Journal of Multivariate Analysis 173 (2019) 416–434

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Low-rank model with covariates for count data with missing values

Geneviève Robin [a,b,*], Julie Josse [a,b], Éric Moulines [a,b], Sylvain Sardy [c]

[a] CMAP, École Polytechnique, route de Saclay, 91128 Palaiseau Cedex, France
[b] XPOP, INRIA, 1 rue Honoré d'Estienne d'Orves, Bâtiment Alan-Turing, Campus de l'École Polytechnique, 91120 Palaiseau, France
[c] Section de mathématiques, Université de Genève, 2-4, rue du Lièvre, CH 1211, Genève 4, Switzerland

ARTICLE INFO

Article history:
Received 2 October 2018
Received in revised form 15 April 2019
Accepted 15 April 2019
Available online 22 April 2019

AMS 2010 subject classifications:
primary 62H12
secondary 62F12

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

ABSTRACT

A complete methodology called LORI (Low-Rank Interaction), including a Poisson model, an algorithm, and an automatic selection of the regularization parameter, is proposed for the analysis of frequency tables with covariates, including an upper bound on the estimation error. A simulation study with synthetic data suggests that LORI improves empirically on state-of-the-art methods in terms of estimation and imputation. Illustrations show how the method can be interpreted through visual displays with the analysis of a well-known plant abundance data set, and the LORI outputs are seen to be consistent with known results. The relevance of the methodology is also demonstrated through the analysis of a waterbirds abundance contingency table from the French national agency for wildlife and hunting management. The method is available in the `R` package `lori` on the Comprehensive Archive Network (CRAN).

1. Introduction

Let Y be an $n \times p$ observation matrix of counts. Let also $R \in \mathbb{R}^{n\times K_1}$ and $C \in \mathbb{R}^{p\times K_2}$ be matrices containing row and column covariates, respectively. In our ecological application in Section 6, the rows of the contingency table represent ecological sites, and the columns represent years. For $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, p\}$, Y_{ij} is the number of waterbirds counted at site i during year j. The row feature $R_{i\ell}$, $\ell \in \{1, \ldots, K_1\}$, embeds geographical information about site i (latitude, longitude, distance to coast, etc.) while the column feature $C_{j\ell}$, $\ell \in \{1, \ldots, K_2\}$, codes meteorological characteristics of year j (precipitation, etc.). In addition, some entries of Y are missing. For example ecological sites are sometimes inaccessible because of meteorological or political conditions, and therefore cannot be counted.

Frequency tables of this sort are often analyzed using low-rank models [11,16,17,21,23,24], imposing a low-rank structure to an underlying parameter matrix. We assume a probabilistic framework with independent entries Y_{ij} following a Poisson model, viz.

$$Y_{ij} \sim \mathcal{P}(e^{X^*_{ij}}), \tag{1}$$

and focus on the estimation of X^* based on a low-rank assumption. The generalized additive main effects and multiplicative interaction model, or row–column model (see, e.g., [16,21]), in which

$$X^*_{ij} = \mu^* + \alpha^*_i + \beta^*_j + \Theta^*_{ij}, \quad \operatorname{rank}(\Theta^*) \le \min(n-1, p-1), \tag{2}$$

* Corresponding author at: CMAP, École Polytechnique, route de Saclay, 91128 Palaiseau Cedex, France.
E-mail address: genevieve.robin@polytechnique.edu (G. Robin).

https://doi.org/10.1016/j.jmva.2019.04.004
0047-259X/

is adequate for the purpose. In this model, 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; see p. 87 in Section 4.1.2 of [26].

To incorporate covariates in this framework, a natural idea is to express the row and column effects α_i^* and β_j^* as regression terms on the covariates. In other words, for $\mu^* \in \mathbb{R}$, $\alpha^* \in \mathbb{R}^{K_1}$, $\beta^* \in \mathbb{R}^{K_2}$ and $\Theta^* \in \mathbb{R}^{n \times p}$,

$$X_{ij}^* = \mu^* + \underbrace{\sum_{k=1}^{K_1} R_{ik}\alpha_k^*}_{\text{row effect}} + \underbrace{\sum_{l=1}^{K_2} C_{jl}\beta_l^*}_{\text{column effect}} + \Theta_{ij}^*, \quad \operatorname{rank}(\Theta^*) \leq \min(n-1, p-1). \tag{3}$$

Such an extension is useful in practice for two main reasons. First, estimated covariates coefficients (and in particular their signs) can be used to determine whether the studied covariates have positive or negative effects on the counts; this is particularly useful in ecology to check whether meteorological, geographical or political conditions favor or endanger species. Second, when the proportion of missing values is large, which is often the case in bird monitoring, incorporating (relevant) covariates can improve the imputation significantly.

Models related to (3) have been considered for statistical ecology applications in [5,6]. However, to the best of our knowledge, their theoretical and empirical properties have not been thoroughly studied. In contrast, the literature on convex low-rank matrix estimation is abundant and benefits from a substantial theoretical background, although few software packages with ready-made solutions are available for practitioners, and applications for count data outside image analysis [8,37,42] and recommendation systems [22] have not been attempted. The purpose of this paper is to develop a complete methodology for the inference of model (3), bridging the gap between convex low-rank matrix completion and model-based count data analysis.

After detailing related work in Section 1.1, we introduce in Section 2 a general model which includes (3). We propose an estimation procedure through the minimization of a data fitting term penalized by the nuclear norm of the matrix Θ, which acts as a convex relaxation of the rank constraint. Building up on existing results on nuclear norm regularized loss functions, we derive statistical guarantees in Section 2.1. In particular, we provide an upper bound for the Frobenius norm of the estimation error. In Section 3, we propose an optimization algorithm, and two methods to choose the regularization parameter automatically.

We provide a simulation study in Section 4 revealing that LORI outperforms state-of-the-art methods when the proportion of missing values is large and the interactions are of significant order compared to the main effects. In Section 5, we show on plant abundance data with side information how the results of our procedure can be interpreted through visual displays. In particular, the arising interpretation is consistent with known results from the original study [10]. In Section 6, we use LORI to analyze a water-birds abundance data set from the French national agency for wildlife and hunting management (ONCFS).

The proofs of the statistical guarantees are postponed to the Appendix, and the method is available as an R package [40] called `lori` (LOw-Rank Interaction) on the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=lori.

1.1. Related work

Model (3) is closely related to other models previously suggested in the statistical ecology literature to analyze frequency tables with row and column covariates. For instance, [6] and [5] suggested the model

$$X_{ij}^* = \mu^* + \alpha_i^* + \beta_j^* + \epsilon_{RC} R_i C_j, \tag{4}$$

where R_i with $i \in \{1, \ldots, n\}$ is a row trait and C_j with $j \in \{1, \ldots, p\}$ is a column trait. The interaction between covariates is modeled 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 main difference with model (3) is that we incorporate the covariates in the main effects rather than the interactions, which leads to different interpretations. In terms of estimation properties, the main advantage of (3) is that, as long as $K_1 \leq n$ and $K_2 \leq p$, we estimate fewer parameters. This is an important point for us because in many applications we are interested in (see, e.g., Section 6), a large proportion of entries is missing, limiting the amount of available data. Finally, model (4) was developed with the aim of testing significant associations between covariates, and its theoretical and empirical estimation properties, as far as we know, were not studied.

In the low-rank matrix completion literature, related approaches for count matrix recovery and dimensionality reduction can be embedded within the framework of low-rank exponential family estimation [12,25,33,34,36] as well as its Bayesian counterpart [22,38]. In terms of statistical guarantees, the theoretical performance of nuclear norm penalized estimators for Poisson denoising has been studied by Cao and Xie [8], who derive uniform bounds on the empirical error risk. Estimation rates are also given by Lafond [31], who obtained optimal bounds for matrix completion in the exponential family. These two papers do not account for possible covariates.

More recently, Chiquet et al. [9] developed a probabilistic PCA framework for the exponential family, where covariates can be included in the parameter space. Fithian and Mazumder [18] present a variety of low-rank problems, including the generalized nuclear norm penalty [3], that can be used to include row and column covariates. Similar estimation

problems were also considered, e.g., in [1,2]. However, to the best of our knowledge, these papers did not provide statistical guarantees and the practical advantages of such extensions compared to classical low-rank methods have not been thoroughly studied.

2. General model and estimation

We introduce a general version of the model described in the previous section. First, we relax the Poisson model and replace it with the following assumption on the distribution of Y_{ij} for $(i,j) \in \{1,\dots,n\} \times \{1,\dots,p\} = [n] \times [p]$.

H 1. *The random variables $Y = \{Y_{i,j}\}_{(i,j)\in[n]\times[p]}$ are independent and there exist $\gamma > 0$, $\sigma_- > 0$ and $\sigma_+ < \infty$ such that for all $i \in [n]$ and $j \in [p]$,*

$$e^{-\gamma} \le \mathrm{E}(Y_{ij}) \le e^{\gamma} \quad \text{and} \quad \sigma_-^2 \le \mathrm{var}(Y_{ij}) \le \sigma_+^2.$$

We define X^*_{ij} for all $(i,j) \in [n] \times [p]$ by

$$X^*_{ij} = \operatorname*{argmin}_{x\in\mathbb{R}} \{-\mathrm{E}(Y_{ij})x + \exp(x)\}\,.$$

In other words, we do not assume that the random variable Y_{ij} follows a Poisson distribution. The target parameter X^*_{ij} minimizes the Kullback–Leibler divergence between the distribution of Y_{ij} and a Poisson distribution. Let us also generalize the decomposition introduced in (3). Let S_1 and S_2 be fixed linear subspaces of $\mathbb{R}^n$ and $\mathbb{R}^p$, respectively. Let P_1 and P_2 be the orthogonal projection matrices on S_1 and S_2, $\mathcal{P}^{\perp} : X \in \mathbb{R}^{n\times p} \mapsto P_1 X P_2^{\top}$, $\mathcal{P} : X \in \mathbb{R}^{n\times p} \mapsto X - \mathcal{P}^{\perp}(X)$, $\mathcal{X}_0 \subset \{X \in \mathbb{R}^{n\times p} : \mathcal{P}^{\perp}(X) = 0\}$ and $\mathcal{T} = \{X \in \mathbb{R}^{n\times p} : \mathcal{P}(X) = 0\}$. We denote

$$r = \max\Big[\{\mathrm{rank}(A) \,:\, A \in \mathcal{X}_0\}\Big]. \tag{5}$$

Consider the following decomposition

$$X^* = X^*_0 + \Theta^*, \quad X^*_0 \in \mathcal{X}_0, \Theta^* \in \mathcal{T}. \tag{6}$$

For any integer $m \ge 1$, let 1_m denote the vector of 1s of length m. Model (3) is included in (6) by setting $S_1 = \{u \in \mathbb{R}^n : 1_n^{\top} u = 0\}$, $S_2 = \{v \in \mathbb{R}^p : 1_p^{\top} v = 0\}$, and

$$\mathcal{X}_0 = \left\{ \left(\mu + \sum_{k=1}^{K_1} R_{ik}\alpha_k + \sum_{k=2}^{K_2} C_{jk}\beta_k \right)_{(i,j)\in[n]\times[p]} : \mu \in \mathbb{R}, \alpha \in \mathbb{R}^{K_1}, \beta \in \mathbb{R}^{K_2} \right\}.$$

The dimension of this subspace is at most $K_1 + K_2 + 1$ and the rank of a matrix in $\mathcal{X}_0$ is less than 3.

Finally, we consider a setting with missing observations. Denote by $\Omega \subset [n] \times [p]$ the set of observed entries, i.e., $(i,j) \in \Omega$ if and only if Y_{ij} is observed. Define also the random variables (ω_{ij}) such that $\omega_{ij} = 1$ if Y_{ij} is observed and $\omega_{ij} = 0$ otherwise. We assume that Y is independent of the ω_{ij}s, and we assume a Missing Completely At Random (MCAR) scenario [35] in which the (ω_{ij})s are mutually independent Bernoulli random variables. For $(i,j) \in [n] \times [p]$, we denote $\pi_{ij} = \Pr(\omega_{ij} = 1)$. We assume that the probability of observing any entry is positive, i.e., there exists $\pi > 0$ such that

$$\min\Big\{\pi_{ij}(i,j) \in [n] \times [p]\Big\} = \pi > 0\,.$$

For $j \in [p]$, denote by $\pi_{\cdot j} = \pi_{1j} + \cdots + \pi_{nj}$ the probability of observing an element in the jth column (up to a normalizing constant). Similarly, for $i \in [n]$, denote by $\pi_{i\cdot} = \pi_{i1} + \cdots + \pi_{ip}$ the probability of observing an element in the ith row (up to a normalizing constant). We define the following upper bound:

$$\max\Big[\{\pi_{i\cdot} \,:\, i \in [n]\} \cup \{\pi_{\cdot j} j \in [p]\}\Big] \le \beta\,. \tag{7}$$

We can now define our data-fitting term, viz.

$$\mathcal{L}(X) = \sum_{(i,j)\in[n]\times[p]} \omega_{ij}\{-Y_{ij}X_{ij} + \exp(X_{ij})\}. \tag{8}$$

Let $\|\cdot\|$ be the operator norm (the largest singular value), $\|\cdot\|_\infty$ be the infinity norm (the largest entry in absolute value) and $\|\cdot\|_*$ be the nuclear norm (the sum of singular values). Our estimator of model (3), for a given regularization parameter $\lambda > 0$, is the minimizer of the data-fitting term (8) penalized by the nuclear norm of Θ, viz.

$$\begin{aligned} &(\hat{X}_0, \hat{\Theta}) \quad \in \operatorname{argmin} \quad \mathcal{L}(X_0 + \Theta) + \lambda\|\Theta\|_*, \\ &\text{such that} \quad \|X_0 + \Theta\|_\infty \le \gamma, (i,j) \in [n] \times [p] \\ &\qquad\qquad X_0 \in \mathcal{X}_0, \quad \Theta \in \mathcal{T}. \end{aligned} \tag{9}$$

Denote

$$\nabla\mathcal{L}(X) = \sum_{(i,j)\in[n]\times[p]} \omega_{ij}\{-Y_{ij} + \exp(X_{ij})\}E_{ij},$$

where the E_{ij}s are the matrices of the canonical basis of $\mathbb{R}^{n\times p}$, the gradient of $\mathcal{L}$ at X. Denote also $\partial^2\mathcal{L}/\partial x_{ij}^2$ the second derivative of $\mathcal{L}$ with respect to the (i,j)th coordinate. Consider the following condition.

H 2. *The function $\mathcal{L}$ is strongly convex and smooth on $[-\gamma-\varepsilon, \gamma+\varepsilon]^{n\times p}$ for some $\varepsilon > 0$. There exist $\sigma_- > 0$ and $\sigma_+ < \infty$ such that for all $X \in [-\gamma-\varepsilon, \gamma+\varepsilon]^{n\times p}$ and $(i,j) \in [n]\times[p]$, $\sigma_-^2 \le \partial^2\mathcal{L}(X)/\partial x_{ij}^2 \le \sigma_+^2$.*

2.1. Statistical guarantees

We now derive an upper bound on the Frobenius estimation error of estimator (9). Let the variables ϵ_{ij} with $(i,j) \in [n]\times[p]$ be i.i.d. Rademacher random variables independent of Y and Ω and define

$$\Sigma_R = \sum_{(i,j)\in[n]\times[p]} \epsilon_{ij}\omega_{ij}E_{ij}. \tag{10}$$

Theorem 1. *Assume **H** 1–2, and $\lambda \ge 2\|\nabla\mathcal{L}(X^*)\|$. Then for all $n, p \ge 1$, with probability at least $1 - 8/(n+p)$,*

$$\|X^* - \hat{X}\|_F^2 \le \frac{C}{\pi^2}\left[\left\{\frac{\lambda^2}{\sigma_-^4} + (\mathrm{E}\|\Sigma_R\|)^2\gamma^2\right\}\{\mathrm{rank}(\Theta^*) + r\} + \ln(n+p)\right],$$

where r, γ are defined in (5) and (9), C is a numerical constant whose value can be found in the proof and which is independent of n, p and X^*.

Proof. See Appendix. □

We then control $\mathrm{E}\,\|\Sigma_R\|$, and compute a value of λ such that the condition $\lambda \ge 2\,\|\nabla\mathcal{L}(X^*)\|$ holds with high probability. We will need the following additional assumption on the distribution of the counts.

H 3. *There exists $\delta > 0$ such that for all $(i,j) \in [n]\times[p]$, $\mathrm{E}\{\exp(|Y_{ij}|/\delta)\} < \infty$.*

Define the following quantities, with C^* a numerical constant defined in Lemma E and r, β and γ defined in (5), (7) and (9) respectively:

$$\Phi_1 = 48\sigma_+^2\beta\ln(n+p), \quad \Phi_2 = 36\delta^2(e-1)^2\ln^1\left(1 + 8\delta^2\frac{np}{\beta\sigma_-^2}\right)\ln^2(n+p),$$

$$\Phi_3 = 4C^{*2}\max\left[\beta^2, \ln\{\min(n,p)\}\right].$$

Theorem 2. *Assume H 1–H 3 and set*

$$\lambda = \max\left\{4\sigma_+\sqrt{3\beta\ln(n+p)}, 12\delta(e-1)\ln\left(1 + 8\delta^2\frac{np}{\beta\sigma_-^2}\right)\ln(n+p)\right\}.$$

Then with probability at least $1 - 10/(n+p)$,

$$\|X^* - \hat{X}\|_F^2 \le C\left[\{\max(\Phi_1,\Phi_2) + \Phi_3\}\left\{\mathrm{rank}(\Theta^*) + r\right\} + \ln(n+p)\right]/\pi^2, \tag{11}$$

where C is a numerical constant independent of n, p and X^*.

Proof. See Appendix. □

We recover an upper bound of order $\mathrm{rank}(\Theta^*)\beta/\pi^2$, which is classical in low-rank matrix estimation and completion [27,31] and equal to $\mathrm{rank}(\Theta^*)\max(n,p)/\pi$ when the sampling is almost uniform ($c_1\pi \le \pi_{ij} \le c_2\pi$). The additional term $r\beta/\pi^2$ accounts for explicit modeling of the covariates in the main effects. The constant term appearing in bound (11) grows linearly with the upper bound σ_+^2 and quadratically with the inverse of σ_-^2. This means that by relaxing Assumption 1 to allow $\mathrm{var}(Y_{ij})$ to grow as fast as $\ln(n+p)$ or decrease as fast as $1/\ln(n+p)$, we only lose a log-polynomial factor in bound (11).

3. Algorithm and selection of λ

3.1. Optimization algorithm

In this section, we propose an algorithm to solve (9) for the initial model

$$X^*_{ij} = \mu^* + \sum_{k=1}^{K_1} R_{ik}\alpha^*_k + \sum_{l=1}^{K_2} C_{jl}\beta^*_l + \Theta^*_{ij},$$

using alternating minimization [13], which consists in updating μ, α, β and Θ alternatively, each time along a descent direction. Note that in the algorithm and the entire numerical section, we relax the constraint $|\mu + R_{i,.}\alpha + C_{j,.}\beta + \Theta_{ij}| \leq \gamma$. Indeed, this constraint is mainly required to obtain statistical guarantees and we observed that in practice, for γ sufficiently large, this constraint is never reached. Denote

$$\mathcal{F}(\mu, \alpha, \beta, \Theta) = \mathcal{L}\{(\mu + R_{i,.}\alpha + C_{j,.}\beta)_{i,j} + \Theta\},$$

and $\nabla_\Theta \mathcal{F}$ the gradient of $\mathcal{F}$ with respect to Θ defined by $(\nabla_\Theta \mathcal{F}(\mu, \alpha, \beta, \Theta))_{ij} = -Y_{ij} + \exp(\mu + R_{i,.}\alpha + C_{j,.}\beta + \Theta_{ij})$ if $\omega_{ij} = 1$ and $(\nabla_\Theta \mathcal{F}(\mu, \alpha, \beta, \Theta))_{ij} = 0$ otherwise. At every iteration we solve three sub-problems. The sub-problem in μ can be solved in closed form at each iteration; the updates in α and β can be done simultaneously by estimating a Poisson generalized linear model (which can be done using standard algorithms implemented in available libraries); the update in Θ is along the proximal gradient direction, with a step size tuned using backtracking line search. Denote by $\mathcal{D}_\lambda$ the soft-thresholding operator of singular values at level λ; see Section 2 in [7]. The procedure is sketched in Algorithm 1.

Algorithm 1: Alternating minimization for problem (9)

1 **Initialize** $\mu^{[0]}$, $\alpha^{[0]}$, $\beta^{[0]}$, $\Theta^{[0]}$
2 **For** $t \in \{0, \ldots, T-1\}$
- a) $\mu^{[t+1]} \in \operatorname{argmin} \mathcal{F}(\mu, \alpha^{[t]}, \beta^{[t]}, \Theta^{[t]})$,
- b) $(\alpha^{[t+1]}, \beta^{[t+1]}) \in \operatorname{argmin} \mathcal{F}(\mu^{[t]}, \alpha, \beta, \Theta^{[t]})$,
- c) $\tau = 1$,
- d) $\Theta^{[t+1]} = \mathcal{D}_\lambda[\Theta^{[t]} - \tau\mathcal{P}\{\nabla_\Theta \mathcal{F}(\mu^{[t]}, \alpha^{[t]}, \beta^{[t]}, \Theta^{[t]})\}]$
- e) While $\mathcal{F}(\mu^{[t+1]}, \alpha^{[t+1]}, \beta^{[t+1]}, \Theta^{[t+1]}) + \lambda\|\Theta^{[t+1]}\|_* > \mathcal{F}(\mu^{[t+1]}, \alpha^{[t+1]}, \beta^{[t+1]}, \Theta^{[t]}) + \lambda\|\Theta^{[t]}\|_*$:
  - $\tau = \tau/2$
  - $\Theta^{[t+1]} = \mathcal{D}_\lambda[\Theta^{[t]} - \tau\mathcal{P}\{\nabla_\Theta \mathcal{F}(\mu^{[t]}, \alpha^{[t]}, \beta^{[t]}, \Theta^{[t]})\}]$.

Output $\mu^{[T]}$, $\alpha^{[T]}$, $\beta^{[T]}$, $\Theta^{[T]}$

Note that if $K_1 + K_2 > |\Omega|$, with $|\Omega|$ denoting the cardinality of Ω, the update in α and β does not have a unique solution. However in our targeted applications, typically $K_1 + K_2 \ll |\Omega|$. In the package `lori`, we additionally implemented a warm-start strategy [19], which consists in solving (9) for a large value of λ, then sequentially decreasing λ and solving the new problem using the previous estimate as a starting point. Thus, our implementation solves (9) for the entire regularization path at once. Note that, even though our theoretical guarantees require a MCAR mechanism, the estimation method still holds when entries are missing at random; see Section 1.3 in [35]. Its imputation properties are illustrated in an ecological application in Section 6.

3.2. Automatic selection of λ

A common way to select the regularization parameter is cross-validation, which consists in erasing a fraction of the observed cells in Y, estimating a complete parameter matrix $\hat{X}$ for a range of λ values, and choosing the parameter λ that minimizes the imputation error. This can be performed directly using LORI without modifying the code. Indeed, let $(\tilde{\omega}_{ij})$ denote the weights in $\{0, 1\}$ indicating which entries are observed after removing some of them for cross-validation, and denote

$$\tilde{\mathcal{F}}(\mu, \alpha, \beta, \Theta) = \sum_{(i,j)\in[n]\times[p]} \tilde{\omega}_{ij}\{-Y_{ij}(\mu + R_{i,.}\alpha + C_{j,.}\beta + \Theta_{ij}) + \exp(\mu + R_{i,.}\alpha + C_{j,.}\beta + \Theta_{ij})\}.$$

The optimization problem becomes

$$(\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\Theta}) \in \operatorname*{argmin}_{\Theta\in\mathcal{T}} \quad \tilde{\mathcal{F}}(\mu, \alpha, \beta, \Theta) + \lambda\|\Theta\|_*,$$

which can be solved using the method described in Section 3.1; see Algorithm 1. However, cross-validation is computationally costly. We suggest an alternative method to cross-validation, inspired by [14] and the work of Giacobino et al.

Table 1
Estimation error (RMSE) of regression coefficients $\sqrt{\|\hat{\alpha}-\alpha^*\|_2^2+\|\hat{\beta}-\beta^*\|_2^2}$ of LORI and a Poisson GLM, for decreasing values of $\tau = \|\Theta\|_F/\|X_0\|_F$.

τ		Mean of RMSE∗100	Standard deviation of RMSE∗100
1	LORI	52	1.6
	GLM	143	29
0.5	LORI	17	0.9
	GLM	22	1.0
0.25	LORI	4.3	0.37
	GLM	5.3	0.33
0.1	LORI	1.7	0.34
	GLM	2.6	0.34
0	LORI	0.88	0.2
	GLM	0.87	0.19

[20] on quantile universal threshold. In Theorem 3, we define the so-called null-thresholding statistic of estimator (9), 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 $\lambda \geq \lambda_0(Y)$.

Theorem 3 (*Null-thresholding Statistic*)**.** *The estimated interaction matrix $\hat{\Theta}^{\lambda}$ for a regularization parameter λ is null if and only if $\lambda \geq \lambda_0(Y)$, where $\lambda_0(Y)$ is the null-thresholding statistic*

$$\lambda_0(Y) = \|\nabla \mathcal{L}(\hat{X}_0)\|, \quad \textit{where } \hat{X}_0 \in \operatorname*{argmin}_{X \in \mathcal{X}_0} \mathcal{L}(X). \tag{12}$$

Proof. See Appendix A.1.

Here, $\|\cdot\|$ is the operator norm (the largest singular value). 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 hypotheses

$$\mathcal{H}_0 : \Theta^* = 0 \ \text{ against the alternative } \ \mathcal{H}_1 : \Theta^* \neq 0 \tag{13}$$

and test whether the parameter matrix X^* can be explained only in terms of linear combinations of the measured covariates. For a probability $\varepsilon \in (0,1)$, consider the upper ε-quantile λ_ε of the null-thresholding statistics, namely that satisfies $\Pr_{\mathcal{H}_0}\{\lambda_0(Y) > \lambda_\varepsilon\} < \varepsilon$. The test, which consists in comparing the statistics $\lambda_0(Y)$ to λ_ε, is of level $1-\varepsilon$ for (13). This can be seen as an alternative to the chi-squared test for independence which handles covariates. In practice we do not have access to the distribution under the null $\Pr_{\mathcal{H}_0}\{\lambda_0(Y) < \lambda\}$, but perform a parametric bootstrap [15] to compute a proxy $\check{\lambda}_\varepsilon$. In practice we recommend $\varepsilon = 0.05$ and use $\lambda_{\mathsf{QUT}} = \check{\lambda}_{.05}$, and refer to it in what follows as the quantile universal threshold (QUT). This selection of the regularization parameter is essentially the universal threshold of [14] extended to our setting.

4. Simulation study

4.1. Estimation

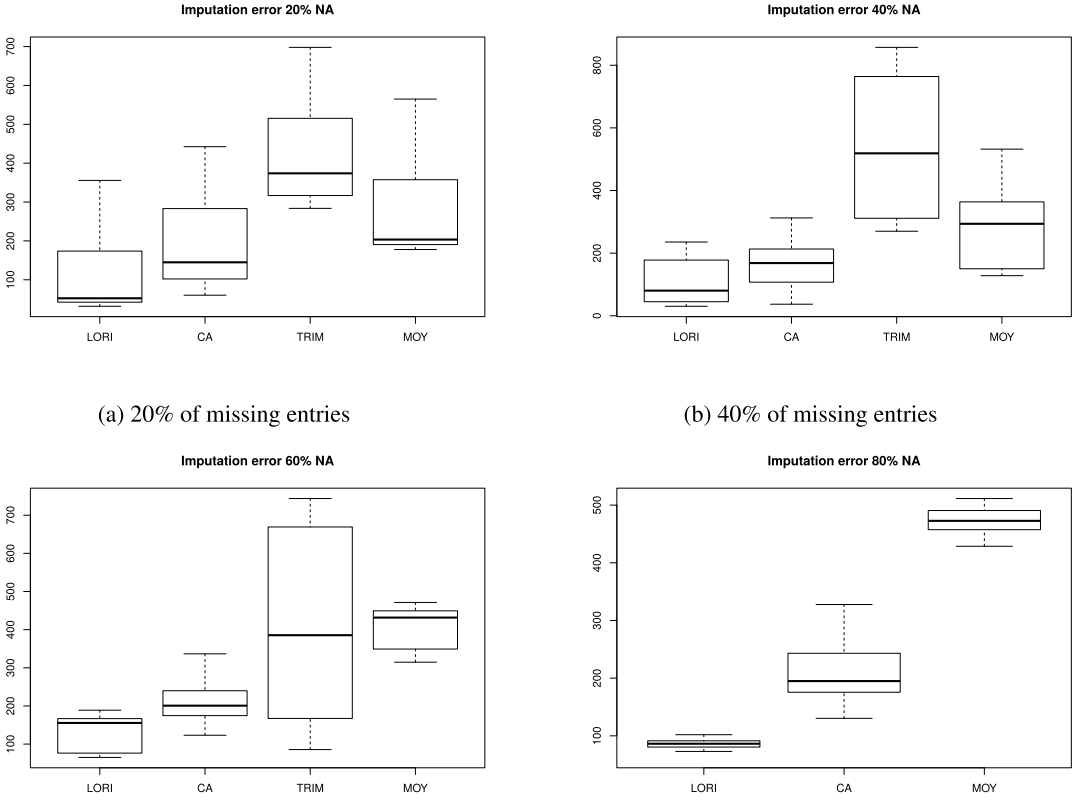
We simulate $Y \in \mathbb{N}^{300\times 30}$ under model (1)–(3), with $R \in \mathbb{R}^{500\times 3}$ and $C \in \mathbb{R}^{300\times 4}$ drawn from multivariate Gaussian distributions with mean 0 and block-diagonal covariance matrices Σ_R and Σ_C, respectively. We set $\mu^* = 1$, $\alpha^* = (2,0,0)$, $\beta^* = (-2,0,0,0)$ and Θ of rank 5. We compare the performance of LORI in terms of estimation of the regression coefficients α and β, and compare it to a standard Poisson Generalized Linear Model (GLM) estimated with the `glm` function in R. We repeat the experiment 100 times for decreasing values of the ratio $\tau = \|\Theta\|_F/\|X_0\|_F$, where $X_0 = (\mu + R_i\alpha + C_j\beta)_{ij}$ is fixed. We look at the Root Mean Square Error (RMSE) for the estimation of X_0; the results are given in Table 1, where we observe that LORI and the Poisson GLM are equivalent for $\tau = 0$, and that LORI outperforms the GLM for non-zero interactions, with a gap widening as τ increases.

Second, we compare LORI to a convex low-rank matrix estimation procedure with a Poisson loss function and where covariates are not modeled (e.g., [31]), in terms of the relative estimation error $\|\hat{X}-X^*\|_F/\|X^*\|_F$ because $\|X\|_F$ varies with τ. We refer to this competitor as the "Poisson LRM". Again, we reproduce the experiment 100 times for decreasing values of the ratio $\tau = \|\Theta\|_F/\|X_0\|_F$.

In Table 2, we observe that LORI achieves lower errors than the Poisson LRM, which is expected as we simulated under the LORI model. As τ decreases, i.e., the size of the main effects increases relative to the interactions, both errors decrease as well, and the gap between LORI and the Poisson LRM widens, indicating that modeling covariates explicitly improves the estimation.

Table 2
Estimation error (relative RMSE) of parameter matrix $\|\hat{X} - X^*\|_F / \|X^*\|_F$ of LORI and a Poisson GLM, for decreasing values of $\tau = \|\Theta\|_F / \|X_0\|_F$.

τ		Mean of relative RMSE∗100	Standard deviation of relative RMSE∗100
1	LORI	82	1.8
	Poisson LRM	95	4.4
0.5	LORI	40	0.25
	Poisson LRM	50	0.17
0.25	LORI	24	0.12
	Poisson LRM	40	0.12
0.1	LORI	11	0.081
	Poisson LRM	36	1.5
0	LORI	4.3	0.1
	Poisson LRM	34	1.2



(a) 20% of missing entries

(b) 40% of missing entries

(c) 60% of missing entries

(d) 80% of missing entries

Fig. 1. Average imputation error $\sum_{(i,j)\in\Omega}(Y_{ij} - \hat{Y}_{ij})^2/|\Omega|$.

4.2. Imputation

Using the same simulation scheme, we now compare LORI in terms of missing values imputation to Correspondence Analysis (CA) and Trends & Indices for Monitoring data (TRIM), a method based on a Poisson log-linear model used to impute bird abundance data [39]. To do so we erase an increasing proportion of entries in the data and impute them using LORI, CA and TRIM, replicating the experiment 100 times. We also impute the missing values using the column means, as a baseline referred to as "MOY". We observe on Fig. 1 that LORI performs best, which is expected as we simulate under the LORI model. Moreover, the gap widens as the percentage of missing values increases. In particular, the error of TRIM for 80% of missing entries is not represented because the method fails. Note that we used the default parameters of the R package `rtrim`.

Table 3
Main effect of the Aravo environment characteristics estimated with LORI. The regularization parameter is tuned using QUT.

Aspect	Slope	PhysD	Snow
0.04	0.07	−0.02	−0.07

Table 4
Main effect of the Aravo species traits estimated with LORI. The regularization parameter is tuned using QUT.

Height	Spread	Angle	Area	Thick	SLA	Nmass	Seed
0.09	−0.24	−0.18	−0.20	−0.11	−0.17	0.18	−0.12

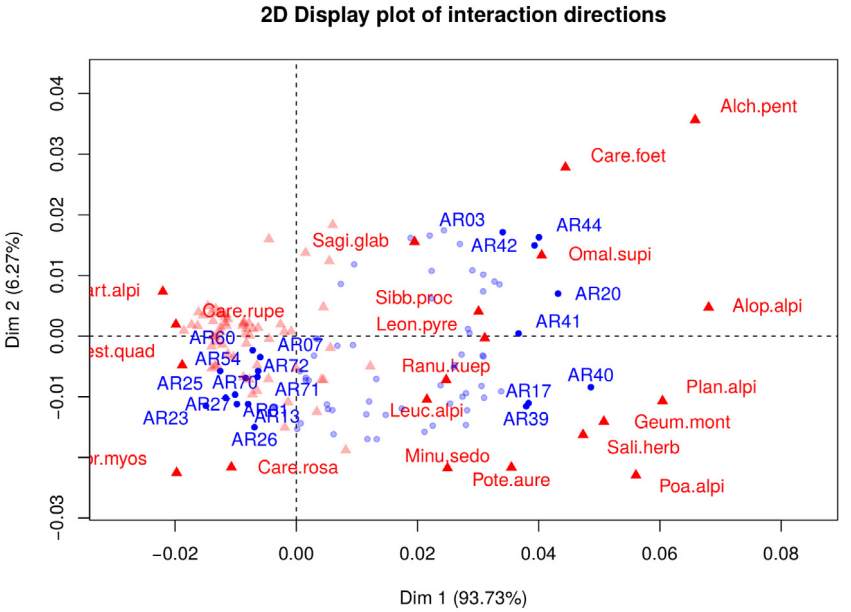


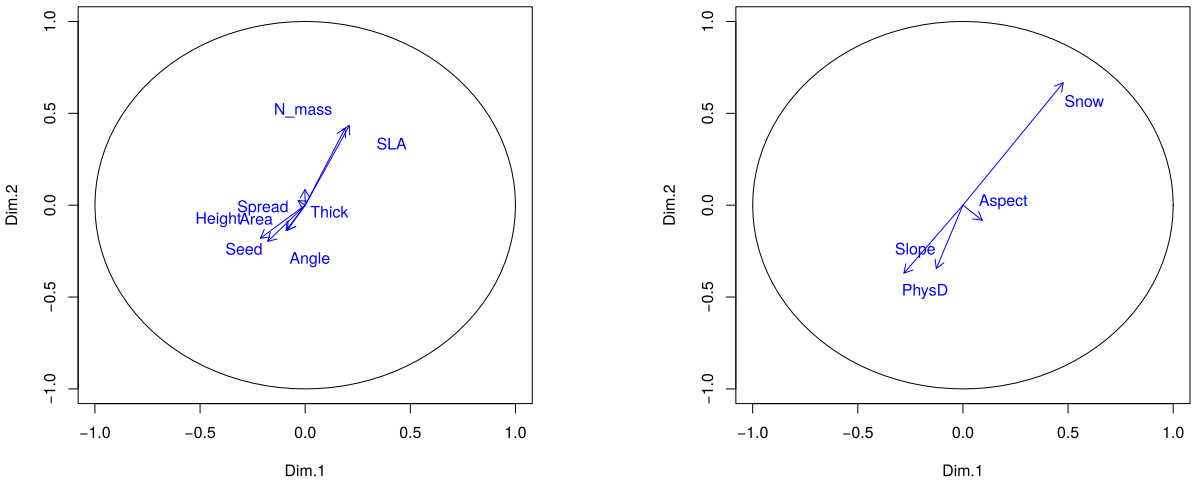
Fig. 2. Display of the two first dimensions of interaction estimated with LORI. Environments are represented with blue points and species with red triangles.

5. Analysis of the aravo data

The Aravo data set [10] counts the abundance of 82 species of alpine plants in 75 sites in France; covariates about environmental characteristics and species are also available. We focus on eight species traits providing physical information about plants (height, spread, etc.), and four environmental variables giving geographical and meteorological information about sites. We apply LORI after scaling the covariates and tuning the regularization parameter with the QUT method. This results in estimates for the main effects of the environment characteristics α and of the species traits β, and of the interaction matrix Θ.

The main effects of environment characteristics are given in Table 3 and the main effects of the species traits in Table 4. First we observe that overall, species traits have larger effects than environment characteristics on the observed abundances. In particular, the mass-based leaf nitrogen content (Nmass) has a large positive effect, which seems to indicate that plants with a large Nmass tend to be more abundant across all environments. In contrast, the maximum lateral spread of clonal plants (Spread), area of single leaf (Area) leaf elevation angle estimated at the middle of the lamina (Angle) and specific leaf area (SLA) have large negative effects on abundance.

The estimated rank of the interaction matrix $\hat{\Theta}$ (number of singular values above 10^{-6}) is 2. The environments (rows) and species (columns) can be visualized on a biplot (see Section 2.5 in [41]), where rows and columns are represented simultaneously in a normalized Euclidean space. In such plots, the dimensions of the Euclidean space are given by the principal directions of $\hat{\Theta}$, scaled by the square root of the singular values of $\hat{\Theta}$. Fig. 2 shows such a display, which can be interpreted in terms of distance between points: a species and an environment that are close interact highly, and two species or two environments that are close have similar profiles. Justifications for such a distance interpretation can be found in Section 2.5 of [41] or Section 2 of [17].



(a) Environment covariates

(b) Species traits

Fig. 3. Correlation between the two first dimensions of interaction and the covariates (the covariates are not used in the estimation).

Table 5
Main effect of the sites characteristics estimated with LORI. The regularization parameter is tuned using QUT.

Agricultural surface	Latitude	Dist. to town	Dist. to coast	Surface
0.09	−0.21	0.09	−0.48	0.20

Table 6
Main effect of the years characteristics estimated with LORI.

Spring N/O	Spring N/E	Winter S/O	Winter S/E
−0.05	−0.03	−0.05	−0.03

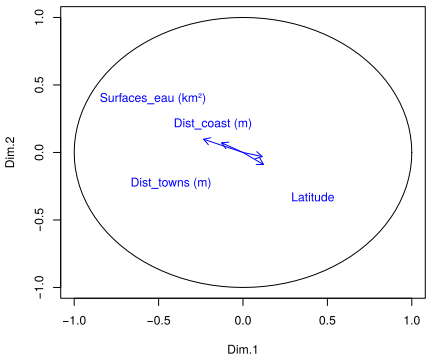
We can then look at the relations between the known traits and the interaction directions of $\hat{\Theta}$. Fig. 3a shows that the two first directions of interaction are correlated with the species covariates; the correlation is particularly high for the Nmass and SLA variables. Thus, on Fig. 2, the two directions separate the plants with large SLA and Nmass (top right corner) from those with small SLA and Nmass (bottom left corner). Then, Fig. 3a shows that the directions of interaction are also correlated with the environment covariates, and particularly with the mean snowmelt date (Snow). Thus, on Fig. 2, the two directions separate the late melting environments (top right corner) from the early melting environments (bottom left corner).

Combining the interpretation of Figs. 2, 3a and 3b, we deduce that plants with large Nmass and SLA interact highly with late melting sites (large value of Snow). This was in fact the main result obtained in the original study by Choler [10] (see, e.g., the summary of findings in the abstract), which advocates the good properties of LORI in terms of interpretation.

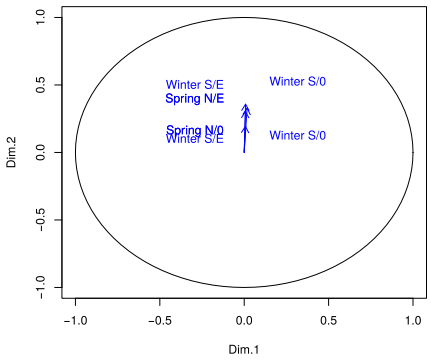
6. Using covariates to impute ecological data

The waterbirds data count the abundance of migratory water-birds in 785 wetland sites (across the five countries in North Africa), between 1990 and 2017 [43]. 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, meteorological anomalies, latitude and longitude. The count table contains a large amount of missing entries (70%), but the covariate matrices which contain respectively six covariates about the 785 sites and eight covariates about the 18 years, are fully observed. Our method allows to take advantage of the available covariates to provide interpretation for spatio-temporal patterns. As a by-product, it produces an imputed contingency table.

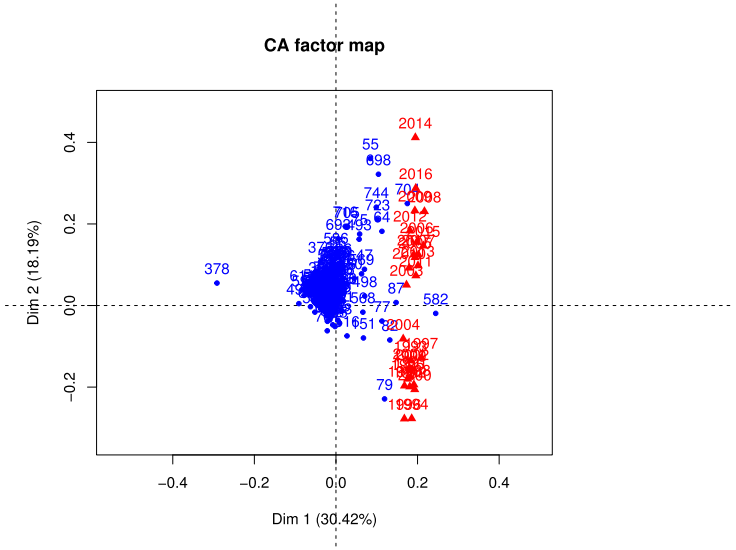
Tables 5 and 6 show the estimated main effects of some of the sites and years characteristics. Sites with large latitudes are associated to smaller counts, as well as sites which are far from the coast. Sites which are located far from towns, and sites with large water surfaces are associated to larger counts. The four year covariates given in Table 6 concern meteorological anomalies. The associated coefficients are all negative, indicating that more important anomalies are associated to smaller abundances.



(a) Correlation between sites characteristics and the two first directions of interaction.



(b) Correlation between year characteristics and the two first directions of interaction.



(c) Display of the two first dimensions of interaction estimated with LORI. Environments are represented with blue points and years with red triangles.

Fig. 4. Visual display of LORI results for the water-birds data.

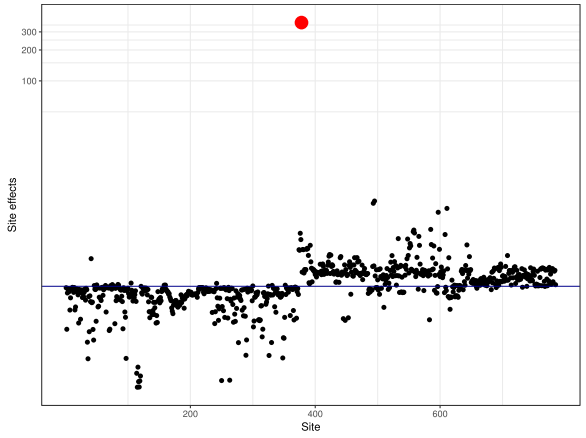
The sites and years can also be displayed using the same visual tools as described in Section 5. Figs. 4a and 4b show the correlations between the covariates and the directions of interaction. On the two-dimensional display on Fig. 4c, the first dimension is correlated with geographical characteristics, and the second dimension with meteorological anomalies. We observe a very clear temporal gradient along the second dimension, indicating that over time, meteorological anomalies increase (in the sense of a summary anomaly variable embodied by the second direction). One of the site (378) lays out of the point cloud, and corresponds to a site with very large surface.

LORI also returns counts estimates which can be used to compute an estimation of the total yearly abundances (i.e. counts estimates summed across sites). To better assess the temporal trend, one can decompose the estimated counts into three factors corresponding to the site effects, year effects and interactions, respectively. Indeed, for $(i, j) \in [n] \times [p]$, one can write

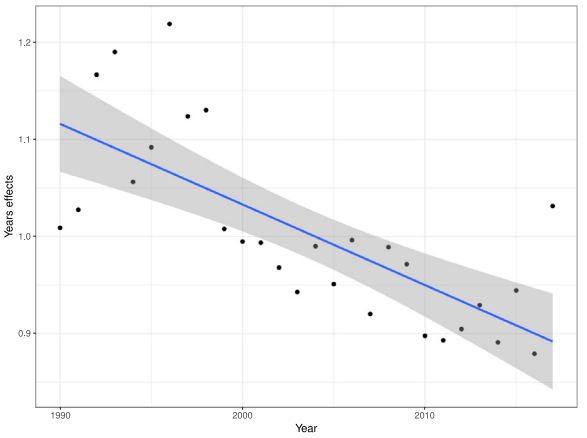
$$\exp(\hat{X}_{ij}) = \exp(\hat{\mu}) \exp(R_{i,\cdot}\hat{\alpha}) \exp(C_{j,\cdot}\hat{\beta}) \exp(\hat{\Theta}_{ij}).$$

Fig. 5 shows the last three factors of this decomposition separately.

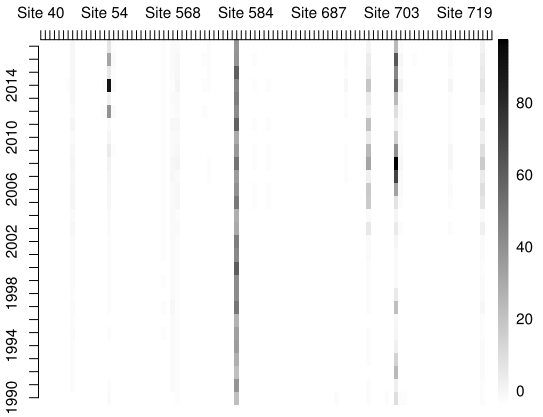
On Fig. 5a we see that most sites have multiplicative effects around 1 on the count scale. One site (site 378, large red dot) stands out; again, it corresponds to an extremely large site (6000 km^2, five times larger than the second, 300 times larger than the mean). In this respect, the row effects act as normalization factors accounting for surface. We also observe tenuous levels along the x-axis, corresponding to sites of different countries. On Fig. 5b we observe a decreasing temporal trend. This means that, all other things being equal, later years tend to produce smaller abundances. As illustrated in Fig. 4b, this temporal trend can be associated with the effects of meteorological anomalies. Note that the temporal effects



(a) Total site effects in count scale ($\exp(R_i\hat{\alpha})$, for $i \in \{1, \ldots, n\}$). One site has $\exp(R_i\hat{\alpha}) \geq \exp(5)$ (large red point). Horizontal line: $\exp(R_i\alpha) = 1$.



(b) Total year effects in count scale ($\exp(C_j\hat{\beta})$, for $j \in \{1, \ldots, p\}$). Blue line: LOESS (standard deviation in gray).



(c) Interaction in count scale ($\exp(\hat{\Theta}_{ij})$) for 30 sites (to improve the display, the rest of the sites take value extremely close to 1).

Fig. 5. Decomposition of the estimated counts into multiplicative site effects (top left), year effects (top right) and interactions (bottom).

(top right) are smaller in amplitude than the spatial effects (top left). Indeed, more variability is observed between sites in a given year, than between years for a given site.

Finally, looking at the interaction matrix on Fig. 5c, we see that the interaction is mainly driven by a few sites which interact more or less highly with every year. In particular site 582 presents large interactions with every year, and corresponds to Al Massira dam in Morocco and is the most abundant site (4 million birds in total, twice as much as the second most abundant, 120 more than the mean). Here, the large abundance is not explained by the geographical covariates available (but maybe by other unmeasured factors such as protection legislations), and thus the extreme abundance is captured in the interaction rather than the main effects. This profile was also visible on Fig. 4c where Site 582 lies within the cloud of years, indicating large interactions through small Euclidean distances.

Site 704 also presents large interactions, but they are not constant throughout the years. It corresponds to Ichkeul National Park in Tunisia, which is a major site for most species; the abundances are very large in Ichkeul compared to other sites. For several years including 2007, however, bad weather conditions prevented ornithologists from counting birds correctly; thus, reported counts are significantly lower than expected. This explains the drop in the interaction in 2007 for Ichkeul, corresponding to an outlier behavior. Again, such a profile could not be highlighted without modeling interactions. This illustrates one of the advantages of LORI for such bird abundance data compared to state-of-the-art

methods such as those of Pannekoek and van Strien [39] which do not model interactions. In particular, in most cases the interaction terms absorb outlying values (small or large), and indirectly account for the over-dispersion which is known to occur in birds abundance data.

7. Discussion

We conclude by discussing some opportunities for further research. First, 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 zero and a small number of large interactions are selected. Second, it would be useful to develop a multiple imputation procedure based on LORI, to provide confidence regions for the estimated parameters. 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.

Acknowledgments

The authors thank Trevor Hastie, Edgar Dobriban, Olga Klopp, Kevin Bleakey and Stéphane Dray for their very helpful comments on this manuscript. We would like to thank all the organizations involved in the Mediterranean Waterbirds Network (MNW) who provided the waterbird data set used in this article: the GREPOM/BirdLife Morocco, the Direction générale des forêts (Algeria), the AAO/BirdLife Tunisia, the Libyan Society for Birds, the Egyptian Environment Affairs Agency, the Office national de la chasse et de la faune sauvage (ONCFS, France) and the Institut de recherche pour la conservation des zones humides méditerranéennes de la Tour du Valat (France). We are also grateful to all the field observers who participated in the North African IWC, thereby making this data set so rich, and to Pierre Defos du Rau and Laura Dami for their help. This work was funded by the École Polytechnique Data Science Initiative and the Swiss National Science Foundation.

Appendix. Proofs

Proof of Theorem 1

We will first derive an upper bound for $\sum_{(i,j)\in\Omega}(\hat{X}_{ij} - X^*_{ij})^2$, then control $\|\hat{X} - X^*\|_F^2$ by

$$\|\hat{X} - X^*\|_F^2 \leq \sum_{(i,j)\in\Omega} (\hat{X}_{ij} - X^*_{ij})^2 + \mathsf{D},$$

with D a residual term defined later on. By definition of $\hat{X} = \hat{X}_0 + \hat{\Theta}$, $\mathcal{L}(\hat{X}) + \lambda\|\hat{\Theta}\|_* \leq \mathcal{L}(X^*) + \lambda\|\Theta^*\|_*$. Using the strong convexity of $\mathcal{L}$ and subtracting $\langle\nabla\mathcal{L}(X^*), \hat{X} - X^*\rangle$ on both sides of this inequality, we obtain

$$\sigma_-^2 \sum_{(i,j)\in\Omega} (\hat{X}_{ij} - X^*_{ij})^2/2 \leq \underbrace{-\langle\nabla\mathcal{L}(X^*), \hat{X} - X^*\rangle}_{\mathrm{I}} + \underbrace{\lambda(\|\Theta^*\|_* - \|\hat{\Theta}\|_*)}_{\mathrm{II}}. \tag{A.1}$$

We will bound separately the two terms on the right-hand side of (A.1).

Given a matrix $X \in \mathbb{R}^{n\times p}$, we denote $\mathcal{S}_1(X)$ (resp. $\mathcal{S}_2(X)$) the span of left (resp. right) singular vectors of X. Let $P^{\perp}_{\mathcal{S}_1(X)}$ (resp. $P^{\perp}_{\mathcal{S}_2(X)}$) be the orthogonal projector in $\mathbb{R}^n$ on $\mathcal{S}_1(X)^{\perp}$ (resp. in $\mathbb{R}^p$ on $\mathcal{S}_2(X)^{\perp}$). We define the projection operator $\mathcal{P}_X^{\perp} : \tilde{X} \mapsto P^{\perp}_{\mathcal{S}_1(X)}\tilde{X}P^{\perp}_{\mathcal{S}_2(X)}$ in $\mathbb{R}^{n\times p}$, and $\mathcal{P}_X : \tilde{X} \mapsto \tilde{X} - P^{\perp}_{\mathcal{S}_1(X)}\tilde{X}P^{\perp}_{\mathcal{S}_2(X)}$. Using the fact that

$$|\langle\nabla\mathcal{L}(X^*), \hat{X} - X^*\rangle| \leq \|\hat{X} - X^*\|_*\|\nabla\mathcal{L}(X^*)\|$$

and the triangle inequality, we get

$$\mathrm{I} \leq \|\nabla\mathcal{L}(X^*)\| \left(\|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\|_* + \|\mathcal{P}^{\perp}_{\Theta^*}(\hat{\Theta} - \Theta^*)\|_* + \|\hat{X}_0 - X^*_0\|_*\right). \tag{A.2}$$

Then, Lemma A(ii), applied to $\hat{\Theta}$ and Θ^*, results in

$$\mathrm{II} \leq \lambda\{\|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\|_* - \|\mathcal{P}^{\perp}_{\Theta^*}(\hat{\Theta} - \Theta^*)\|_*\}. \tag{A.3}$$

Plugging inequalities (A.2) and (A.3) into (A.1), we obtain

$$\sigma_-^2 \sum_{(i,j)\in\Omega} (\hat{X}_{ij} - X^*_{ij})^2 \le 2\{\lambda + \|\nabla\mathcal{L}(X^*)\|\} \times \|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\|_* + 2\{\|\nabla\mathcal{L}(X^*)\| - \lambda\} \times \|\mathcal{P}^{\perp}_{\Theta^*}(\hat{\Theta} - \Theta^*)\|_* + 2\|\nabla\mathcal{L}(X^*)\| \times \|\hat{X}_0 - X^*_0\|_*. \tag{A.4}$$

We now use the condition $\lambda \ge 2\,\|\nabla\mathcal{L}(X^*)\|$ in (A.4) to deduce that

$$\sigma_-^2 \sum_{(i,j)\in\Omega} (\hat{X}_{ij} - X^*_{ij})^2 \le 3\lambda\|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\|_* + \lambda\|\hat{X}_0 - X^*_0\|_*.$$

Furthermore, we have that $P_1(\hat{X}_0 - X^*_0)P_2^\top = P_1\hat{X}_0P_2^\top - P_1X^*_0P_2^\top = 0$, which implies that $\hat{X}_0 - X^*_0 \in \mathcal{X}_0$ and, by definition of r, $\mathrm{rk}(\hat{X}_0 - X^*_0) \le r$.

In addition to the fact that $\|\hat{X}_0 - X^*_0\|_F \le \|\hat{X} - X^*\|_F$, and given that $\hat{X}_0 - X^*_0$ is the orthogonal projection of $\hat{X} - X^*$ on $\mathcal{X}_0$, we find $\|\hat{X}_0 - X^*_0\|_* \le \sqrt{r}\,\|X^* - \hat{X}\|_F$, which together with Lemma A(iii) and $\|\hat{\Theta} - \Theta^*_0\|_F \le \|\hat{X} - X^*\|_F$ (also by orthogonal projection of $\hat{X} - X^*$, this time on $\mathcal{T}$) yields

$$\sigma_-^2 \sum_{(i,j)\in[n]\times[p]} \omega_{ij}(\hat{X}_{ij} - X^*_{ij})^2 \le \lambda\left\{3\sqrt{2\,\mathrm{rank}(\Theta^*)} + \sqrt{r}\right\}\|\hat{X} - X^*\|_F. \tag{A.5}$$

We now derive the upper bound

$$\|\hat{X} - X^*\|_F^2 \le \sum_{(i,j)\in[n]\times[p]} \omega_{ij}(\hat{X}_{ij} - X^*_{ij})^2 + \mathsf{D}.$$

Define $\eta = 72\ln(n+p)/\{\pi\ln(6/5)\}$,

$$\Sigma(\omega, X) = \sum_{(i,j)\in[n]\times[p]} \omega_{ij}X_{ij}^2$$

and the set

$$\mathcal{C}(\eta,\rho) = \left\{X \in \mathbb{R}^{n\times p} : \|X\|_\infty \le 1,\, \|X\|_* \le \sqrt{\rho}\,\|X\|_F,\, \mathrm{E}\left[\Sigma(\omega, X)\right] > \eta\right\}.$$

Lemma B shows that whenever $\hat{X} - X^*$ belongs to $\mathcal{C}(\eta,\rho)$ (for ρ and D defined later on), a restricted strong convexity property of the form

$$\|\hat{X} - X^*\|_F^2 \le \sum_{(i,j)\in[n]\times[p]} \omega_{ij}(\hat{X}_{ij} - X^*_{ij})^2 + \mathsf{D}$$

holds. We consider the following two cases.

Case 1. If $\sum_{(i,j)\in[n]\times[p]} \pi_{ij}(\hat{X}_{ij} - X^*_{ij})^2 \le \eta$, then $\|\hat{X} - X^*\|_2^2 \le \eta/\pi$ and the result of Theorem 1 (11) is proved.

Case 2. If $\sum_{(i,j)\in[n]\times[p]} \pi_{ij}(\hat{X}_{ij} - X^*_{ij})^2 > \eta$, we show that $(\hat{X} - X^*)/2\gamma \in \mathcal{C}\{\eta,\, 16\,\mathrm{rank}(\Theta^*) + 8r\}$. Using (A.4), $\sigma_-^2 \sum_{(i,j)\in\Omega}(\hat{X}_{ij} - X^*_{ij})^2 \ge 0$ and $\|\nabla\mathcal{L}(X^*)\| \le \lambda/2$, we obtain that

$$\|\mathcal{P}^{\perp}_{\Theta^*}(\hat{\Theta} - \Theta^*)\|_* \le 3\,\|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\|_* + \|\hat{X}_0 - X^*_0\|_*.$$

Furthermore,

$$\begin{aligned}\|\hat{X} - X^*\|_* &\le \|\mathcal{P}^{\perp}_{\Theta^*}(\hat{\Theta} - \Theta^*)\|_* + \|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\|_* + \|\hat{X}_0 - X^*_0\|_* \\ &\le 4\,\|\mathcal{P}_{\Theta^*}(\hat{\Theta} - \Theta^*)\|_* + 2\,\|\hat{X}_0 - X^*_0\|_* \\ &\le 2\sqrt{2\,\mathrm{rank}(\Theta^*)}\,\|\hat{\Theta} - \Theta^*\|_F + 2\sqrt{r}\,\|\hat{X}_0 - X^*_0\|_F.\end{aligned}$$

Using the fact that, because $\hat{X}_0 - X^*_0$ and $\hat{\Theta} - \Theta^*$ are both orthogonal projections of $\hat{X} - X^*$, $\|\hat{\Theta} - \Theta^*\|_F \le \|\hat{X} - X^*\|_F$ and $\|\hat{X}_0 - X^*_0\|_F \le \|\hat{X} - X^*\|_F$, we deduce that

$$\|\hat{X} - X^*\|_* \le (2\sqrt{2\,\mathrm{rank}(\Theta^*)} + 2\sqrt{r}) \times \|\hat{X} - X^*\|_F.$$

Then, using the identity $2ab \le a^2 + b^2$, we have

$$2\sqrt{2\,\mathrm{rank}(\Theta^*)} + 2\sqrt{r} = \sqrt{8\,\mathrm{rank}(\Theta^*) + 4r + 8\sqrt{2}\sqrt{\mathrm{rank}(\Theta^*)}\sqrt{r}} \le \sqrt{16\,\mathrm{rank}(\Theta^*) + 8r}.$$

Finally, we conclude that $\|\hat{X}-X^*\|_* \le \sqrt{16\,\mathrm{rank}(\Theta^*)+8r}\,\|\hat{X}-X^*\|_F$, and $(\hat{X}-X^*)/2\gamma \in \mathcal{C}\{\eta, 16\,\mathrm{rank}(\Theta^*)+8r\}$. Thus, Lemma B implies that with probability at least $1-8/(n+p)$,

$$\sum_{(i,j)\in[n]\times[p]} \omega_{ij}(\hat{X}_{ij}-X^*_{ij})^2 \ge \frac{1}{2}\,\mathrm{E}\left\{\sum_{(i,j)\in[n]\times[p]} \omega_{ij}(\hat{X}_{ij}-X^*_{ij})^2\right\} - 384\gamma^2[\{16\,\mathrm{rank}(\Theta^*)+8r\}(\mathrm{E}\|\Sigma_R\|)^2+8]/\pi\,. \tag{A.6}$$

Combining (A.6) and (A.5) we obtain

$$\begin{aligned}&\pi\|\hat{X}-X^*\|_F^2/2 - 384\gamma^2[\{16\,\mathrm{rank}(\Theta^*)+8r\}(\mathrm{E}\|\Sigma_R\|)^2+8]/\pi\\ &\quad\le \lambda\left(3\sqrt{2\,\mathrm{rank}(\Theta^*)}+\sqrt{r}\right)\|\hat{X}-X^*\|_F/\sigma_-^2.\end{aligned}$$

Moreover, using the identity $ab \le a^2+b^2/4$, we obtain

$$\lambda\left(3\sqrt{2\,\mathrm{rank}(\Theta^*)}+\sqrt{r}\right)\|\hat{X}-X^*\|_F/\sigma_-^2 \le \lambda^2\left(3\sqrt{2\,\mathrm{rank}(\Theta^*)}/(\pi\sigma_-^4)+\sqrt{r}\right)^2 + \pi\|\hat{X}-X^*\|_F^2/4.$$

Using the identity $2ab \le a^2+b^2$, we also have

$$\left(3\sqrt{2\,\mathrm{rank}(\Theta^*)}+\sqrt{r}\right)^2 \le 36\,\mathrm{rank}(\Theta^*)+2r.$$

Finally, we obtain the result, viz.

$$\|\hat{X}-X^*\|_F^2 \le \left\{144\lambda^2/(\pi^2\sigma_-^4)+24576\gamma^2(\mathrm{E}\|\Sigma_R\|)^2/\pi^2\right\}\{\mathrm{rk}(\Theta^*)+r\}+12288/\pi^2. \tag{A.7}$$

Lemma A. *The following statements hold for all M and M' in $\mathbb{R}^{n\times p}$.*

(i) $\|M+\mathcal{P}_M^\perp(M)\|_* = \|M\|_* + \|\mathcal{P}_M^\perp(M)\|_*$.
(ii) $\|M\|_* - \|M'\|_* \le \|\mathcal{P}_M(M-M')\|_* - \|\mathcal{P}_M^\perp(M-M')\|_*$.
(iii) $\|\mathcal{P}_M(M-M')\|_* \le \sqrt{2\mathrm{rk}(M)}\|M-M'\|_F$.

Proof. See, e.g., Lemma 16 in [31]. □

Lemma B. *Let $\eta = 72\ln(n+p)/\{\pi\ln(6/5)\}$ and $\rho>0$ and define*

$$\varsigma = 96\{\rho(\mathrm{E}\|\Sigma_R\|)^2+8\}/\pi\,. \tag{A.8}$$

With probability at least $1-8/(n+p)$, for all $X\in\mathcal{C}(\eta,\rho)$ we get

$$|\Sigma(\omega,X)-\mathrm{E}\{\Sigma(\omega,X)\}| \le \mathrm{E}\{\Sigma(\omega,X)\}/2+\varsigma,$$

with Σ_R defined in (10).

Proof. Consider the event

$$\mathcal{B} = \left\{\sup_{X\in\mathcal{C}(\eta,\rho)}\left[|\Sigma(\omega,X)-\mathrm{E}\{\Sigma(\omega,X)\}| - \mathrm{E}\{\Sigma(\omega,X)\}/2\right] > \varsigma\right\}.$$

Define also, for $\ell\in\mathbb{N}_*$,

$$\mathcal{S}_\ell = \left\{X\in\mathcal{C}(\eta,\rho) : \kappa^{\ell-1}\eta < \mathrm{E}\{\Sigma(\omega,X)\} < \kappa^\ell\eta\right\},$$

for $\kappa=6/5$ and $\eta = 72\ln(n+p)/\{\pi\ln(6/5)\}$. On $\mathcal{B}$, there exist $\ell\ge1$ and $X\in\mathcal{C}(\eta,\rho)$ such that $X\in\mathcal{C}(\eta,\rho)\bigcap\mathcal{S}_\ell$, and

$$|\Sigma(\omega,X)-\mathrm{E}\{\Sigma(\omega,X)\}| > \mathrm{E}\{\Sigma(\omega,X)\}/2+\varsigma > \kappa^{\ell-1}\eta/2+\varsigma = 5\kappa^\ell\eta/12+\varsigma. \tag{A.9}$$

For $T>0$, define the set

$$\mathcal{C}(\eta,\rho,T) = \{X\in\mathcal{C}(\eta,\rho) : \mathrm{E}\{\Sigma(\omega,X)\}\le T\}$$

and the event

$$\mathcal{B}_l = \left\{\sup_{X\in\mathcal{C}(\eta,\rho,\kappa^\ell\eta)} |\Sigma(\omega,X)-\mathrm{E}\{\Sigma(\omega,X)\}| > 5\kappa^\ell\eta/12+\varsigma\right\}.$$

It follows from (A.9) that $\mathcal{B}\subset\bigcup_{\ell=1}^\infty\mathcal{B}_\ell$; thus, it is enough to estimate the probability of the events $\mathcal{B}_\ell$ for $\ell\in\mathbb{N}$, and then apply the union bound. Such an estimation is given in the following lemma, adapted from Lemma 10 in [28]. Define

$$Z_T = \sup_{X\in\mathcal{C}(\eta,\rho,T)} |\Sigma(\omega,X)-\mathrm{E}\{\Sigma(\omega,X)\}|\,.$$

Lemma C implies that

$$\Pr(\mathcal{B}) \leq \sum_{\ell=1}^{\infty} \Pr(\mathcal{B}_\ell) \leq 4 \sum_{\ell=1}^{\infty} \exp(-\pi \kappa^\ell \eta/72) \leq 8/(n+p),$$

which concludes the proof. □

Lemma C. *Under the assumptions of Theorem 1,*

$$\Pr\left(Z_T \geq 5T/12 + \varsigma\right) \leq 4e^{-\pi T/72}, \tag{A.10}$$

where ς is defined in (A.8).

Proof. We use the following concentration inequality due to Talagrand and a symmetrization argument. Recall the statement of Talagrand's concentration inequality. Let $f : [-1,1]^m \mapsto \mathbb{R}$ be a convex Lipschitz function with Lipschitz constant L, $\Xi_1, \ldots, \Xi_m$ be independent random variables taking values in $[-1,1]$, and $Z = f(\Xi_1, \ldots, \Xi_m)$. Then, for any $t \geq 0$, $\Pr\{|Z - \mathrm{E}(Z)| \geq 16L + t\} \leq 4e^{-t^2/2L^2}$. For $\mathbf{x} = (x_{ij})$, $(i,j) \in [n] \times [p]$, we apply this result to the function

$$f(\mathbf{x}) = \sup_{X \in \mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j)\in[n]\times[p]} (x_{ij} - \pi_{ij}) X_{ij}^2 \right|,$$

which is Lipschitz with Lipschitz constant $\sqrt{T/\pi}$. We find

$$\begin{aligned}
&|f(x_{11}, \ldots, x_{np}) - f(z_{11}, \ldots, z_{np})| \\
&\quad = \left| \sup_{X \in \mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j)\in[n]\times[p]} (x_{ij} - \pi_{ij}) X_{ij}^2 \right| - \sup_{X \in \mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j)\in[n]\times[p]} (z_{ij} - \pi_{ij}) X_{ij}^2 \right| \right| \\
&\quad \leq \sup_{X \in \mathcal{C}(\eta,\rho,T)} \left| \left| \sum_{(i,j)\in[n]\times[p]} (x_{ij} - \pi_{ij}) X_{ij}^2 \right| - \left| \sum_{(i,j)\in[n]\times[p]} (z_{ij} - \pi_{ij}) X_{ij}^2 \right| \right| \\
&\quad \leq \sup_{X \in \mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j)\in[n]\times[p]} (x_{ij} - \pi_{ij}) X_{ij}^2 - \sum_{(i,j)\in[n]\times[p]} (z_{ij} - \pi_{ij}) X_{ij}^2 \right| \\
&\quad \leq \sup_{X \in \mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j)\in[n]\times[p]} (x_{ij} - z_{ij}) X_{ij}^2 \right| \\
&\quad \leq \sup_{X \in \mathcal{C}(\eta,\rho,T)} \sqrt{(i,j) \in \sum_{[n]\times[p]} (x_{ij} - z_{ij})^2/\pi_{ij}} \sqrt{\sum_{(i,j)\in[n]\times[p]} \pi_{ij} X_{ij}^4} \\
&\quad \leq \sup_{X \in \mathcal{C}(\eta,\rho,T)} \sqrt{\pi^{-1}} \sqrt{\sum_{(i,j)\in[n]\times[p]} (x_{ij} - z_{ij})^2} \sqrt{\sum_{(i,j)\in[n]\times[p]} \pi_{ij} X_{ij}^2} \\
&\quad \leq \sqrt{T/\pi} \sqrt{\sum_{(i,j)\in[n]\times[p]} (x_{ij} - z_{ij})^2},
\end{aligned}$$

where we have used $||a| - |b|| \leq |a - b|$, $\|X\|_\infty \leq 1$ and $\mathrm{E}\{\Sigma(\omega, X)\} \leq T$. Thus, Talagrand's inequality and the identity $\sqrt{T/\pi} \leq T/(2 \times 96) + 96/(2\pi)$ together yield

$$\Pr\left\{Z_T \geq \mathrm{E}(Z_T) + 768/\pi + T/12 + t\right\} \leq 4e^{-t^2\pi/2T}.$$

Taking $t = T/6$ we get

$$\Pr\left\{Z_T \geq \mathrm{E}(Z_T) + 768/\pi + 3T/12\right\} \leq 4e^{-\pi T/72}. \tag{A.11}$$

Now we bound the expectation $\mathrm{E}(Z_T)$ using a symmetrization argument; see Section 7.2 in [32]. Let (ϵ_{ij}) be an iid Rademacher sequence. We have

$$\mathrm{E}(Z_T) \leq 2\mathrm{E}\left(\sup_{X \in \mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j)\in[n]\times[p]} \epsilon_{ij} \omega_{ij} X_{ij}^2 \right| \right),$$

Then, the contraction inequality (see Theorem 2.2 in [29]) yields

$$\mathrm{E}(Z_T) \leq 8\,\mathrm{E}\left(\sup_{X\in\mathcal{C}(\eta,\rho,T)}\left|\sum_{(i,j)\in[n]\times[p]}\epsilon_{ij}\omega_{ij}X_{ij}\right|\right) = 8\,\mathrm{E}\left(\sup_{X\in\mathcal{C}(\eta,\rho,T)}|\langle\Sigma_R, X\rangle|\right),$$

where Σ_R is defined in (10). For $X \in \mathcal{C}(\eta, \rho, T)$ we have that $\|X\|_* \leq \sqrt{\rho T/\pi}$. Then, by duality between the nuclear and operator norms, we obtain

$$\mathrm{E}(Z_T) \leq 8\,\mathrm{E}\left(\sup_{\|X\|_*\leq\sqrt{\rho T/\pi}}|\langle\Sigma_R, X\rangle|\right) \leq 8\sqrt{\rho T/\pi}\,\mathrm{E}\|\Sigma_R\|.$$

Combined with (A.11) and using $8\sqrt{\rho T/\pi}\,\mathrm{E}\|\Sigma_R\| \leq T/6 + 3\times 8^2\rho(\mathrm{E}\|\Sigma_R\|)^2/(2\pi)$, we finally obtain (A.10) using the definition of ς in (A.8). □

Proof of Theorem 2

Theorem 2 derives from Theorem 1 and combining the two following steps: (i) computing a value of λ such that the condition $\lambda \geq 2\,\|\nabla\mathcal{L}(X^*)\|$ holds with high probability and (ii) controlling $\mathrm{E}\,\|\Sigma_R\|$. Let us start with (i). Define the random matrices $Z_{ij} = \omega_{ij}\{-Y_{ij} + \exp(X^*_{ij})\}E_{ij}$ and the quantity

$$\sigma_Z^2 = \max\left(\frac{1}{np}\left\|\sum_{i=1}^n\sum_{j=1}^p \mathrm{E}(Z_{ij}Z_{ij}^\top)\right\|,\ \frac{1}{np}\left\|\sum_{i=1}^n\sum_{j=1}^p \mathrm{E}(Z_{ij}^\top Z_{ij})\right\|\right).$$

Lemma D. *Under the assumptions of Theorem 2,*

$$\frac{\sigma_-^2\beta}{np} \leq \sigma_Z^2 \leq \frac{\sigma_+^2\beta}{np}. \tag{A.12}$$

Proof. For all $(i, j) \in [n]\times[p]$, $Z_{ij}Z_{ij}^\top = \omega_{ij}\{-Y_{ij}+\exp(X^*_{ij})\}^2E_{ij}E_{ij}^\top$, and $\mathrm{E}(Z_{ij}Z_{ij}^\top) = \mathrm{E}(\omega_{ij})\mathrm{E}[\{-Y_{ij}+\exp(X^*_{ij})\}^2]E_{ij}E_{ij}^\top$, which is a diagonal matrix with 0 everywhere except for the ith element of its diagonal, where its value is $\mathrm{E}(\omega_{ij})\mathrm{E}[\{-Y_{ij}+\exp(X^*_{ij})\}^2]$. Thus

$$\sum_{(i,j)\in[n]\times[p]}\mathrm{E}(Z_{ij}Z_{ij}^\top)$$

is also a diagonal matrix, and the ith element of its diagonal is $\sum_{j=1}^p \mathrm{E}(\omega_{ij})\mathrm{E}[\{-Y_{ij} + \exp(X^*_{ij})\}^2]$. We obtain

$$\frac{1}{np}\left\|\sum_{(i,j)\in[n]\times[p]}\mathrm{E}(Z_{ij}Z_{ij}^\top)\right\| = \frac{1}{np}\max_{i\in[n]}\sum_{j=1}^p \mathrm{E}(\omega_{ij})\mathrm{E}[\{-Y_{ij} + \exp(X^*_{ij})\}^2].$$

Using the fact that $\mathrm{E}(Y_{ij}) = \exp(X^*_{ij})$ and $\sigma_-^2 \leq \mathrm{var}(Y_{ij}) \leq \sigma_+^2$, we then get

$$\frac{\sigma_-^2}{np}\max_{i\in[n]}\sum_{j=1}^p \mathrm{E}(\omega_{ij}) \leq \frac{1}{np}\left\|\sum_{(i,j)\in[n]\times[p]}\mathrm{E}(Z_{ij}Z_{ij}^\top)\right\| \leq \frac{\sigma_+^2}{np}\max_{i\in[n]}\sum_{j=1}^p \mathrm{E}(\omega_{ij}). \tag{A.13}$$

Using the same arguments, we also obtain

$$\frac{\sigma_-^2}{np}\max_{j\in[p]}\sum_{i=1}^n \mathrm{E}[\omega_{ij}] \leq \frac{1}{np}\left\|\sum_{(i,j)\in[n]\times[p]}\mathrm{E}[Z_{ij}^\top Z_{ij}]\right\| \leq \frac{\sigma_+^2}{np}\max_{j\in[p]}\sum_{i=1}^n \mathrm{E}[\omega_{ij}].$$

Combining (A.13) and (A.13), we deduce that

$$\frac{\sigma_-^2}{np}\max\left\{\max_{i\in[n]}\sum_{j=1}^p \mathrm{E}(\omega_{ij}), \max_{j\in[p]}\sum_{i=1}^n \mathrm{E}(\omega_{ij})\right\} \leq \sigma_Z^2 \leq \frac{\sigma_+^2}{np}\max\left\{\max_{i\in[n]}\sum_{j=1}^p \mathrm{E}(\omega_{ij}), \max_{j\in[p]}\sum_{i=1}^n \mathrm{E}(\omega_{ij})\right\},$$

which concludes the proof. □

Note that $\mathrm{E}(Z_{ij}) = 0$ for all $(i,j) \in [n] \times [p]$ and $\nabla \mathcal{L}(X^*) = \sum_{i=1}^n \sum_{j=1}^p Z_{ij}$. We use an extension of Theorem 4 in [30] to rectangular matrices via self-adjoint dilation; see, e.g., Example 2.6 in [44]. Let $\Xi_1, \ldots, \Xi_m$ be m independent $(n \times p)$-matrices satisfying $\mathrm{E}(\Xi_i) = 0$ and

$$\inf\{K > 0 : \mathrm{E}\{\exp(\|\Xi_i\|/K)\} \le e\} < M$$

for some constant M and for all $i \in \{1, \ldots, m\}$. Define

$$\sigma^2 = \max\left\{ \frac{1}{m} \left\| \sum_{i=1}^m \mathrm{E}(\Xi_i \Xi_i^\top) \right\|, \quad \frac{1}{m} \left\| \sum_{i=1}^m \mathrm{E}(\Xi_i^\top \Xi_i) \right\| \right\},$$

and $\bar{U} = M \ln(1 + 2M^2/\sigma^2)$. Then, for $t\bar{U} \le 2(e-1)\sigma^2 m$,

$$\Pr\left(\left\| \frac{1}{m} \sum_{i=1}^m \Xi_i \right\| \ge t \right) \le 2(n+p) \exp\left(-\frac{t^2}{4m\sigma^2 + 2\bar{U}t/3} \right)$$

and for $t\bar{U} > 2(e-1)\sigma^2 m$,

$$\Pr\left(\left\| \frac{1}{m} \sum_{i=1}^m \Xi_i \right\| \ge t \right) \le 2(n+p) \exp\left\{ -\frac{t}{(e-1)\bar{U}} \right\}.$$

Under Assumption 3 we may apply this result with $m = np$, $(\Xi_1, \ldots, \Xi_m) = (Z_{11}, \ldots, Z_{np})$, $M = 2\delta$, $\sigma^2 = \sigma_Z^2$ and $\bar{U} = 2\delta \ln(1 + 8\delta^2/\sigma_Z^2)$. Indeed, the matrices Z_{ij}, $i \in [n]$, $j \in [p]$ are independent: the entries Y_{ij} are independent, as well as the Bernoulli random variables ω_{ij}; furthermore, the random variables ω_{ij} are also independent of Y. Taking

$$t \ge \max\left\{ 2\sigma_Z \sqrt{3np \ln(n+p)}, 6\delta(e-1) \ln(1 + 8\delta^2/\sigma_Z^2) \ln(n+p) \right\}$$

and using Lemma D, we get that with probability at least $1 - 1/(n+p)$,

$$\|\nabla \mathcal{L}(X^*)\| \le \max\left[2\sigma_+ \{3\beta \ln(n+p)\}^{1/2}, 6\delta(e-1) \ln\{1 + 8\delta^2 np/(\beta\sigma_-^2)\} \ln(n+p) \right].$$

Thus, taking λ as in Theorem 2 ensures that $\lambda \ge 2\|\nabla \mathcal{L}(X^*)\|$ with probability at least $1 - 1/(n+p)$.

We now turn to (ii) and control $\mathrm{E}\,\|\Sigma_R\|$ with the following lemma.

Lemma E. *There exists an absolute constant C^* such that the following inequality holds:*

$$\mathrm{E}(\|\Sigma_R\|) \le C^* \left(\sqrt{\beta} + \sqrt{\ln m} \right).$$

The proof of this lemma depends on the following extension to rectangular matrices via self-adjoint dilation of Corollary 3.3 in [4].

Proposition A. *Let A be an $n \times p$ matrix with A_{ij} independent centered bounded random variables. Then, there exists a universal constant C^* such that*

$$\mathrm{E}(\|A\|) \le C^* \left\{ \sigma_1 \vee \sigma_2 + \sigma_* \sqrt{\ln(n \wedge p)} \right\},$$

$$\sigma_1 = \max_{i \in [n]} \sqrt{\sum_{j=1}^p \mathrm{E}(A_{ij}^2)}, \quad \sigma_2 = \max_{j \in [p]} \sqrt{\sum_{i=1}^n \mathrm{E}(A_{ij}^2)}, \quad \sigma_* = \max_{i \in [n], j \in [p]} |A_{ij}|.$$

Proof of Lemma E. Applying Proposition A to Σ_R with $\sigma_1 \vee \sigma_2 \le \sqrt{\beta}/|\Omega|$ and $\sigma_* \le 1$ we obtain

$$\mathrm{E}(\|\Sigma_R\|) \le C^* \left\{ \sqrt{\beta} + \sqrt{\ln(n \wedge p)} \right\}.$$

This completes the argument. □

Combining (i) and (ii) with (A.7) and a union bound argument, we obtain the conclusion of Theorem 2.

A.1. Proof of Theorem 3

In what follows, we denote for $X_0 \in \mathcal{X}_0$ and $\Theta \in \mathcal{T}$ $\mathcal{F}^\lambda(X_0, \Theta) = \mathcal{L}(X_0 + \Theta) + \lambda\|\Theta\|_*$. We establish below that $\lambda_0(Y)$ defined in (12) is equal to

$$\lambda_0(Y) = \min_\lambda \quad 0 \in \partial_\Theta \{\mathcal{F}^\lambda(\hat{X}_0, \Theta) + \chi_{\mathcal{T}}(\Theta)\}\Big|_{\Theta=0},$$

where for $\mathcal{K} \subset \mathbb{R}^{n\times p}$, $\chi_{\mathcal{K}}(X)$ is the characteristic function of the set $\mathcal{K}$, equal to 0 on $\mathcal{K}$ and $+\infty$ elsewhere, and $\hat{X}_0 = \operatorname{argmin}_{X\in\mathcal{X}_0} \mathcal{L}(X)$; see (12). The subdifferential of the objective function $\mathcal{F}^\lambda$ with respect to Θ is given by

$$\partial_\Theta \mathcal{F}^\lambda(\hat{X}_0, 0) = \nabla\mathcal{L}(\hat{X}_0 + \Theta)\Big|_{\Theta=0} + \lambda\partial_\Theta \|\Theta\|_* \Big|_{\Theta=0} + \partial_\Theta \chi_{\mathcal{T}}(\Theta)\Big|_{\Theta=0}.$$

$0 \in \partial_\Theta \chi_{\mathcal{T}}(\Theta)\,|_{\Theta=0}$. Lemma F ensures that $0 \in \partial\mathcal{F}^\lambda(\Theta)\,|_{\Theta=0}$ if and only if

$$0 \in \left\{\nabla\mathcal{L}(\hat{X}_0) + \lambda W : \|\mathcal{P}_{\mathcal{T}}(W)\| \le 1\right\}.$$

This is equivalent to $\lambda \ge |\mathcal{P}_{\mathcal{T}}\{\nabla\mathcal{L}(\hat{X}_0)\}||$. Additionally, at the optimum $\hat{X}_0$, we have $\mathcal{P}_{\mathcal{T}}\{\nabla\mathcal{L}(\hat{X}_0)\} = \nabla\mathcal{L}(\hat{X}_0)$, which concludes the proof.

Lemma F. *Let $g : \mathcal{T} \to \mathbb{R}_+$ be the function defined by $g(A) = \|A\|_*$ for $A \in \mathcal{T}$. Then*

$$\partial g(0) = \left\{W \in \mathbb{R}^{n\times p} : \|\mathcal{P}_{\mathcal{T}}(W)\| \le 1\right\}.$$

Proof. By definition of the subdifferential, we need to prove that for all $W \in \mathbb{R}^{n\times p}$, $\|\mathcal{P}_{\mathcal{T}}(W)\| < 1$, and for all $B \in \mathcal{T}$, $g(B) \ge g(0) + \langle W, B - 0\rangle$. First, $B \in \mathcal{T}$ implies $\langle W, B\rangle = \langle \mathcal{P}_{\mathcal{T}}(W), B\rangle$. Therefore, $\|\mathcal{P}_{\mathcal{T}}(W)\| \le 1$ is a sufficient condition for $W \in \partial g(0)$. Now assume that $\|\mathcal{P}_{\mathcal{T}}(W)\| > 1$ and let $\mathcal{P}_{\mathcal{T}}(W) = U\Sigma V^\top$, where U and V are orthogonal matrices of left and right singular vectors, and $\Sigma_{11} = \|\mathcal{P}_{\mathcal{T}}(W)\| > 1$. Define $B = U\tilde{\Sigma}V^\top$, $\tilde{\Sigma}_{11} = 1$ and $\tilde{\Sigma}_{ij} = 0$ elsewhere; note that with this definition, $B \in \mathcal{T}$. We have $g(B) = 1$ and $\langle \mathcal{P}_{\mathcal{T}}(W), B\rangle = \Sigma_{11} > g(B)$. Therefore, $\|\mathcal{P}_{\mathcal{T}}(W)\| > 1 \Rightarrow W \notin \partial g(0)$, from which we conclude. □

References

[1] J. Abernethy, F. Bach, T. Evgeniou, J.-P. Vert, A new approach to collaborative filtering: Operator estimation with spectral regularization, J. Mach. Learn. Res. 10 (2009) 803–826.
[2] D. Agarwal, B.-C. Chen, Regression-based latent factor models, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '09, ACM, New York, 2009, pp. 19–28.
[3] R. Angst, C. Zach, M. Pollefeys, The generalized trace-norm and its application to structure-from-motion problems, in: Proceedings of the 2011 International Conference on Computer Vision, in: ICCV '11, IEEE Computer Society, Washington, DC, 2011, pp. 2502–2509.
[4] A.S. Bandeira, R. van Handel, Sharp nonasymptotic bounds on the norm of random matrices with independent entries, Ann. Probab. 44 (2016) 2479–2506.
[5] C.J. ter Braak, P. Peres-Neto, S. Dray, A critical issue in model-based inference for studying trait-based community assembly and a solution, Peer J. 5 (2017) e2885.
[6] A.M. Brown, D.I. Warton, N.R. Andrew, M. Binns, G. Cassis, H. Gibb, The fourth-corner solution: Using predictive models to understand how species traits interact with the environment, Methods Ecol. Evol. 5 (2014) 344–352.
[7] J.-F. Cai, E.J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (2010) 1956–1982.
[8] Y. Cao, Y. Xie, Poisson matrix recovery and completion, IEEE Trans. Signal Process. 64 (2016) 1609–1620.
[9] J. Chiquet, M. Mariadassous, S. Robin, Variational inference for probabilistic Poisson PCA, Ann. Appl. Statist. 12 (2018) 2674–2698.
[10] P. Choler, Consistent shifts in Alpine plant traits along a mesotopographical gradient, Arctic Antarct. Alpine Res. 37 (2005) 444–453.
[11] R. Christensen, Log-Linear Models, Springer, New York, 2010.
[12] M. Collins, S. Dasgupta, R.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, in: NIPS'01, MIT Press, Cambridge, MA, 2001, pp. 617–624.
[13] I. Csiszár, G. Tusnády, Information geometry and alternating minimization procedures, Stat. Decis. Supplement Issue 1 (1984).
[14] D.L. Donoho, I.M. Johnstone, Ideal spatial adaptation via wavelet shrinkage, Biometrika 81 (1994) 425–455.
[15] B. Efron, Bootstrap methods: Another look at the jackknife, Ann. Statist. 7 (1979) 1–26.
[16] A. de Falguerolles, Log-bilinear biplots in action, in: J. Blasius, M. Greenacre (Eds.), Visualization of Categorical Data, Academic Press, San Diego, CA, 1998, pp. 527–539.
[17] W. Fithian, J. Josse, Multiple correspondence analysis and the multilogit bilinear model, J. Multivariate Anal. 157 (2017) 87–102.
[18] W. Fithian, R. Mazumder, Flexible low-rank statistical modeling with missing data and side information, Statist. Sci. 33 (2018) 238–260.
[19] J. Friedman, T. Hastie, H. Höfling, R.J. Tibshirani, Pathwise coordinate optimization, Ann. Appl. Stat. 1 (2007) 302–332.
[20] C. Giacobino, S. Sardy, J. Diaz Rodriguez, N. Hengartner, Quantile universal threshold, Electron. J. Statist. 11 (2017) 4701–4722.
[21] 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, Ann. Statist. 13 (1985) 10–69.
[22] P. Gopalan, F.J.R. Ruiz, R. Ranganath, D. Blei, Bayesian nonparametric Poisson factorization for recommendation systems, in: AISTATS, 2014, pp. 275–283.
[23] J. Gower, S. Lubbe, N. l. Roux, Understanding Biplots, Wiley, New York, 2011.
[24] M. Greenacre, Theory and Applications of Correspondence Analysis, Academic Press, Cambridge, MA, 1984.
[25] J. Josse, S. Wager, Bootstrap-based regularization for low-rank matrix estimation, J. Mach. Learn. Res. 17 (2016) 1–29.
[26] M. Kateri, Contingency Table Analysis, Springer, New York, 2014.
[27] O. Klopp, Noisy low-rank matrix completion with general sampling distribution, Bernoulli 20 (2014) 282–303.
[28] O. Klopp, Matrix completion by singular value thresholding: Sharp bounds, Electron. J. Statist. 9 (2015) 2348–2369.
[29] V. Koltchinskii, Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery, Springer, New York, 2011.
[30] V. Koltchinskii, A remark on low rank matrix recovery and noncommutative bernstein type inequalities, Inst. Math. Stat. Collect. 9 (2013) 213–226.
[31] J. Lafond, Low rank matrix completion with exponential family noise, in: J. Mach. Learn. Res. Workshop and Conference Proceedings, Vol. 40, 2015, pp. 1–18.
[32] M. Ledoux, The Concentration of Measure Phenomenon, American Mathematical Society, Providence, RI, 2001.

[33] J. de Leeuw, Principal component analysis of binary data by iterated singular value decomposition, Comput. Statist. Data Anal. 50 (2006) 21–39.
[34] J. Li, D. Tao, Simple exponential family PCA, IEEE Trans. Neural Netw. Learn. Syst. 24 (2013) 485–497.
[35] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, Wiley, New York, 2002.
[36] L.T. Liu, E. Dobriban, A. Singer, EPCA: High dimensional exponential family PCA, Ann. Appl. Statist. 12 (2018) 2121–2150.
[37] F. Luisier, T. Blu, M. Unser, Image denoising in mixed Poisson-Gaussian noise, IEEE Trans. Image Process. 20 (2011) 696–708.
[38] S. Mohamed, Z. Ghahramani, K.A. Heller, Bayesian exponential family PCA, in: D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (Eds.), Advances in Neural Information Processing Systems, Vol. 21, Curran Associates, Inc., 2009, pp. 1089–1096.
[39] J. Pannekoek, A. van Strien, Trim 3 Manual (Trends & Indices for Monitoring Data), Statistics Netherlands, 2001.
[40] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016.
[41] M. de Rooij, W.J. Heiser, Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data, Psychometrika 70 (2005) 99–122.
[42] J. Salmon, Z. Harmany, C. Deledalle, R. Willett, Poisson noise reduction with non-local PCA, J. Math. Imaging Vis. 48 (2014) 279–294.
[43] M. Sayoud, H. Salhi, B. Chalabi, A. Allali, M. Dakki, A. Qninba, M.E. Agbani, H. Azafzaf, C. Feltrup-Azafzaf, H. Dlensi, N. Hamouda, W.A.L. Ibrahim, H. Asran, A.A. Elnoor, H. Ibrahim, K. Etayeb, E. Bouras, W. Bashaimam, A. Berbash, C. Deschamps, J. Mondain-Monval, A. Brochet, S. Véran, P.D. de 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, Biol. Conserv. 206 (2017) 11–20.
[44] J.A. Tropp, User-friendly tail bounds for sums of random matrices, Found. Comput. Math. 12 (2012) 389–434.