finally proves (iii). □

Lemma 6. *Assume $\lambda > 2\,\|\nabla\Phi_Y(X^*\|$. Then,*

$$\left\|P^{\perp}_{\Theta^*}(\Theta^* - \hat{\Theta})\right\|_* \leq 3\left\|P_{\Theta^*}(\Theta^* - \hat{\Theta})\right\|_* + \|R\alpha^* - R\hat{\alpha}\|_* + \left\|(C\beta^* - C\hat{\beta})^T\right\|_*.$$

Proof. The result stems from the convexity of Φ_Y and Lemma 5(ii). □

On the one hand, Assumption 1 ensures the strong convexity of Φ_Y with constant $\sigma^2_{\text{MIN}}/m_1m_2$ and implies

$$\frac{\sigma^2_{\text{MIN}}\left\|X^* - \hat{X}\right\|^2_F}{2m_1m_2} \leq \Phi_Y(\hat{X}) - \Phi_Y(X^*) - \langle\nabla\Phi_Y(X^*), \hat{X} - X^*\rangle.$$

On the other hand by definition of the estimator $\hat{X}$,

$$\Phi_Y(\hat{X}) - \Phi_Y(X^*) \leq \lambda\left(\|\Theta^*\|_* - \left\|\hat{\Theta}\right\|_*\right).$$

Substracting $\langle\nabla\Phi_Y(X^*), \hat{X} - X^*\rangle$ on both side and in conjunction with the strong convexity inequality we obtain

$$\frac{\sigma^2_{\text{MIN}}\left\|X^* - \hat{X}\right\|^2_F}{2m_1m_2} \leq -\langle\nabla\Phi_Y(X^*), \hat{X} - X^*\rangle + \lambda\left(\|\Theta^*\|_* - \left\|\hat{\Theta}\right\|_*\right). \tag{14}$$

We now bound separately the two terms on the right hand side. First, the duality of $\|\cdot\|_*$ and $\|\cdot\|$ and the triangular inequality give

$$\begin{aligned} -\langle\nabla\Phi_Y(X^*), \hat{X} - X^*\rangle \leq\ & \|\nabla\Phi_Y(X^*)\| \times \\ & \left(\left\|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\right\|_* + \left\|\mathcal{P}^{\perp}_{\Theta^*}(\hat{\Theta} - \Theta^*)\right\|_* + \|R\hat{\alpha} - R\alpha^*\|_* + \left\|(C\hat{\beta} - C\beta^*)^T\right\|_*\right). \end{aligned} \tag{15}$$

Then, Lemma 5 (ii) applied to $\hat{X}$, results in

$$\|\Theta^*\|_* - \left\|\hat{\Theta}\right\|_* \leq \left\|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\right\|_* - \left\|\mathcal{P}^{\perp}_{\Theta^*}(\hat{\Theta} - \Theta^*)\right\|_*. \tag{16}$$

Plugging inequalities (15) and (16) in (14), and using the condition $\lambda \geq 2\,\|\nabla\Phi_Y(X^*)\|$, we finally obtain

$$\begin{aligned} \frac{\sigma^2_{\text{MIN}}\left\|X^* - \hat{X}\right\|^2}{m_1m_2} \quad & \\ \leq 3\lambda\left\|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\right\|_* + \lambda\left(\|R\hat{\alpha} - R\alpha^*\|_* + \left\|(C\hat{\beta} - C\beta^*)^T\right\|_*\right). & \end{aligned} \tag{17}$$

Then, $\mathrm{rk}(R\hat{\alpha} - R\alpha^*) \leq K_1$, $\mathrm{rk}(C\hat{\beta} - C\beta^*) \leq K_2$ implies $\|R\hat{\alpha} - R\alpha^*\|_* + \left\|(C\hat{\beta} - C\beta^*)'\right\|_* \leq \left(\sqrt{K_1} + \sqrt{K_2}\right)\left\|X^* - \hat{X}\right\|_F$, which together with Lemma 5 (iii) and $2(a^2 + b^2) \geq (a + b)^2$ yield Proposition 4.

8.3 Proof of Theorem 2

We assume without loss of generality that $m_1 \geq m_2$. Recall $N = r(m_1 - K_1) + K_1 m_2 + K_2 m_1 - K_1 K_2$ and denote $\eta \in (0, 1/8)$. We define

$$\kappa = \min\left(\frac{1}{2}, \frac{\eta^{1/2} N^{1/2}}{2\gamma \sigma_{\text{MAX}} (m_1 m_2)^{1/2}}\right), \quad \gamma = \min\{\gamma_{\text{MIN}}, \gamma_{\text{MAX}}\}. \tag{18}$$

Let $\mathcal{B}_R = (u_1, \ldots, u_{K_1}, \ldots, u_{m_1})$ be an orthonormal basis of $\mathbb{R}^{m_1}$ such that $(u_1, \ldots, u_{K_1})$ is an orthonormal basis of the range of R, and $\mathcal{B}_C = (v_1, \ldots, v_{K_2}, \ldots, v_{m_2})$ an orthonormal basis of $\mathbb{R}^{m_2}$ such that $(v_1, \ldots, v_{K_2})$ is an orthonormal basis of the range of C. With these notations, $\left(u_i v_j^T\right)_{i,j}$, $i = 1, \ldots, m_1$ and $j = 1, \ldots, m_2$ is an orthonormal basis of $\mathbb{R}^{m_1 \times m_2}$ and any matrix $X \in \mathbb{R}^{m_1 \times m_2}$ can be decomposed as follows:

$$X = \underbrace{\sum_{1 \leq k \leq K_1} \sum_{1 \leq j \leq m_2} t_{kj} u_k v_j^T}_{R\alpha} + \underbrace{\sum_{K_1+1 \leq i \leq m_1} \sum_{1 \leq k \leq K_2} z_{ik} u_i v_k^T}_{(C\beta)^T} + \underbrace{\sum_{K_1+1 \leq k \leq m_1} \sum_{K_2+1 \leq \ell \leq m_2} w_{k\ell} u_k v_\ell^T}_{\Theta}.$$

We now define the following set of matrices for $r \geq 1$:

$$\begin{aligned} \mathcal{L} = \Big\{ L \in \mathbb{R}^{m_1 \times m_2} \mid \quad & (t_{kj}) \in \{0, \kappa\gamma\}, k = 1, \ldots, K_1, \quad j = 1, \ldots, m_2; \\ & (z_{i\ell}) \in \{0, \kappa\gamma\}, i = K_1 + 1, \ldots, m_1, \quad \ell = 1, \ldots K_2; \\ & (w_{k\ell}) \in \{0, \kappa\gamma\}, k = K_1 + 1, \ldots, m_1, \quad \ell = K_2 + 1, \ldots, K_2 + r \Big\}. \end{aligned} \tag{19}$$

We also define the set $\tilde{\mathcal{L}}$ as follows. For integers n and m, denote by $n[m]$ the value of n modulo m. For every $L \in \mathcal{L}$ we define $\tilde{L} \in \tilde{\mathcal{L}}$ as:

$$\tilde{L} = \sum_{1 \leq q \leq k_1} \sum_{1 \leq j \leq m_2} t_{kj} u_k v_j^T + \sum_{1 \leq i \leq m_1} \sum_{1 \leq s \leq k_2} z_{is} u_i v_k^T + \underbrace{\sum_{K_1+1 \leq k \leq m_1} \sum_{K_2+1 \leq \ell \leq m_2} \tilde{w}_{k\ell} u_k v_\ell^T}_{\tilde{\Theta}},$$

where for $k = K_1 + 1, \ldots, m_1$ and with κ and γ defined in (18) we have set

$$\tilde{w}_{k\ell} = w_{k\ell[r]}, \quad \ell = K_2 + 1, \ldots, m_2.$$

For all $\tilde{L} \in \tilde{\mathcal{L}}$, the corresponding $\tilde{\Theta}$ is of rank at most r, and this is also true for the difference between any two elements of $\tilde{\mathcal{L}}$. The Varshamov-Gilbert bound [Tsybakov, 2008, Lemma 2.9] guarantees that there exists a subset $\mathcal{A} \subset \tilde{\mathcal{L}}$ of cardinality

$$\text{Card}(\mathcal{A}) \geq 2^{N/8}$$

containing the null matrix with zero in all entries, such that for any two elements X^1, X^2 of $\mathcal{A}$,

$$\begin{aligned} \left\|X^1 - X^2\right\|_F^2 &\geq \frac{(r(m_1 - K_1)(m_2 - K_2)/r + K_1 m_2 + K_2 m_1 - K_1 K_2)\, \kappa^2 \gamma^2}{8} \\ &\geq \frac{(m_1 m_2)\, \kappa^2 \gamma^2}{8}. \end{aligned} \tag{20}$$

Recall that for $X \in \mathbb{R}^{m_1 \times m_2}$ we denote by $\mathbb{P}_X$ the law of $m_1 \times m_2$ independant random Poisson variables with means $\exp(X_{ij})$. Denote by $\mathbb{P}_0$ the law of $m_1 \times m_2$ independant random Poisson variables with means 1. Let us compute the Kullback-Leibler divergence between $\mathbb{P}_X$ and $\mathbb{P}_0$:

$$\mathrm{KL}(\mathbb{P} \,||\, \mathbb{P}_0) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_1} \left(\exp(X_{ij}) - 1 - X_{ij}\right).$$

Using the fact that $X_{ij} = 0$ or $X_{ij} = \kappa\gamma$, that $x \mapsto \exp(x) - 1 - x$ is gradient Lipschitz on $(0, \gamma)$ with constant σ_{MAX}, and the definition of κ, we obtain

$$KL(\mathbb{P}_X \,||\, \mathbb{P}_0) \leq \frac{\eta N}{8} \leq \eta \log_2(\mathrm{Card}(\mathcal{A})),$$

and thus:

$$\frac{1}{\mathrm{Card}(\mathcal{A})} \sum_{X \in \mathcal{A}} \mathrm{KL}(\mathbb{P}_X \,||\, \mathbb{P}_0) \leq \eta \log_2(\mathrm{Card}(\mathcal{A})). \tag{21}$$

Equations (20) and (21) guarantee that we can use [Tsybakov, 2008, Theorem 2.5], which gives that

$$\inf_{\hat{X}} \sup_{X \in \mathcal{F}(r,\gamma)} \mathbb{P}_X \left(\frac{\left\| \hat{X} - X \right\|_F^2}{m_1 m_2} > \min\left\{ \gamma^2, \frac{\eta N}{m_1 m_2 \sigma_{\text{MAX}}^2} \right\} \right) \geq d(\eta, m_1), \tag{22}$$

where

$$d(\eta, m_1) = \frac{1}{1 + 2^{-N/16}} \left(1 - 2\eta - \frac{1}{2} \frac{\eta^{1/2}}{(N \log(2))^{1/2}} \right).$$

8.4 Proof of Proposition 3

We start by some preparatory lemmas and notations. Denote $\mathcal{K}_{\mathcal{T}}$ the image of $\mathcal{K}$ by projector $\mathcal{T}$. For some $X \in \mathcal{K}$, let $f_X : \mathcal{K}_{\mathcal{T}} \to \mathbb{R}_+$ be the function such that $f_X(A) = \mathbb{1}_{\mathcal{K}}(X + A)$. Let $g : \mathcal{V}^{\perp} \to \mathbb{R}_+$ be the function defined by $g(A) = \|A\|_*$ for $A \in \mathcal{V}^{\perp}$.

Lemma 7. $0 \in \partial f_X(A) \mid_{A=0}$.

Proof. $X \in \mathcal{K}$ implies that $f(0) = 0$, moreover, for all $B \in \mathcal{K}_{\mathcal{T}}$, $f(0) + \langle 0, (B - A) \rangle = 0$. By definition of the subdifferential we only need to prove that $f(B) \geq 0$ for all $B \in \mathcal{K}_{\mathcal{T}}$, which is straightforward with the definition of f. □

Lemma 8. $\partial g(0) = \{W \in \mathbb{R}^{m_1 \times m_2}, \|\mathcal{T}(W)\| < 1\}$.

Proof. By definition of the subdifferential we need to prove that for all $W \in \mathbb{R}^{m_1 \times m_2}$, $\|\mathcal{T}(W)\| < 1$, and for all $B \in \mathcal{V}^{\perp}$, $g(B) \geq g(0) + \langle W, B - 0 \rangle$. First $B \in \mathcal{V}^{\perp}$ implies $\langle W, B \rangle = \langle \mathcal{T}(W), B \rangle$, therefore $\|\mathcal{T}(W)\| \leq 1$ is a sufficient condition for $W \in \partial g(0)$. Now assume $\|\mathcal{T}(W)\| > 1$ and let $\mathcal{T}(W) = U \Sigma V^T$, where U and V are orthogonal matrices of left and right singular vectors, and $\Sigma_{11} = \|\mathcal{T}(W)\| > 1$. Let us define $B = U \tilde{\Sigma} V^T$, $\tilde{\Sigma}_{11} = 1$ and $\tilde{\Sigma}_{ij} = 0$ elsewhere; not that with this definition $B \in \mathcal{V}^{\perp}$. We have $g(B) = 1$ and $\langle \mathcal{T}(W), B \rangle = \Sigma_{11} > g(B)$. Therefore $\|\mathcal{T}(W)\| > 1 \Rightarrow W \notin \partial g(0)$, from which we conclude

$$\partial g(0) = \left\{ W \in \mathbb{R}^{m_1 \times m_2}, \|\mathcal{T}(W)\| < 1 \right\}.$$

□

We now proceed to the proof of Proposition 3. In what follows for some $X \in \mathbb{R}$, we write with a small abuse of notations $\Theta = \mathcal{T}(X)$ and $X_0 = X - \mathcal{T}(X)$. We define $\Phi_Y^\lambda(X_0, \Theta) = \Phi_Y(X_0, \Theta) + \lambda \|\Theta\|_*$. The zero thresholding statistic is formally defined by

$$\lambda_0(Y) = \min_\lambda \quad 0 \in \partial\{\Phi_Y^\lambda(X_0, \Theta) + \mathbb{1}_\mathcal{K}(X_0 + \Theta)\} \mid_{\Theta=0},$$

where $\mathbb{1}_\mathcal{K}(X_0 + \Theta)$ is the indicator function of $\mathcal{K}$, equal to 0 on $\mathcal{K}$ and $+\infty$ elsewhere. Under the constraint $\Theta = 0$, we get

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

while the subdifferential of the objective function Φ_Y^λ with respect to Θ at $\Theta = 0$ is given by

$$\partial_\Theta \Phi_Y^\lambda \mid_{\Theta=0} = -\frac{1}{m_1 m_2}\left(Y - \exp(X_0)\right) + \lambda \partial_\Theta \|\Theta\|_* \mid_{\Theta=0} + \partial_\Theta \mathbb{1}_\mathcal{K}(X_0 + \Theta) \mid_{\Theta=0}.$$

Lemma 7 guarantees that $0 \in \partial_\Theta \mathbb{1}_\mathcal{K}(X_0 + \Theta) \mid_{\Theta=0}$, and Lemma 8 ensures that $0 \in \partial \Phi_Y^\lambda(\Theta) \mid_{\Theta=0}$ if and only if

$$0 \in -\frac{1}{m_1 m_2}(Y - \exp(\hat{X}_0)) + \lambda W, \ \|\mathcal{T}(W)\| < 1.$$

This is equivalent to $\lambda \geq (m_1 m_2)^{-1} \left\|\mathcal{T}(Y - \exp(\hat{X}_0))\right\|$. Additionally, at the optimum $\hat{X}_0$, we have $\mathcal{T}(Y - \exp(\hat{X}_0)) = Y - \exp(\hat{X}_0)$, which concludes the proof.

Acknowledgements

The authors thank Trevor Hastie, Edgar Dobriban, Olga Klopp and Kevin Bleakey for their very helpful comments.

References

A. Agresti. *Categorical Data Analysis, 3rd Edition*. Wiley, 2013.

J. Bigot, C. Deledalle, and D. Féral. Generalized sure for optimal shrinkage of singular values in low-rank matrix denoising. *arXiv:1605.07412*, 2016.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–22, 2011.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Yang Cao and Yao Xie. Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing*, 64(6), March 2016.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011. ISSN 0924-9907. doi: 10.1007/s10851-010-0251-1. URL http://dx.doi.org/10.1007/s10851-010-0251-1.

Philippe Choler. Consistent shifts in alpine plant traits along a mesotopographical gradient. *Arctic, Antarctic, and Alpine Research*, 37(4):444–453, 1 2005. doi: 10.1214/12-AOS986. URL http://dx.doi.org/10.1657/1523-0430(2005)037[0444:CSIAPT]2.0.CO;2.

R. Christensen. *Log-Linear Models*. Springer-Verlag, New York., 2010.

M. Collins, S. Dasgupta, and R.E. Schapire. A generalization of principal component analysis to the exponential family. In *Advances in Neural Information Processing Systems*. MIT Press, 2001.

M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-Bit Matrix Completion. *ArXiv e-prints*, September 2012.

A. de Falguerolles. Log-bilinear biplot in action. In J. Blasius and M. . Greenacre, editors, *Visualisation of categorical data*, pages 527–533. Academic Press, 1998.

Jan de Leeuw. Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data Analysis*, 50(1):21–39, 2006.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

Bradley Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26, 1979.

Andrey Feuerverger, Yu He, and Shashi Khatri. Statistical significance of the netflix challenge. *Statist. Sci.*, 27(2):202–231, 05 2012. doi: 10.1214/11-STS368. URL http://dx.doi.org/10.1214/11-STS368.

Mário A. T. Figueiredo and José M. Bioucas-Dias. Restoration of poissonian images using alternating direction optimization. *Trans. Img. Proc.*, 19(12):3133–3145, December 2010. ISSN 1057-7149. doi: 10.1109/TIP.2010.2053941. URL http://dx.doi.org/10.1109/TIP.2010.2053941.

William Fithian and Julie Josse. Multiple correspondence analysis & the multilogit bilinear model. *Journal of Multivariate Analysis*, 2017.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 12 2007. doi: 10.1214/07-AOAS131. URL http://dx.doi.org/10.1214/07-AOAS131.

M. Gavish and D. L. Donoho. Optimal shrinkage of singular values. *arXiv:1405.7511*, 2014a.

Matan Gavish and David L Donoho. The optimal hard threshold for singular values is 4/sqrt (3). *IEEE Transactions on Information Theory*, 60(8), 2014b.

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

Roland Glowinski and Americo Marrocco. Sur l'approximation, par éléments finis d'ordre 1, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires. *C. R. Acad. Sci. Paris Sér. A*, 278:1649–1652, 1974.

L. A. Goodman. The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Annals of Statistics*, 13:10–69, 1985.

P. Gopalan, F.J.R. Ruiz, R. Ranganath, and D.M Blei. Bayesian nonparametric poisson factorization for recommendation systems. *In AISTATS*, pages 275–283, 2014.

J. Gower, S. Lubbe, and N. le Roux. *Understanding Biplots*. John Wiley & Sons, 2011.

T. Hastie, R. Mazumder, J. Lee, and R. Zadeh. Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *ArXiv e-prints*, October 2014.

T. Hastie, R. Mazumder, J. Lee, and R. Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *Journal in Machine Learning Research*, 2015.

T Jeong, H Woo, and S Yun. Frame-based poisson image restoration using a proximal linearized alternating direction method. *Inverse Problems*, 29(7):075007, 2013. URL http://stacks.iop.org/0266-5611/29/i=7/a=075007.

Julie Josse and Sylvain Sardy. Adaptive shrinkage of singular values. *Statistics and Computing*, pages 1–10, 2015.

Julie Josse and Stefan Wager. Bootstrap-based regularization for low-rank matrix estimation. *Journal of Machine Learning Research*, 17(124):1–29, 2016.

Maria Kateri. *Contingency Table Analysis*. Springer New York, 2014.

Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

Jean Lafond. Low rank matrix completion with exponential family noise. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 40:1–18, 2015.

J. Li and D. Tao. Simple exponential family PCA. *IEEE Transactions on Neural Networks and Learning Systems*, 24(3):485–497, 2013.

Roderick. J. A. Little and Donald. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons series in probability and statistics, New-York, 1987, 2002.

L.T. Liu, E. Dobriban, and A. Singer. epca: High dimensional exponential family pca. *arXiv:1611.05550*, 2016.

S. Mohamed, Z. Ghahramani, and K. A. Heller. Bayesian exponential family pca. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1089–1096. 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

M. Raginsky, R. M. Willett, Z. T. Harmany, and R. F. Marcia. Compressed sensing performance bounds under poisson noise. *IEEE Transactions on Signal Processing*, 58(8):3990–4002, Aug 2010. ISSN 1053-587X. doi: 10.1109/TSP.2010.2049997.

J. Salmon, Z. Harmany, C.A. Deledalle, and R. Willett. Poisson noise reduction with non-local pca. *Journal of Mathematical Imaging and Vision*, 48(2):279–294, 2014.

Andrey A Shabalin and Andrew B Nobel. Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis*, 118:67–76, 2013.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *arXiv preprint arXiv:1410.0342*, 2014.