Low-rank model with covariates for count data analysis

Geneviève Robin, Julie Josse, Éric Moulines, Sylvain Sardy

Low-rank Interaction Contingency Tables with Covariates

Geneviève Robin[1], Julie Josse[1], Éric Moulines[1]

Sylvain Sardy[2]

[1]Center of Applied Mathematics, École Polytechnique, XPOP INRIA
[2]Department of Mathematics, Université de Genève

March 20, 2018

Abstract

Contingency tables are collected in many scientific and engineering tasks including image processing, single-cell RNA sequencing and ecological studies. Low-rank methods are extensively used to analyse them, in particular for estimation purposes. However, common estimation methods based on probabilistic models 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 negative 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. Our results can be straightforwardly extended to general exponential family models. A complete methodology is proposed, including an algorithm based on the alternating direction method of multipliers, and an automatic selection of the regularization parameter. The method can also be applied when the table contains missing values. The simulation study reveals that our estimator compares favourably to competitors. Then, analysing two ecological data sets, we show how to interpret the model using graphical tools. The method is available in the R package `lori`.

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

1 Introduction

1.1 Model and contributions

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-bilinear 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, is commonly used to describe the structure of the matrix X^* and defined by

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

where $\text{rk}(\Theta^*)$ denotes the rank of Θ^* and $K \leq \min(m_1 - 1, m_2 - 1)$. In these models, μ^* is an offset, the terms which only depend on the index of the row or column (α^*_i and β^*_j) are called *main effects*, and the terms which depend on both (here Θ^*_{ij}) are called *interactions* [Kateri, 2014, Section 4.1.2, p.87].

Our first contribution is to introduce an extension of the *row-column model* (1) by incorporating general covariates and interactions between them, as well as residual interaction terms. More formally, let $R \in \mathbb{R}^{m_1 \times K_1}$ and $C \in \mathbb{R}^{m_2 \times K_2}$ be matrices of known row and column covariates respectively, and $\mu^* \in \mathbb{R}^{K_1 \times K_2}$, $\alpha^* \in \mathbb{R}^{K_2 \times m_1}$ and $\beta^* \in \mathbb{R}^{K_1 \times m_2}$ matrices of unknown parameters. We model the matrix X^* as follows:

$$X^* = R\mu^* C^\top + {\alpha^*}^\top C^\top + R\beta^* + \Theta^*, \tag{2}$$

with $C^\top$ denoting the transpose of matrix C. In the ecology example of Section 5, 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 the column features C code physical traits about species like height or mass. In model (2), $R\mu^* C^\top$ incorporates *interactions* between covariates, $R\beta^*$ (resp. $C\alpha^*$) contains *interactions* between environment covariates and species (resp. species covariates and environments). Although they look like main effects, these two terms indeed correspond to *interactions* since the linear combination of the environments covariates (the main effect) is different for every species, *i.e*

$$(R\beta^*)_{ij} = \sum_{k=1}^{K_1} R_{ik}\beta^*_{kj}.$$

Lastly, Θ^* corresponds to the interactions unexplained by the known covariates R and C; by a small abuse of terminology, we will refer to Θ^* as the *interaction matrix*. The classical log-bilinear model (1) is a special case of our model using the coding

$$R = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \updownarrow m_1 \ , C = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \updownarrow m_2 \ ,$$

$$X = \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix} \mu \quad (1 \quad \dots \quad 1) + \begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_{m_1} \end{pmatrix} \quad (1 \quad \dots \quad 1) + \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix} \quad (\beta_1 \quad \dots \quad \beta_{m_2}) + \Theta.$$

The paper is organized as follows. After discussing related works in Section 1.2, we define in Section 2 an estimator for model (2), through the minimization of

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

$$\Phi_Y(X) = -\frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \left(Y_{ij} X_{ij} - \exp(X_{ij}) \right), \tag{3}$$

penalized by the nuclear norm of the matrix Θ, which acts as a convex relaxation of the rank constraint. In Section 2.2, another contribution is to derive an upper bound for the estimation risk that holds for a number of generalized linear models, and a minimax lower bound. In Section 3, we propose an optimization algorithm based on the *alternating descent method of multipliers* [Boyd et al., 2011], and two methods to choose the regularization parameter automatically. We show in Section 4 on simulated data that our procedure compares favourably to competitors, and highlight the interpretability of the method with two applications in ecology in Sections 5 and 6. The proofs are given in Section 8. The methods is available as an R package [R Core Team, 2016] called `lori` (Low-rank Interaction) on the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=lori. The code for the experiments is also publicly available at https://github.com/genevievelrobin/lori.

1.2 Related work

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]. Poisson matrix estimation has also been considered via singular value shrinkage, extending the Gaussian setting [Shabalin and Nobel, 2013, Gavish and Donoho, 2017, 2014, Josse and Sardy, 2016]. Bigot et al. [2017] have studied optimal singular value shrinkage in the exponential family, while Liu et al. [2016] have suggested a new shrinkage for covariance matrix estimation.

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., 2014]. Estimation rates are also given in Lafond [2015], where optimal bounds are proved for matrix completion in the exponential family. None of the methods in this body of literature account for available covariates.

Note also the work of Fithian and Mazumder [2013], which presents a variety of low-rank problems including the generalized nuclear norm penalty [Angst et al., 2011], that can be used to include row and column covariates. However the authors focus on scalable estimation algorithms and provide neither statistical guarantees nor automatic selection of the regularization parameters.

Finally, our contribution has some connections with methods suggested in the statistical ecology literature to analyse contingency tables with row and column covariates. Brown et al. [2014] and ter Braak et al. [2017] suggested the following

model

$$X^*_{ij} = \mu^* + \alpha^*_i + \beta^*_j + \epsilon_{RC} R_i C_j, \tag{4}$$

with R_i, $1 \leq i \leq m_1$ a row trait and C_j, $1 \leq j \leq m_1$ a column trait. The interaction between covariates modelled by $\epsilon_{RC} R_i C_j$, where ϵ_{RC} is an unknown parameter measuring the strength of the interaction between the two traits. The extension to the case where several covariates are present is straightforward. Note that the interaction term $\epsilon_{RC} R_i C_j$ corresponds for row i, column j and row and column covariates k and l to the interaction term $R_{ik}\mu_{kl}C_{jl}$ in our model (2). Consequently, the first difference is that our model is more complex and allows modelling of interactions between species and environments which might be caused by unmeasured traits, through the matrix Θ^*. Second, model (4) was developed with the aim of testing significant associations between covariates. On the contrary we study the interaction between species and environments after discarding any effect (main effect and interaction) of the covariates. In Section 3, we introduce a method to test for the existence of such interactions. Finally, to the best of our knowledge the theoretical properties in terms of estimation of the models derived in Brown et al. [2014] and ter Braak et al. [2017] have not yet been studied. We can also mention the "fourth-corner" [Legendre et al., 1997] and RLQ [Dolédec et al., 1996] methods, which also aim at testing the associations between covariates. Contrary to the method we develop here, they are both defined without referring to a probabilistic framework.

2 Estimator and theoretical results

2.1 Notation and estimator

Along this article we will denote, for $A \in \mathbb{R}^{m_1 \times m_2}$, $\|A\|_*$ the sum of the singular values of A (the nuclear norm), $\|A\|_F$ the Frobenius norm, $\|A\|$ the largest singular value (the operator norm), and $\|A\|_\infty$ the largest entry in absolute value. We also use the following notation related to the covariates R and C:

$$\begin{aligned} \mathcal{V}_R^\perp &= \left\{A \in \mathbb{R}^{K_2 \times m_1}; AR = 0\right\}, \\ \mathcal{V}_C^\perp &= \left\{A \in \mathbb{R}^{K_1 \times m_2}; AC = 0\right\}, \\ \mathcal{V}^\perp &= \left\{A \in \mathbb{R}^{m_1 \times m_2}; A^\top R = 0, AC = 0\right\}. \end{aligned} \tag{5}$$

Consider now the following assumption.

H 1. *There exist* $\gamma_{min} > -\infty$ *and* $\gamma_{max} < \infty$ *such that for all* $i = 1, \ldots, m_1$ *and* $j = 1, \ldots, m_2$,

$$\gamma_{min} \leq \log E(Y_{ij}) \leq \gamma_{max}.$$

Moreover there exist $\sigma_{min} > 0$ *and* $\sigma_{max} < \infty$ *such that for all* $i = 1, \ldots, m_1$ *and* $j = 1, \ldots, m_2$,

$$\sigma^2_{min} \leq \mathrm{var}(Y_{ij}) \leq \sigma^2_{max}.$$

In Assumption 1, the bounds on the expectations γ_{min} and γ_{max} guarantee the existence of a solution to the convex program (7) defined below, and the convergence of the algorithm; they are therefore present in our implementation. On the contrary, constants σ_{min} and σ_{max} are only required to guarantee the theoretical results.

Consider the compact set $\mathcal{K} = [\gamma_{\min}, \gamma_{\max}]^{m_1 \times m_2}$. We define our estimator, for a given regularization parameter $\lambda > 0$, as the minimizer of the following penalized negative log-likelihood, with $\phi_Y(\mu, \alpha, \beta, \Theta) = \Phi_Y(R\mu C^\top + \alpha^\top C^\top + R\beta + \Theta)$, and Φ_Y is the negative log-likelihood defined in (3).

$$\begin{aligned} (\tilde{\mu}^\lambda, \tilde{\alpha}^\lambda, \tilde{\beta}^\lambda, \tilde{\Theta}^\lambda) \quad &= \operatorname{argmin} \quad \phi_Y(\mu, \alpha, \beta, \Theta) + \lambda \left\| \Theta \right\|_*, \\ \text{such that} \quad & R\mu C^\top + \alpha^\top C^\top + R\beta + \Theta \in \mathcal{K}, \\ & \alpha \in \mathcal{V}_R^\perp, \beta \in \mathcal{V}_C^\perp, \Theta \in \mathcal{V}^\perp, \end{aligned} \tag{6}$$

where $\alpha \in \mathcal{V}_C^\perp$, $\beta \in \mathcal{V}_R^\perp$ and $\Theta \in \mathcal{V}^\perp$ are identifiability constraints, with $\mathcal{V}_C^\perp$, $\mathcal{V}_R^\perp$ and $\mathcal{V}^\perp$ defined in (5). The problem is neither jointly convex nor separable in μ α, β and Θ. We first re-parametrize (6), which allows us to derive theoretical results, while simplifying the optimization (see Section 3). Let Π_R and Π_C be the orthogonal projection matrices on the linear span of the columns of R and C respectively. Let

$$\mathcal{T} : X \in \mathbb{R}^{m_1 \times m_2} \mapsto (I - \Pi_R) X (I - \Pi_C)^\top$$

be the projection operator on the subspace $\mathcal{V}^\perp$. Consider the reformulated problem

$$\hat{X}^\lambda, \hat{\Theta}^\lambda = \underset{\substack{X \in \mathcal{K} \\ \Theta = \mathcal{T}(X)}}{\operatorname{argmin}} \quad \Phi_Y(X) + \lambda \left\| \Theta \right\|_*, \tag{7}$$

Program (7) yields the same solution in Θ and X as problem (6): $\hat{\Theta}^\lambda = \tilde{\Theta}^\lambda$ and $\hat{X}^\lambda = \tilde{\mu}^\lambda + \tilde{\alpha}^{\lambda\top} C^\top + R\tilde{\beta}^\lambda + \tilde{\Theta}^\lambda$. Moreover, the identifiability constraint $\mathcal{T}(X) = \Theta$ ensures that we can compute $R\hat{\mu}^\lambda C^\top, \hat{\alpha}^{\lambda\top} C^\top$ and $R\hat{\beta}^\lambda$ a posteriori based on $\hat{X}^\lambda$ and $\hat{\Theta}^\lambda$ only, by applying simple projections. The parameters $\hat{\mu}^\lambda, \hat{\alpha}^\lambda$ and $\hat{\beta}^\lambda$ can also be recovered, whenever R and C have full rank, *i.e.* $R^\top R$ and $C^\top C$ are invertible.
Problem (7) 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 $\mathcal{K}$. These two properties guarantee the existence and uniqueness of the solution of (7).

Note that estimator (7) 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

$$\hat{X}^\lambda = \operatorname{argmin}_X \quad L(X; Y) + \lambda \left\| X \right\|_*,$$

where L is a loss function. The main difference with (7) 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.

2.2 Statistical guarantees

We now derive an upper bound on the Frobenius estimation error of estimator (7). Denote $M = \max(m_1, m_2)$ and $m = \min(m_1, m_2)$.

H 2 (Sub-exponentiality). *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*

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

Let Assumptions 1 and 2 hold, and $m_1+m_2 \geq \max\left\{\delta^2(2\sigma_{max}^2\sigma_{min}^2)^{-1}, (4\delta^2/\sigma_{max}^2)^4\right\}$. Then with probability at least $1-(m_1+m_2)^{-1}$,

$$\frac{\left\|X^* - \hat{X}^\lambda\right\|_F^2}{m_1 m_2} \leq \left(\frac{4\sigma_{max}^2}{\sigma_{min}^4}\right)\frac{M\left(\mathrm{rk}(\Theta^*)+K_1+K_2\right)\log(m_1+m_2)}{m_1 m_2}. \tag{8}$$

Proof. See Section 8.1. □

The constant term appearing in bound (8) grows linearly with the upper bound σ_{max}^2 and quadratically with the inverse of σ_{min}^2. This means that by relaxing Assumption 1 to allow $\mathrm{var}(Y_{ij})$ to grow as fast as $\log(m_1+m_2)$ or decrease as fast as $1/\log(m_1+m_2)$, we only lose a log-polynomial factor in the bound. Furthermore, the explicit form of the data-fitting term Φ_Y does not appear in this result. This bound therefore holds for a number of other generalized linear models, including the binomial and exponential ones. Note also that the upper bound in Theorem 1 is parametric.

Let us now derive a lower bound on the Frobenius estimation error. Define $\gamma = \min(|\gamma_{\mathsf{min}}|, |\gamma_{\mathsf{max}}|)$, where γ_{min} and γ_{max} are defined in Assumption 1. For an integer $r \leq m$ define $\mathcal{F}(r,\gamma)$ the set of matrices

$$\mathcal{F}(r,\gamma) = \bigcup_{\substack{R\in\mathbb{R}^{m_1\times K_1}\\ C\in\mathbb{R}^{m_2\times K_2}}} \left\{X\in\mathbb{R}^{m_1\times m_2} : \mathrm{rk}\left((I-\Pi_R)X(I-\Pi_C)^\top\right)\leq r, \|X\|_\infty \leq \gamma\right\}.$$

For $X\in\mathbb{R}^{m_1\times m_2}$, denote by $\mathbb{P}_X$ the law of $m_1\times m_2$ independent random Poisson variables with means $\exp(X_{ij})$, $i=1,\dots,m_1$ and $j=1,\dots,m_2$.

Theorem 2. *There exist absolute constants $\eta>0$ and $C>0$ such that for all $m_1, m_2 \geq 2$, $1\leq r\leq m$*

$$\inf_{\hat{X}}\sup_{X\in\mathcal{F}(r,\gamma)} \mathbb{P}_X\left(\frac{\left\|\hat{X}-X\right\|_F^2}{m_1 m_2} > C\psi\frac{rM+K_1 m_1+K_2 m_2}{m_1 m_2}\right)\geq\eta,$$

where the infimum is computed over all estimators and

$$\psi = \min\left(\min(\gamma,\sigma_{max})^2, \frac{1}{\sigma_{max}^2}\right).$$

Proof. See Section 8.3. □

Comparing the upper bound of Theorem 1 and the lower bound of Theorem 2, we see that our rates are minimax optimal up to constant and logarithmic terms, whenever the dimensions m_1 and m_2 are of the same order of magnitude, *i.e.* when the ratio $\max(m_1,m_2)/\min(m_1,m_2)$ is either constant or a logarithmic in m_1+m_2.

3 Algorithm and selection of λ

3.1 Optimization algorithm

We solve (7) using the *alternating directions method of multipliers* [Glowinski and Marrocco, 1974], whose convergence stems from Boyd et al. [2011, Theorem 3.2.1]. 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

$$\mathcal{L}_\tau(X, \Theta, \Gamma) = \Phi_Y(X) + \lambda \|\Theta\|_* + \langle \Gamma, \mathcal{T}(X) - \Theta \rangle + \frac{\tau}{2} \|\mathcal{T}(X) - \Theta\|_F^2 , \qquad (9)$$

where $\langle ., . \rangle$ denotes the trace scalar product on $\mathbb{R}^{m_1 \times m_2}$. At iteration ℓ, the algorithm consists in updating separately the primal variables X, Θ, and the dual variable Γ to maximize (9) according to the following equations:

$$\begin{aligned} X^{\ell+1} &= \operatorname{argmin}_{X \in \mathcal{K}} \quad \mathcal{L}_\tau(X, \Theta^\ell, \Gamma^\ell) \\ \Theta^{\ell+1} &= \operatorname{argmin}_{\Theta \in \mathcal{K}_\mathcal{T}} \quad \mathcal{L}_\tau(X^{\ell+1}, \Theta, \Gamma^\ell) \\ \Gamma^{\ell+1} &= \Gamma^\ell + \tau(\mathcal{T}(X^{\ell+1}) - \Theta^{\ell+1}). \end{aligned} \qquad (10)$$

The function Φ_Y and $\|.\|_*$ are closed, proper and convex on $\mathbb{R}^{m_1 \times m_2}$. This ensures the resolvability of the minimization problems defined in update (10). Moreover Φ_Y is twice differentiable, so the optimization in X can be done using Newton's method. The update of Θ can itself be done in closed form and involves singular value decomposition and thresholding [Cai et al., 2010], with

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

and $\mathcal{D}_{\lambda/\tau}$ is the soft-thresholding operator of singular values at level λ/τ.
To speed up the procedure, 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 all the singular values of Θ to 0 (see Section 3); 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].
A possible substitute to the alternating direction method of multipliers is alternating minimization, which consists in minimizing the objective in (6) 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

$$\begin{aligned} \Theta^{\ell+1} \quad &= \operatorname{argmin}_\Theta \quad \phi_Y(\mu^\ell, \alpha^\ell, \beta^\ell, \Theta) + \lambda \|\Theta\|_* , \\ \text{such that} \quad & R\mu^\ell C^\top + {\alpha^\ell}^\top C^\top + R\beta^\ell + \Theta \in \mathcal{K}, \quad \Theta \in \mathcal{V}^\perp. \end{aligned}$$

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.

3.2 Automatic selection of λ

Let us now describe two methods to select the regularization parameter: cross-validation and *quantile universal threshold*. Cross-validation consists in erasing a fraction of the observed cells 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. 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

$$\left(\hat{X}^\lambda, \hat{\Theta}^\lambda\right) = \underset{\substack{X \in \mathcal{K} \\ \Theta = \mathcal{T}(X)}}{\operatorname{argmin}} \quad \Phi_{\Omega(Y)}(X) + \lambda \left\|\Theta\right\|_*, \tag{11}$$

which can be solved using the method described in Section 3.1. Repeating this procedure N times for a grid of λ, we select the value λ_{CV} that minimizes the prediction squared error. 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). Problem (11) can also be used to complete contingency tables with missing values, as shown in an ecological application in Section 6.

We suggest an alternative method to cross-validation, inspired by Donoho and Johnstone [1994] and the work of Giacobino et al. [2017] on *quantile universal threshold*. In Proposition 1 below, we define the so-called *null-thresholding statistic* of estimator (2), 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 1 in Section 8.2.

Proposition 1 (Null-thresholding statistic)**.** *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 null-thresholding statistic*

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

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

$$\mathbf{H}_0 : \Theta^* = 0 \quad \text{against the alternative} \quad \mathbf{H}_1 : \Theta^* \neq 0 \tag{12}$$

which actually boils down to testing whether the parameter matrix X^* can be explained only in terms of linear combinations of the measured covariates. For $0 < \varepsilon < 1$, consider a value λ_ε that satisfies $\mathbb{P}_{\mathbf{H}_0}\left(\lambda_0(Y) > \lambda_\varepsilon\right) < \varepsilon$. The test which consists in comparing the statistics $\lambda_0(Y)$ to λ_ε is of level $1 - \varepsilon$ for (12). This can be seen as an alternative to the χ^2 test for independence, which can furthermore handle covariates. In practice we do not have access to the distribution under the null $\mathbb{P}_{\mathbf{H}_0}\left(\lambda_0(Y) < \lambda\right)$, but perform parametric bootstrap [Efron, 1979] to compute a proxy $\check{\lambda}_\varepsilon$. We define $\lambda_{\mathsf{QUT}} := \check{\lambda}_{.05}$ the value we use in practice, and refer in what follows to this method of selecting λ as *quantile universal threshold*.

When covariates are included in the model, (12) boils down to testing if the

measured features are sufficient to explain the variability of the data, or if unobserved latent covariates also influence the counts. In Section 4.1 we compare on a simulation study the empirical properties of cross-validation and quantile universal threshold. Both methods give a potential estimate of the rank of the interaction matrix Θ, as singular values smaller than λ are set to zero.

4 Simulation study

4.1 Comparison of cross-validation and quantile universal threshold

We generate a contingency table according to the model $Y \sim \mathcal{P}(\exp(X^*))$, where $\mathcal{P}$ denotes the Poisson distribution, with $X^* = X_0^* + \Theta^*$, $(X_0^*)_{ij} = \mu^* + \alpha_i^* + \beta_j^*$. We draw an offset μ^*, row and column effects α_i^* and β_j^* from a standard normal distribution, and generate $\Theta^* = UDV^T$, where U and V are random orthonormal matrices and $D \in \mathbb{R}^{K \times K}$ is a diagonal matrix containing the singular values of Θ^*. 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 $\text{SNR} = \|\Theta^*\|_* / \|X_0^*\|_*$.
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 $\hat{K}_{\text{QUT}}$ 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 (the row-column model) (1) with the rank found using quantile universal threshold. We observe the same results over 1000 replications.

With the same simulation scheme, we further investigate the performance of our method, referred to as LORI (LOw-Rank Interaction), in different situations. We vary the values of the rank (K =2, 5 and 10) and the signal to noise ratio. Figure 2 highlights three interaction regimes. We observe similar behaviours for the different ranks and comment the case where $K = 2$. In the small interaction regime (Figure 2, top left, $\mathsf{SNR} = 0.2$), 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, $\mathsf{SNR} = 0.7$) 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 in-

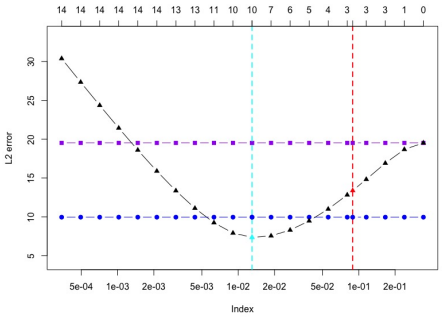


Figure 1: ℓ_2 loss (black triangles) of the estimator as a function of λ; $m_1 = 20$, $m_2 = 15$, $K = 3$. Comparison of λ_{CV} (cyan dashed line) and λ_{QUT} (red dashed line) with the independence model (purple squares) and the *row-column model* with oracle rank (blue points). The rank of Θ is written along the top for each λ.

teraction setting (Figure 2, top right $\mathsf{SNR} = 1.7$), quantile universal threshold overestimates the rank (here 6 instead of 2), and the row-column model fails

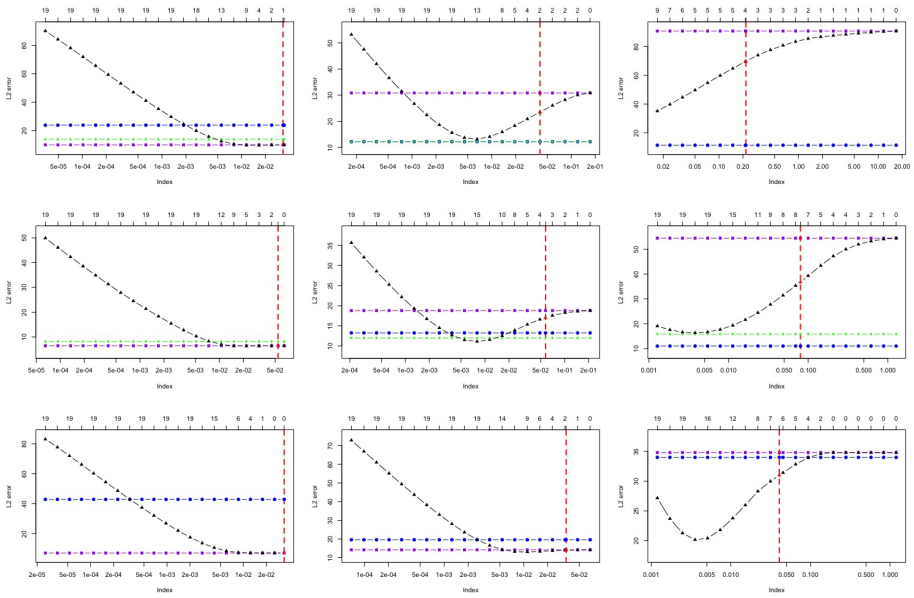


Figure 2: 50×20 matrices. Comparison of the ℓ_2 error of LORI (black triangles) with the independence model (purple squares), the row-column model with oracle rank (blue points) and with rank $\hat{K}_{\mathsf{QUT}}$ (green diamonds). Results are drawn for a grid of λ with λ_{QUT} (red dot). The rank 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 $\mathsf{SNR} = 0.2, 0.7, 1.7$.

to calculate the maximum likelihood estimation (possibly because of numerical issues that occur in available R libraries).

4.2 Simulating covariates

We simulate under model (2) as follows with $R \in \mathbb{R}^{100\times 2}$ and $C \in \mathbb{R}^{50\times 3}$ coming from a mixture of multivariate Gaussian distributions with different means and equal covariance matrices. More precisely, for the first half of the environments, $R_1 \in \mathbb{R}^{50\times 2}$ is drawn from $\mathcal{N}(\mu_R^1, \Sigma_R)$, and for the second half, $R_2 \in \mathbb{R}^{50\times 2}$ is drawn from $\mathcal{N}(\mu_R^2, \Sigma_R)$, with $\mu_R^1 = (1,1)$ and $\mu_R^2 = (2,2)$. Similarly, for the first half of the species, $C_1 \in \mathbb{R}^{25\times 3}$ is drawn from $\mathcal{N}(\mu_C^1, \Sigma_C)$, and for the second half, $C_2 \in \mathbb{R}^{25\times 3}$ is drawn from $\mathcal{N}(\mu_C^2, \Sigma_C)$, with $\mu_C^1 = (1,1,1)$ and $\mu_C^2 = (1,0,2)$. We consider two correlation structures. First, the case where Σ_R and Σ_C are identity matrices, *i.e.* the covariates are independent, and second, a case where covariates are correlated, with

$$\Sigma_R = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix} \text{ and } \Sigma_C = \begin{pmatrix} 1 & 0.3 & 0.2 \\ 0.3 & 1 & 0.3 \\ 0.2 & 0.3 & 1 \end{pmatrix}.$$

The parameters in μ, α and β are then i.i.d. Gaussian variables with mean 1 and variance 1. The interaction matrix Θ is generated as in Section 4.1. Then, its rows and columns are projected on the linear subspaces orthogonal to the columns of R and C respectively. We generate 100 samples and apply our method, using quantile universal threshold to select λ. We also apply Correspondence Analysis [Greenacre, 1984], which is a component-based method to analyse contingency tables, and can be used to estimate the parameter matrix. The results are compared in terms of relative estimation error $\left\|\hat{X} - X^*\right\|_F / \|X^*\|_F$, and given in Tables 1 and 2. LORI achieves lower errors, which is expected as it takes into account covariates, but the errors are especially lower when the covariates are independent. This can be explained by the fact that, when covariates are correlated, not taking them into account leads to losing less information than when they are independent. Correspondence Analysis also has very large errors in a few cases where count values are either large or have many zeros. LORI does not exhibit such behaviour due to the regularization.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
LORI	0.17	0.24	0.30	0.32	0.36	0.60
CA	20.91	48.20	140.04	3.0e4	614.14	2.3e6

Table 1: Independent covariates - Comparison of relative estimation errors of LORI, using λ_{QUT} and Correspondence Analysis (CA) on simulations with covariates, over 100 replications.

5 Analysis of the Aravo data

The Aravo dataset [Choler, 2005] measures the abundance of 82 species of alpine plants in 75 sites in France. The data consist of a contingency table collecting

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
LORI	0.46	0.99	1.00	1.51	1.02	7.86
CA	1.41	1.64	1.74	5.12e5	1.86	5.04e7

Table 2: Correlated covariates - Comparison of relative estimation errors of LORI, using λ_{QUT} and Correspondence Analysis (CA) on simulations with covariates, over 100 replications.

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.
We first compare the simple model (1) where the covariates are not taken into account with our model (2), to see how the incorporation of covariates impacts the interpretation. Figures 3 and 4 show visualizations of the data in the selected dimensions of interaction, defined by the first singular vectors scaled to the eigenvalues 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

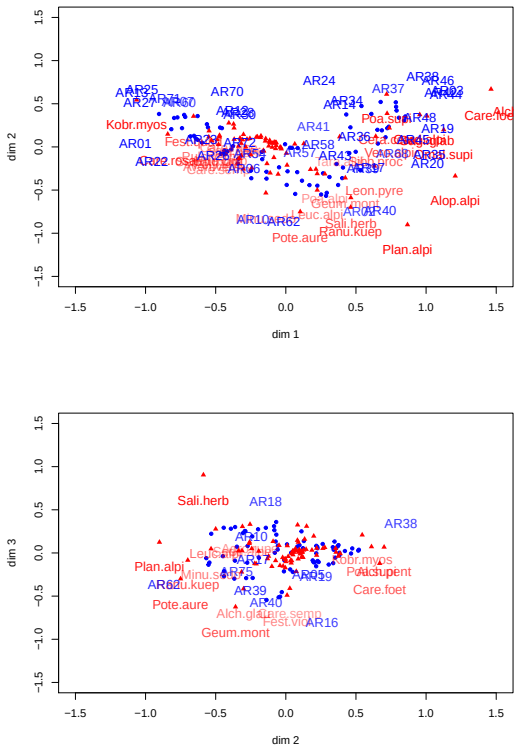


Figure 3: Display of the three first dimensions of interaction estimated with model (1). Environments are represented in blue and species in red.

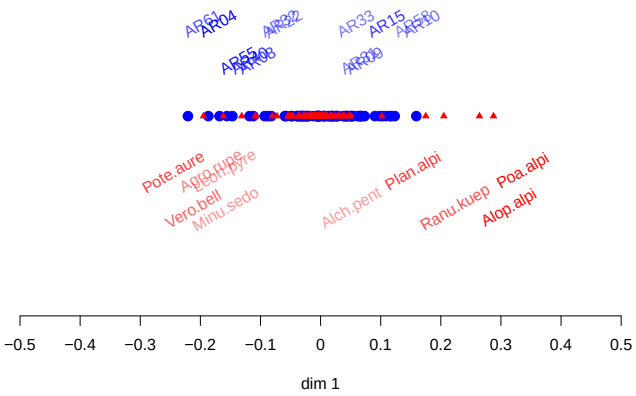


Figure 4: One-dimensional display of the first dimension of interaction estimated with model (2). Environments are represented in blue and species in red.

Figure 3 where we do not use the covariates, we find a rank of 3 for the interaction, while in Figure 4 after incorporating the covariates we find a rank 1. This suggests that an additional unknown variable summarizes the remaining interactions.

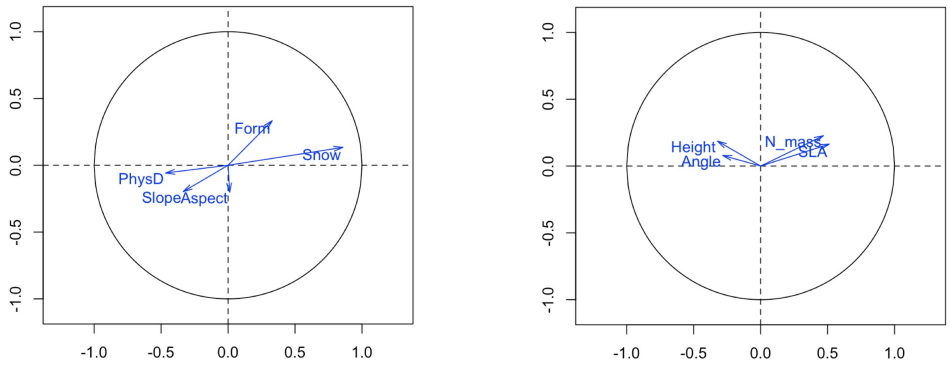
In the case of model (1) with no covariates, we can look at the relations between the known traits and the interaction directions of $\hat{\Theta}^{\lambda}$. Figure 5a shows that environment covariates and the two first directions of interaction are correlated. The first direction is correlated with 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 left graph 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 5b, 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). This confirms the fact that our method leads to different interpretations and suggests that the covariates are not sufficient to explain the variability of the counts.

6 Using covariates to impute ecological data

The water-birds data count the abundance of migratory water-birds in 722 wetland sites (across the 5 countries in North Africa), between 1990 and 2016

[Sayoud et al., 2017]. One of the objectives is to assess the effect of time on species abundances, to monitor the populations and assess wetlands conservation policies. Ornithologists have also recorded side information concerning the sites and years, which may influence the counts. For instance, the political situation in a country, the latitude and longitude. The contingency table contains a large amount of missing entries (70%), but the covariate matrices which contain respectively 6 covariates about the 722 sites and 8 covariates about the 17 years, are fully observed. Our method allows to skip the missing values to perform the analysis and to take advantage of the available covariates, to provide interpretation for temporal patterns. As a by-product, it gives an imputed contingency table that used the available counts and covariates.

Figure 6 displays for a given site i, the row vector $(R\mu C^\top + R\beta)_{i,\cdot}$ versus the year. One line corresponds to one site. It represents the evolution of the impact of the sites covariates R on the counts over time. A line which is high on average corresponds to an "abundant" site with respect to the measured covariates, *i.e.* the covariate values taken by the site are associated to higher counts. Peaks indicate years when the site covariates have a larger effect on the counts. Sites presenting the same positive peak indicate that the corresponding year yields high counts, for example a year where the meteorological conditions were particularly good. Sites presenting opposite peaks indicate an interaction between year and site covariates. For example, a political measure may foster the sites close to urban centres but disadvantage the more isolated ones.
Figure 6 highlights sites with similar temporal trends and three particular years 1992, 1995 and 1996. To further investigate this behaviour, one can wonder if the observed association between the site covariates and the years can be explained by the year covariates. Figure 7 separates the terms $R\mu C^\top$ and $R\beta$. The three peaks observed in Figure 6 are also observed on the right plot of Figure 7 which means that they can be explained by characteristics of the years which are not measured in the year covariates C. This is in accordance with the fact that the rank of the interaction matrix Θ is estimated at 10, meaning that 10 latent features might be at work. Interestingly, we observe in 2012,



(a) Environment covariates

(b) Species traits

Figure 5: Correlation between the two first dimensions of interaction and the covariates (the covariates are not used in the estimation).

two opposite peaks in $R\mu C^\top$ and $R\beta$ which seem to cancel out in Figure 6, indicating that some year features (respectively measured covariates and latent traits) could have opposite effects on the counts. Thus, our method reveals the "good" characteristics of year 2012 with respect to the measured year covariates.

7 Discussion

We conclude by discussing some opportunities for further research. To select covariates, we could penalize the main effects with an ℓ_1 penalty on α and β. It may be also 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. Secondly, our algorithm directly handles missing values and can be used to impute contingency tables. It would therefore be interesting to extend our theoretical guarantees to the missing data framework and to assess the quality of imputation. Indeed, few methods are available for single and multiple imputation of contingency tables, especially with covariates. The properties of the thresholding test, which can be seen as an alternative to a chi-squared test for independence with covariates, also merit further investigation. In particular, the power could be assessed. Finally, we could also explore whether our model could be extended to more complex models such as the zero-inflated negative binomial models, which are often used in ecological applications.

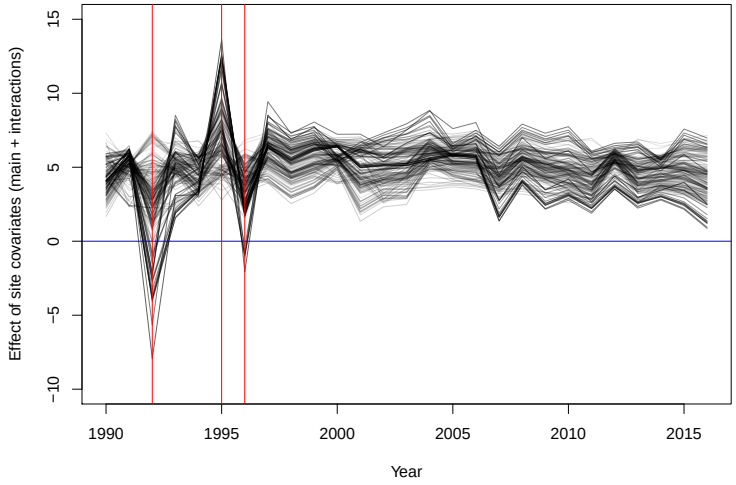


Figure 6: $(R\mu C^\top + R\beta)_{ij}$ versus the year (indexed by j), one line for each site i. The blue line indicates the overall mean of $R\mu C^\top + R\beta$. Only the 50% most variable lines in terms of total variation norm are plotted.

8 Proofs

8.1 Proof of Theorem 1

The proof of Theorem 1 derives from the strong convexity of Φ_Y and tail bounds for the largest singular value of random matrices with sub-exponential 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 2. *Under Assumption 1, assume* $\lambda \geq 2\left\|\nabla\Phi_Y(X^*)\right\|$. *Then*

$$\frac{\left\|X^* - \hat{X}_\lambda\right\|_F^2}{m_1 m_2} \leq \left(\frac{16\lambda^2 m_1 m_2}{\sigma_{min}^4}\right)\{\mathrm{rk}(\Theta^*) + K_1 + K_2\}. \tag{13}$$

We prove this result in Section 8.2.

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

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

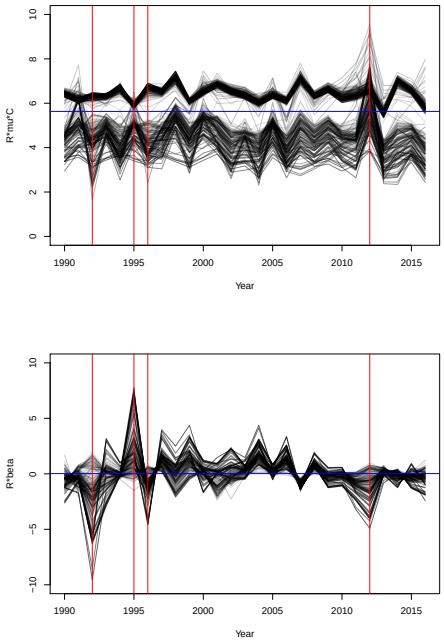


Figure 7: $(R\mu C^\top)_{ij}$ (left) and $(R\beta)_{ij}$ (right) versus the year (indexed by j), one line for each site i. The blue lines indicate the overall means of $R\mu C^\top$ and $R\beta$. Only the 50% most variable lines in terms of total variation norm are plotted.

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

$$\sigma_Z^2 = \max\left(\frac{1}{m_1 m_2}\left\|\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} E\left(Z_{ij}Z_{ij}^T\right)\right\|, \quad \frac{1}{m_1 m_2}\left\|\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} E\left(Z_{ij}^T Z_{ij}\right)\right\|\right). \tag{14}$$

We have

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

Moreover, note that

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

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

$$\frac{M}{m_1 m_2}\sigma_{\min}^2 \leq \sigma_Z^2 \leq \frac{M}{m_1 m_2}\sigma_{\max}^2$$

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

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

The condition

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

ensures that the left term dominates. Then, taking

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

and plugging this value in (13) of Proposition 2 directly gives the result.

8.2 Proof of Proposition 2

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

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

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

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

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

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

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

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 2. *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 1(ii). □

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

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

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

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

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

$$\frac{\sigma^2_{\mathsf{min}}\left\|X^* - \hat{X}\right\|^2_F}{2m_1m_2} \leq -\langle\nabla\Phi_Y(X^*), \hat{X} - X^*\rangle + \lambda\left(\|\Theta^*\|_* - \left\|\hat{\Theta}\right\|_*\right). \tag{15}$$

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

$$\begin{aligned}&-\langle\nabla\Phi_Y(X^*), \hat{X} - X^*\rangle \leq \|\nabla\Phi_Y(X^*)\| \times \\ &\left(\left\|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\right\|_* + \left\|\mathcal{P}^{\perp}_{\Theta^*}(\hat{\Theta} - \Theta^*)\right\|_* + \|R\hat{\alpha} - R\alpha^*\|_* + \left\|(C\hat{\beta} - C\beta^*)^T\right\|_*\right).\end{aligned} \tag{16}$$

Then, Lemma 1 (ii) applied to $\hat{X}$, results in

$$\|\Theta^*\|_* - \left\|\hat{\Theta}\right\|_* \leq \left\|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\right\|_* - \left\|\mathcal{P}^{\perp}_{\Theta^*}(\hat{\Theta} - \Theta^*)\right\|_* . \tag{17}$$

Plugging inequalities (16) and (17) in (15), and using the condition $\lambda \geq 2\|\nabla\Phi_Y(X^*)\|$, we finally obtain

$$\frac{\sigma^2_{\min}\left\|X^* - \hat{X}\right\|^2}{m_1 m_2} \leq 3\lambda\left\|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\right\|_* + \lambda\left(\|R\hat{\alpha} - R\alpha^*\|_* + \left\|(C\hat{\beta} - C\beta^*)^T\right\|_*\right). \tag{18}$$

Then, $\mathrm{rk}(R\hat{\alpha} - R\alpha^*) \leq K_1$ and $\mathrm{rk}(C\hat{\beta} - C\beta^*) \leq K_2$ imply $\|R\hat{\alpha} - R\alpha^*\|_* + \left\|(C\hat{\beta} - C\beta^*)'\right\|_* \leq \left(\sqrt{K_1} + \sqrt{K_2}\right)\left\|X^* - \hat{X}\right\|_F$, which together with Lemma 1 (iii) and $2(a^2 + b^2) \geq (a+b)^2$ yield Proposition 2.

8.3 Proof of Theorem 2

We assume without loss of generality that $m_1 \geq m_2$. We prove separately lower bounds of order K_1/m_2, K_2/m_1 and rM/m_1m_2. To do so, we only need to prove such lower bounds for particular cases of covariates R and C.
We start by proving a lower bound of order K_2/m_1. Consider the following covariate matrices:

$$R = 0, \quad C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{matrix} \uparrow \\ \\ \\ \downarrow \\ \uparrow \\ \\ \\ \downarrow \end{matrix} \begin{matrix} \\ \\ \\ K_2 \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ m_2 - K_2 \end{matrix} .$$

For some $\eta_\alpha \in (0,1)$ we define the following set of matrices of $\mathbb{R}^{K_2 \times m_1}$ where α will lie.

$$\mathcal{A} = \left\{\alpha \in \mathbb{R}^{K_2 \times m_1}; \alpha_{kj} \in \{0, \eta_\alpha \min(\gamma, \sigma_{\max})\}\right\}.$$

Also define $\mathcal{X}_\alpha = \{(C\alpha)^\top; \alpha \in \mathcal{A}\}$. In other words, we consider cases where the parameters β and Θ are both 0. Note that for all $\alpha \in \mathcal{A}$, $X = (C\alpha)^\top \in \mathcal{F}(r, \gamma)$. The Varshamov-Gilbert bound [Tsybakov, 2008, Lemma 2.9] guarantees that there exists a subset $\mathcal{A}_0 \subset \mathcal{A}$ of cardinality $\mathrm{Card}(\mathcal{A}_0) \geq 2^{K_2 m_1/8} + 1$, containing the zero $K_2 \times m_1$ matrix, and such that for any two distinct elements α and α' in $\mathcal{A}_0$,

$$\|\alpha - \alpha'\|_F^2 \geq \frac{K_2 m_1 \eta_\alpha^2 \min(\gamma, \sigma_{\max})^2}{8}.$$

Let $\mathcal{X}_{\alpha,0} = \{(C\alpha)^\top; \alpha \in \mathcal{A}_0\}$. The definition of C implies that for any two elements X and X' in $\mathcal{X}_{\alpha,0}$ we also have

$$\|X - X'\|_F^2 \geq \frac{K_2 m_1 \eta_\alpha^2 \min(\gamma, \sigma_{\max})^2}{8}. \tag{19}$$

Now, the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X)$ satisfies

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

Using the fact that $x \mapsto \exp(x) - 1 - x$ is gradient Lipschitz on $(0, \gamma)$ with constant σ_{max}, and the definition of $\mathcal{X}_{\alpha,0}$, we obtain

$$KL(\mathbb{P}_X, \mathbb{P}_0) \leq K_2 m_1 \sigma_{\mathsf{max}}^2 \eta_\alpha^2 \min(\gamma, \sigma_{\mathsf{max}})^2.$$

Taking $\eta_\alpha = \min\left(1, \frac{1}{8\sqrt{2}\sigma_{\mathsf{max}} \min(\gamma, \sigma_{\mathsf{max}})}\right)$, we obtain that

$$\frac{1}{\mathrm{Card}(\mathcal{X}_{\alpha,0})} \sum_{X \in \mathcal{X}_{\alpha,0}} \mathrm{KL}(\mathbb{P}_X, \mathbb{P}_0) \leq \frac{1}{16} \log_2(\mathrm{Card}(\mathcal{X}_{\alpha,0})). \tag{20}$$

Equations (19) and (20) guarantee that we can use [Tsybakov, 2008, Theorem 2.5], which gives that there exists $\kappa_\alpha > 0$ such that

$$\inf_{\hat{X}} \sup_{X \in \mathcal{F}(r,\gamma)} \mathbb{P}_X \left(\frac{\left\| \hat{X} - X \right\|_F^2}{m_1 m_2} > \frac{K_2}{m_2} \min\left(\frac{\min(\gamma, \sigma_{\mathsf{max}})^2}{16}, \frac{\sigma_{\mathsf{max}}^2}{2048} \right) \right) \geq \kappa_\alpha. \tag{21}$$

Let us now prove a lower bound of order K_1/m_1. Consider the following covariate matrices:

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{matrix} \uparrow \\ \\ \\ \downarrow \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ K_1 \\ \\ \\ \\ \\ \end{matrix}, \quad C = 0.$$

$$\begin{matrix} \\ \\ \\ \\ \uparrow \\ \\ \\ \\ \downarrow \end{matrix} \begin{matrix} \\ \\ \\ \\ m_1 - K_1 \\ \\ \\ \\ \end{matrix}$$

For some $\eta_\beta \in (0, 1)$ we define the following set of matrices of $\mathbb{R}^{K_1 \times m_2}$ where β will lie.

$$\mathcal{B} = \left\{ \beta \in \mathbb{R}^{K_1 \times m_2}; \beta_{kj} \in \{0, \eta_\beta \min(\gamma, \sigma_{\mathsf{max}})\} \right\}.$$

Also define $\mathcal{X}_\beta = \{R\beta; \beta \in \mathcal{B}\}$. In other words, we consider cases where the parameters α and Θ are both 0. Note that for all $\beta \in \mathcal{B}$, $X = R\beta \in \mathcal{F}(r, \gamma)$. The Varshamov-Gilbert bound [Tsybakov, 2008, Lemma 2.9] guarantees that there exists a subset $\mathcal{B}_0 \subset \mathcal{B}$ of cardinality $\mathrm{Card}(\mathcal{B}_0) \geq 2^{K_1 m_2/8} + 1$, containing the zero $K_1 \times m_2$ matrix, and such that for any two distinct elements β and β' in $\mathcal{B}_0$,

$$\|\beta - \beta'\|_F^2 \geq \frac{K_1 m_2 \eta_\beta^2 \min(\gamma, \sigma_{\mathsf{max}})^2}{8}.$$

Let $\mathcal{X}_{\beta,0} = \{R\beta; \beta \in \mathcal{B}_0\}$. The definition of R implies that for any two elements X and X' in $\mathcal{X}_{\beta,0}$ we also have

$$\|X - X'\|_F^2 \geq \frac{K_1 m_2 \eta_\beta^2 \min(\gamma, \sigma_{\mathsf{max}})^2}{8}. \tag{22}$$

Now, the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X)$ satisfies

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

Using the fact that $x \mapsto \exp(x) - 1 - x$ is gradient Lipschitz on $(0, \gamma)$ with constant $\sigma_{\max}$, and the definition of $\mathcal{X}_{\beta,0}$, we obtain

$$KL(\mathbb{P}_X, \mathbb{P}_0) \leq K_1 m_2 \sigma_{\max}^2 \eta_\beta^2 \min(\gamma, \sigma_{\max})^2.$$

Taking $\eta_\beta = \min\left(1, \frac{1}{8\sqrt{2}\sigma_{\max}\min(\gamma, \sigma_{\max})}\right)$, we obtain that

$$\frac{1}{\mathrm{Card}(\mathcal{X}_{\beta,0})} \sum_{X \in \mathcal{X}_{\beta,0}} \mathrm{KL}(\mathbb{P}_X, \mathbb{P}_0) \leq \frac{1}{16} \log_2(\mathrm{Card}(\mathcal{X}_{\beta,0})). \tag{23}$$

Equations (22) and (23) guarantee that we can use [Tsybakov, 2008, Theorem 2.5], which gives that there exists $\kappa_\beta > 0$ such that

$$\inf_{\hat{X}} \sup_{X \in \mathcal{F}(r,\gamma)} \mathbb{P}_X \left(\frac{\left\| \hat{X} - X \right\|_F^2}{m_1 m_2} > \frac{K_1}{m_1} \min\left(\frac{\min(\gamma, \sigma_{\max})^2}{16}, \frac{\sigma_{\max}^2}{2048} \right) \right) \geq \kappa_\beta. \tag{24}$$

Let us now prove a lower bound of order $rM/m_1 m_2$.Consider covariate matrices $R = 0, C = 0$. For some $\eta_\Theta \in (0, 1)$ we define the following set of matrices of $\mathbb{R}^{m_1 \times r}$.

$$\tilde{\mathcal{L}} = \left\{ \tilde{\Theta} \in \mathbb{R}^{m_1 \times r}; \tilde{\Theta}_{kj} \in \left\{ 0, \eta_\Theta \min(\gamma, \sigma_{\max}) \sqrt{\frac{rM}{m_1 m_2}} \right\} \right\},$$

and the associated set of block matrices

$$\mathcal{L} = \left\{ L = (0|\tilde{L}|\ldots|\tilde{L}) \in \mathbb{R}^{m_1 \times m_2} : \tilde{L} \in \tilde{\mathcal{L}} \right\},$$

where 0 denotes the $m_1 \times (m_2 - r\lfloor m_2/r \rfloor)$ zero matrix and $\lfloor x \rfloor$ is the integer part of x. Also define $\mathcal{X}_\Theta = \{\Theta; \Theta \in \mathcal{L}\}$. In other words, we consider cases where the parameters α and β are both 0. Note that for all $\Theta \in \mathcal{L}$, $X = \Theta \in \mathcal{F}(r, \gamma)$. The Varshamov-Gilbert bound [Tsybakov, 2008, Lemma 2.9] guarantees that there exists a subset $\mathcal{L}_0 \subset \mathcal{L}$ of cardinality $\mathrm{Card}(\mathcal{L}_0) \geq 2^{rM/8} + 1$, containing the zero $m_1 \times m_2$ matrix, and such that for any two distinct elements Θ and Θ' in $\mathcal{L}_0$,

$$\|\Theta - \Theta'\|_F^2 \geq \frac{rM}{8} \eta_\Theta^2 \min(\gamma, \sigma_{\max})^2 \frac{rM}{m_1 m_2} \left\lfloor \frac{m_2}{r} \right\rfloor \geq \frac{\eta_\Theta^2}{16} \min(\gamma, \sigma_{\max})^2 \frac{rM}{m_1 m_2}.$$

Let $\mathcal{X}_{\Theta,0} = \{\Theta; \Theta \in \mathcal{L}_0\}$. We have trivially

$$\|X - X'\|_F^2 \geq \frac{\eta_\Theta^2}{16} \min(\gamma, \sigma_{\max})^2 \frac{rM}{m_1 m_2}. \tag{25}$$

Now, the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X)$ satisfies

$$\begin{aligned} KL(\mathbb{P}_X, \mathbb{P}_0) &\leq \left\lfloor \frac{m_2}{r} \right\rfloor rM \frac{rM}{m_1 m_2} \sigma_{\max}^2 \eta_\Theta^2 \min(\gamma, \sigma_{\max})^2 \\ &\leq rM \sigma_{\max}^2 \eta_\Theta^2 \min(\gamma, \sigma_{\max})^2. \end{aligned}$$

Taking $\eta_\Theta = \min\left(1, \frac{1}{8\sqrt{2}\sigma_{\max}\min(\gamma,\sigma_{\max})}\right)$, we obtain that

$$\frac{1}{\mathrm{Card}(\mathcal{X}_{\Theta,0})}\sum_{X\in\mathcal{X}_{\Theta,0}}\mathrm{KL}(\mathbb{P}_X,\mathbb{P}_0) \leq \frac{1}{16}\log_2(\mathrm{Card}(\mathcal{X}_{\Theta,0})). \tag{26}$$

Equations (25) and (26) guarantee that we can use [Tsybakov, 2008, Theorem 2.5], which gives that there exists $\kappa_\Theta > 0$ such that

$$\inf_{\hat{X}}\sup_{X\in\mathcal{F}(r,\gamma)}\mathbb{P}_X\left(\frac{\left\|\hat{X}-X\right\|_F^2}{m_1m_2} > \frac{rM}{m_1m_2}\min\left(\frac{\min(\gamma,\sigma_{\max})^2}{32}, \frac{1}{4096\sigma_{\max}^2}\right)\right) \geq \kappa_\Theta. \tag{27}$$

Theorem 2 follows from equations (21), (24) (27).

Acknowledgements

The authors thank Trevor Hastie, Edgar Dobriban, Olga Klopp, Kevin Bleakey and Stéphane Dray for their very helpful comments on this manuscript, and Pierre Defos du Rau and Laura Dami for giving access to their data and helping with the interpretation.

References

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

Roland Angst, Christopher Zach, and Marc Pollefeys. The generalized trace-norm and its application to structure-from-motion problems. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2502–2509, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-1-4577-1101-5. doi: 10.1109/ICCV.2011.6126536. URL http://dx.doi.org/10.1109/ICCV.2011.6126536.

J. Bigot, C. Deledalle, and D. Féral. Generalized sure for optimal shrinkage of singular values in low-rank matrix denoising. *Journal of Machine Learning Research*, 18:1–50, 2017.

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.

Alexandra M. Brown, David I. Warton, Nigel R. Andrew, Matthew Binns, Gerasimos Cassis, and Heloise Gibb. The fourth-corner solution – using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution*, 5(4):344–352, 2014. ISSN 2041-210X. doi: 10.1111/2041-210X.12163. URL http://dx.doi.org/10.1111/2041-210X.12163.

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.

Y. Cao and Y. Xie. Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing*, 64(6):1609–1620, March 2016. ISSN 1053-587X. doi: 10.1109/TSP.2015.2500192.

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.

Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. A generalization of principal component analysis to the exponential family. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pages 617–624, Cambridge, MA, USA, 2001. MIT Press. URL http://dl.acm.org/citation.cfm?id=2980539.2980620.

M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-Bit Matrix Completion. *Information and Inference: A Journal of the IMA*, 3:189–223, July 2014. doi: 10.1093/imaiai/iau006.

Antoine de Falguerolles. Chapter 35 - log-bilinear biplots in action. In Jörg Blasius and Michael Greenacre, editors, *Visualization of Categorical Data*, pages 527 – 539. Academic Press, San Diego, 1998. ISBN 978-0-12-299045-8. doi: https://doi.org/10.1016/B978-012299045-8/50039-5. URL https://www.sciencedirect.com/science/article/pii/B9780122990458500395.

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

S. Dolédec, D. Chessel, C. J. F. ter Braak, and S. Champely. Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics*, 3(2):143–166, Jun 1996. ISSN 1573-3009. doi: 10.1007/BF02427859. URL https://doi.org/10.1007/BF02427859.

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. *Ann. Statist.*, 7(1):1–26, 01 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

W. Fithian and R. Mazumder. Flexible Low-Rank Statistical Modeling with Side Information. *ArXiv e-prints*, August 2013.

William Fithian and Julie Josse. Multiple correspondence analysis and the multilogit bilinear model. *Journal of Multivariate Analysis*, 157: 87 – 102, 2017. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2017.02.009. URL http://www.sciencedirect.com/science/article/pii/S0047259X1730115X.

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. *IEEE Transactions on Information Theory*, 63(4):2137–2152, April 2017. ISSN 0018-9448. doi: 10.1109/TIT.2017.2653801.

Matan Gavish and David L Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040—-5053, 2014.

C. Giacobino, S. Sardy, J. Diaz Rodriguez, and N. Hengardner. Quantile universal threshold. *Electronic Journal of Statistics*, 11:4701–4722, 2017.

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.

M. Greenacre. *Theory and Applications of Correspondence Analysis*. Acadamic Press, 1984.

Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *J. Mach. Learn. Res.*, 16(1):3367–3402, January 2015. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2789272.2912106.

Julie Josse and Sylvain Sardy. Adaptive shrinkage of singular values. *Statistics and Computing*, 26(3):715–724, May 2016. ISSN 1573-1375. doi: 10.1007/s11222-015-9554-9. URL https://doi.org/10.1007/s11222-015-9554-9.

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.

Pierre Legendre, René Galzin, and Mireille L. Harmelin-Vivien. Relating behavior to habitat: Solutions to thefourth-corner problem. *Ecology*, 78(2):547–562, 1997. ISSN 1939-9170. doi: 10.1890/0012-9658(1997)078[0547:RBTHST]2.0.CO;2. URL http://dx.doi.org/10.1890/0012-9658(1997)078[0547:RBTHST]2.0.CO;2.

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.

Shakir Mohamed, Zoubin Ghahramani, and Katherine 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. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3532-bayesian-exponential-family-pca.pdf.

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.

M.S. Sayoud, H. Salhi, B. Chalabi, A. Allali, M. Dakki, A. Qninba, M.A. El Agbani, H. Azafzaf, C. Feltrup-Azafzaf, H. Dlensi, N. Hamouda, W. Abdel Latif Ibrahim, H. Asran, A. Abu Elnoor, H. Ibrahim, K. Etayeb, E. Bouras, W. Bashaimam, A. Berbash, C. Deschamps, J.Y. Mondain-Monval, A.L. Brochet, S. Véran, and P. Defos du Rau. The first coordinated trans-north african mid-winter waterbird census: The contribution of the international waterbird census to the conservation of waterbirds and wetlands at a biogeographical level. *Biological Conservation*, 206:11 – 20, 2017. ISSN 0006-3207. doi: https://doi.org/10.1016/j.biocon.2016.12.005. URL http://www.sciencedirect.com/science/article/pii/S0006320716309788.

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.

Cajo J.F. ter Braak, Pedro Peres-Neto, and Stéphane Dray. A critical issue in model-based inference for studying trait-based community assembly and a solution. *PeerJ*, 5:e2885, January 2017. ISSN 2167-8359. doi: 10.7717/peerj.2885. URL https://doi.org/10.7717/peerj.2885.

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