Variational inference for probabilistic Poisson PCA

Julien Chiquet, Mahendra Mariadassou, Stephane Robin

VARIATIONAL INFERENCE FOR PROBABILISTIC POISSON PCA

By Julien Chiquet, Mahendra Mariadassou and Stéphane Robin

AgroParisTech, INRA, Université Paris-Saclay

Many application domains such as ecology or genomics have to deal with multivariate non Gaussian observations. A typical example is the joint observation of the respective abundances of a set of species in a series of sites, aiming to understand the co-variations between these species. The Gaussian setting provides a canonical way to model such dependencies, but does not apply in general. We consider here the multivariate exponential family framework for which we introduce a generic model with multivariate Gaussian latent variables. We show that approximate maximum likelihood inference can be achieved via a variational algorithm for which gradient descent easily applies. We show that this setting enables us to account for covariates and offsets. We then focus on the case of the Poisson-lognormal model in the context of community ecology.

1. Introduction. Principal component analysis (PCA) is among the oldest and most popular tool for multivariate analysis. It basically aims at reducing the dimension of a large data set made of continuous variables ([Anderson](#), [2003](#); [Mardia et al.](#), [1979](#)) in order to ease its interpretation and visualization. The methodology basically exploits the dependency structure between the variables to exhibit the few synthetic variables that best summarize the information content of the whole data set: the principal components. From a purely algebraic point-of-view, PCA can be seen as a matrix-factorization problem were the data matrix is decomposed as the product of a loading matrix with a score matrix ([Eckart and Young, 1936a](#)).

For statistical purposes, PCA can also be cast in a probabilistic framework. Probabilistic PCA (pPCA) is a model-based version of PCA originally defined in a Gaussian setting, in which the principal components are treated as hidden variables ([Tipping and Bishop, 1999](#); [Minka, 2000](#)). It is closely related to factor analysis and, as it involves hidden variable, maximum-likelihood estimates (MLE) can be obtained via an EM algorithm ([Dempster et al., 1977](#)). The Gaussian setting is obviously convenient as the dependency

---

structure is entirely described by the covariance matrix but pPCA has since been extended to more general settings.

Indeed, in many applications (Royle and Wikle, 2005; Srivastava and Chen, 2010) Gaussian pPCA needs to be adapted to handle specific measurement types, such as binary or count data. For count data, the multivariate Poisson distribution seems a natural counterpart of the multivariate normal. Still, there is no canonical form (Johnson et al., 1997) and several distributions have been proposed in the literature including Gamma-Poisson (Nelson, 1985) and lognormal-Poisson (Aitchison and Ho, 1989; Izsák, 2008) as an alternative. The latter takes advantage of the properties of the Gaussian distribution to display a larger panel of dependency structure than the former, but maximum likelihood-based parameter inference raises some issues as the MLE of the covariance matrix is not always positive definite.

A series of works have considered the extension of PCA to a broader class of distributions, typically in the exponential family. The matrix factorization point-of-view has been adopted to satisfy a positivity constraint of the parameters (Lafond, 2015), to minimize the Poisson loss function (Cao and Xie, 2015) or more general losses (Lee and Seung, 2001) consistent with exponential family noise. Sparse extensions have also been proposed (Witten et al., 2009; Liu et al., 2016). In a model-based context, Collins et al. (2001) suggest to minimize a Bregman divergence to get estimates of the principal components: the divergence is chosen according to the distribution at hand and a generic alternative minimization scheme is proposed. Salmon et al. (2014) consider a similar framework and use matrix factorization for the minimization of Bregman divergence. In both cases, the principal components are considered as fixed parameters. Mohamed et al. (2009) cast the same model in a Bayesian context and use Monte-Carlo sampling for the inference. Acharya et al. (2015) consider Bayesian inference of the Gamma-Poisson distribution.

As recalled above, in pPCA, principal components are treated as hidden variables. One of the main issue of non-Gaussian pPCA arises from the fact that their conditional distribution given the observed data is often intractable, which hampers the use of an Expectation-Minimization (EM) strategy. Variational approximations (Jaakkola and Jordan, 2000; Wainwright and Jordan, 2008) have become a standard tool to approximate such a conditional distribution. Karlis (2005) use such an approximation for the inference of the one-dimensional lognormal-Poisson model and derive a variational EM (VEM) algorithm. Hall et al. (2011) provide a theoretical analysis of this approximation for the same model and prove the consistency of the estimators. Li and Tao (2010) also use such an approximation to

extend pPCA to the simple exponential family, considering both loadings and scores as random hidden variables. Landgraf and Lee (2015) reframes exponential family PCA as an optimization problem with rank constraints and develops both a convex relaxation and a Maximization-Minimization (MM) algorithm to solve it for binomial and Poisson families. Finally Zhou (2016); Zhou et al. (2012) consider factor analysis in the more complex setting of negative-binomial families. The main difference between previous approaches and ours is that we only consider loading as random hidden variables, whereas we consider the scores as parameter. This has deep consequences on the general properties of the inference algorithm.

Our contribution. In this paper, pPCA is extended to the simple exponential family. We consider the principal components as Gaussian hidden variables to allow a large panel of dependency structures. We rely on a frequentist setting in order to avoid heavy-computing Monte-Carlo sampling often required in Bayesian inference. We use a variational approximation of the conditional distribution of the components given the observed data to derive a variational lower bound of the likelihood. This bound is biconcave, *i.e.* concave in the model parameters and in the variational parameters but not jointly concave in general. We use a quasi-Newton method with box-constraints to optimize the objective function and thus estimate the parameters, in contrast to the EM algorithm traditionally used in this setting. Additionally, the model-based framework allows us to introduce covariates and offsets in the model and to handle missing data at no additional cost. In practical analyses, this enables us to distinguish between correlations that are caused by known covariates from those corresponding to unknown structure and requiring further investigations.

Organization of the paper is as follows: in Section 2 we introduce pPCA for the exponential family and the variational framework considered. Section 3 generalizes the model in the manner of a generalized linear model, in order to handle covariates and offsets. Then, Section 4 is dedicated to the inference and optimization strategy. Section 5 details the special Poisson case and Section 6 devises the visualization, an important issue for non-Gaussian PCA methods. Finally, Section 7 considers applications to two examples from metagenomics: the impact of a pathogenic fungi on microbial communities from tree leaves, and the impact of weaning on piglets gut microbiota.

2. A variational framework for probabilistic PCA in the exponential family. We start this section by stating the probabilistic framework associated to Gaussian probabilistic PCA. Then we show how it can be naturally extended to other exponential families. We finally derive vari-

ational lower bounds for the likelihoods of pPCA which are the building blocks of our inference strategy.

2.1. *Gaussian probabilistic PCA (pPCA).* The probabilistic version of principal component analysis – or pPCA – (Minka, 2000; Mohamed et al., 2009; Tipping and Bishop, 1999) relates a sample of p-dimensional observation vectors $\boldsymbol{Y}_i$ to a sample of q-dimensional vectors of latent variables $\boldsymbol{W}_i$ in the following way:

$$\text{(1)} \qquad \boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{W}_i + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{0}_p, \sigma^2 \boldsymbol{I}_p).$$

The parameter $\boldsymbol{\mu}$ allows the mode to have *main effects*. The $p \times q$ matrix $\boldsymbol{B}$ captures the dependence between latent and observed variables. Furthermore, the latent vectors are conventionally assumed to have independent Gaussian component with unit variance $\sigma^2 = 1$, that is to say, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{0}_q, \boldsymbol{I}_q)$. This ensures that there is no structure in the latent space. Model (1) can thus be restated as $\boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{B}^\intercal + \boldsymbol{I}_p)$.

In the following, we consider an alternative formulation stated in a hierarchical framework. Despite its seemingly more complex statement it lends itself to generalizations. Formally,

$$\text{(2)} \qquad \begin{array}{lll} \text{latent space} & \boldsymbol{W}_i \quad \text{i.i.d.} & \boldsymbol{W}_i \sim \mathcal{N}(\boldsymbol{0}_q, \boldsymbol{I}_q) \\ \text{parameter space} & \boldsymbol{Z}_i | \boldsymbol{W}_i = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{W}_i & \boldsymbol{Z}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{B}^\intercal) \\ \text{observation space} & Y_{ij} | Z_{ij} \quad \text{indep.} & Y_{ij} | Z_{ij} \sim \mathcal{N}(Z_{ij}, \sigma^2) \end{array}$$

In Equation (2), $\boldsymbol{Z}_i$ is a linear transform of $\boldsymbol{W}_i$ and the last layer $\boldsymbol{Y}_i | \boldsymbol{Z}_i$ simply corresponds to *observation noise*. The diagonal nature of the covariance matrix of $\boldsymbol{\varepsilon}_i$ in (2) means that, conditionally on $\boldsymbol{Z}_i$, all components of $\boldsymbol{Y}_i$ are in fact independent. This is why we may consider univariate variables $Y_{ij} | Z_{ij}$ in Formulation (2).

Informally, the *latent* variables $\boldsymbol{W}_i$ (in $\mathbb{R}^q$) are mapped to a linear subspace of the *parameter* space $\mathbb{R}^p$ via the $\boldsymbol{Z}_i$ which are then pushed into the *observation* space using Gaussian emission laws. The main idea of this paper is to replace Gaussian emission laws with more general probability distributions, namely univariate natural exponential families. The focus of the inference is on the main effects $\boldsymbol{\mu}$ and the matrix $\boldsymbol{B}$ that captures the dependence.

Hereafter and unless stated otherwise,

- index i refers to *observations* and ranges in $\{1, \ldots, n\}$,
- index j refers to *variables* and ranges in $\{1, \ldots, p\}$,
- index k refers to *factors* and ranges in $\{1, \ldots, q\}$.

2.2. *Natural Exponential family (NEF).*   The work in this study is based on essential properties of univariate *natural exponential families* (NEF) where the parameter is in canonical form. They include normal distribution with known variance, Poisson distribution, gamma distribution with known shape parameter (and therefore exponential distribution as a particular example) and binomial distribution with known number of trials. The probability density (or mass function) of a NEF can be written

$$(3) \qquad f(x|\lambda) = \exp\left(x\lambda - b(\lambda) - a(x)\right)$$

where λ is the canonical parameter and b and a are known functions. The function b is well known to be convex (and analytic) over its domain and the mean and variance are easily deduced from b as

$$\mathbb{E}_\lambda[X] = b'(\lambda) \quad \text{and} \quad \mathbb{V}_\lambda[X] = b''(\lambda).$$

The canonical link function g is defined such that $g(b'(\lambda)) = \lambda$. The maximum likelihood estimate $\hat{\lambda}$ of λ from a single observation x is given by $\hat{\lambda} = \hat{\lambda}(x) = g(x)$ and satisfies

$$\mathbb{E}_{\hat{\lambda}(x)}[X] = b'(\hat{\lambda}(x)) = x.$$

2.3. *Probabilistic PCA for the exponential family.*   We now extend pPCA from the Gaussian setting to more general NEF. The connection between the two versions is exactly the same as the connection between linear models and generalized linear models (GLM). Intuitively, we assume that *i)* there exists a (low) q-dimensional (linear) subspace in the *natural canonical parameter space* where some latent variable $\boldsymbol{Z}_i$ lie; and *ii)* observations $\boldsymbol{Y}_i$ are generated in the *data space* according to some NEF distribution with parameter $\boldsymbol{Z}$. The latter is linked to $\mathbb{E}[\boldsymbol{Y}_i|\boldsymbol{Z}_i]$ through the canonical link function g. In the Gaussian case, the link function is the identity and the parameter space can be identified with the data space but this is not the case in general for other families.

Formally, we extend Model (2) to

$$(4) \qquad \begin{aligned} \boldsymbol{W}_i \quad &\text{i.i.d.} \quad \boldsymbol{W}_i \sim \mathcal{N}(\boldsymbol{0}_q, \boldsymbol{I}_q) \\ \boldsymbol{Z}_i|\boldsymbol{W}_i = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{W}_i \quad &\boldsymbol{Z}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{B}^\mathsf{T}) \\ Y_{ij}|Z_{ij} \quad &\text{indep.} \quad Y_{ij}|Z_{ij} \sim \exp\left(Y_{ij}Z_{ij} - b(Z_{ij}) - a(Y_{ij})\right) \end{aligned}$$

Note in particular that $g(\mathbb{E}[Y_{ij}|Z_{ij}]) = b'(Z_{ij})$ and that an unconstrained estimate $\tilde{Z}_{ij}$ of Z_{ij} is $\tilde{Z}_{ij} = g(Y_{ij})$. The vector $\boldsymbol{\mu}$ corresponds to main effects, $\boldsymbol{B}$ to *rescaled* loadings in the parameter spaces and $\boldsymbol{W}_i$ to scores of the i-th

observation in the low-dimensional latent subspace of the parameter space. The model specified in (4) is the same as the one specified in (2) but for the last data emission layer.

REMARK 1. $\boldsymbol{B}$ *suffers from two identifiability limitations. First, it is only identifiable through* $\boldsymbol{BB}^\intercal$ *and therefore at best up to rotations in* $\mathbb{R}^q$. *Second, if* $\boldsymbol{B}$ *is of rank* $q' < q$, *the model is degenerate and could be written using a* q'-*dimensional latent space. Note that the first limitation is shared with standard PCA where each principal component (PC) can be arbitrarily flipped without changing the least square criteria. More generally PCA finds a good approximation subspace but without additional constraints, infinitely many bases can be used to parametrize this subspace. The second limitation deprives us from the nestedness properties induced by* Eckart and Young's *theorem in standard PCA: there is no guarantee that the best* q'-*dimensional model can be built easily from the best* q-*dimensional model.*

2.4. *Likelihood.* Note $\boldsymbol{Y}$ (resp. $\boldsymbol{W}$) the $n \times p$ (resp. $n \times q$) matrix obtained by stacking the row-vectors $\boldsymbol{Y}_i^\intercal$ (resp. $\boldsymbol{W}_i^\intercal$). Conversely, for any matrix $\boldsymbol{A}$, $\boldsymbol{A}_i$ refers to the i-th row of $\boldsymbol{A}$ considered as a *column* vector. In matrix expression, $\boldsymbol{Z} = \mathbf{1}_n\boldsymbol{\mu}^\intercal + \boldsymbol{W}\boldsymbol{B}^\intercal$. The observation matrix $\boldsymbol{Y}$ only depends on $\boldsymbol{Z}$ through $\boldsymbol{\mu}$, $\boldsymbol{B}$ and $\boldsymbol{W}$ and the complete log-likelihood is therefore

$$\log p(\boldsymbol{Y}, \boldsymbol{W}; \boldsymbol{\mu}, \boldsymbol{B}) = \sum_{i=1}^n \log p(\boldsymbol{Y}_i | \boldsymbol{W}_i; \boldsymbol{\mu}, \boldsymbol{B}) + \log p(\boldsymbol{W}_i)$$

$$= \sum_{i=1}^n \left[\sum_{j=1}^p Y_{ij}(\mu_j + \boldsymbol{B}_j^\intercal\boldsymbol{W}_i) - b(\mu_j + \boldsymbol{B}_j^\intercal\boldsymbol{W}_i) - a(Y_{ij}) - \sum_{k=1}^q \frac{W_{ik}^2 + \log(2\pi)}{2} \right]$$

which can be stated in the following compact matrix form:

$$(5) \quad \log p(\boldsymbol{Y}, \boldsymbol{W}; \boldsymbol{\mu}, \boldsymbol{B}) = \mathbf{1}_n^\intercal \left[\boldsymbol{Y} \odot (\mathbf{1}_n\boldsymbol{\mu}^\intercal + \boldsymbol{W}\boldsymbol{B}^\intercal) - b(\mathbf{1}_n\boldsymbol{\mu}^\intercal + \boldsymbol{W}\boldsymbol{B}^\intercal) \right] \mathbf{1}_p$$

$$- \frac{\|\boldsymbol{W}\|_F^2}{2} - \frac{nq}{2}\log(2\pi) - K(\boldsymbol{Y}),$$

where the function a and b are applied component-wise to vectors and matrices, $\odot$ is the Hadamard product and $K(\boldsymbol{Y}) = \mathbf{1}_n^\intercal a(\boldsymbol{Y})\mathbf{1}_p$ is a constant depending only on $\boldsymbol{Y}$ and not scaling with q.

We do not know how to integrate out $\boldsymbol{W}$ and therefore cannot derive an analytic expression of $\log p(\boldsymbol{Y}; \boldsymbol{\mu}, \boldsymbol{B})$. Numerical approximation using Hermite-Gauss quadrature or MCMC techniques are possible but likely to become computationally prohibitive as the dimension of the integration

space increases. A standard EM algorithm relying on $\mathbb{E}_{W|Y}[\log p(\boldsymbol{Y}, \boldsymbol{W}; \boldsymbol{\mu}, \boldsymbol{B})]$ is also not possible as it would require at least first and second order of $p(\boldsymbol{W_i}|\boldsymbol{Y_i})$ which are unknown in general. We resort instead to a variational strategy and integrate out $\boldsymbol{W}$ under a tractable approximation of $p(\boldsymbol{W}|\boldsymbol{Y})$.

2.5. *Variational bound of the likelihood.*   Consider any product distribution $\tilde{p} = \otimes_{i=1}^{n}\tilde{p}_i$ on the $\boldsymbol{Z}_i$. The variational approximation relies on maximizing the following lower bound over a tractable set for $\tilde{p}$

$$\log p(\boldsymbol{Y}; \boldsymbol{\mu}, \boldsymbol{B}) \geq J_q(\tilde{p}, \boldsymbol{\mu}, \boldsymbol{B})$$

where

$$J_q(\tilde{p}, \boldsymbol{\mu}, \boldsymbol{B}) := \log p(\boldsymbol{Y}; \boldsymbol{\mu}, \boldsymbol{B}) - KL(\tilde{p}(\boldsymbol{W})\|p(\boldsymbol{W}|\boldsymbol{Y}; \boldsymbol{\mu}, \boldsymbol{B}))$$

(6)
$$= \mathbb{E}_{\tilde{p}}[\log p(\boldsymbol{Y}, \boldsymbol{W}; \boldsymbol{\mu}, \boldsymbol{B}) - \log \tilde{p}(\boldsymbol{W})]$$

$$= \sum_{i=1}^{n} \mathbb{E}_{\tilde{p}_i}[\log p(\boldsymbol{W}_i) + \log p(\boldsymbol{Y}_i|\boldsymbol{W}_i; \boldsymbol{\mu}, \boldsymbol{B}) - \log \tilde{p}_i(\boldsymbol{W}_i)],$$

with term-by-term inequality:

$$\log p(\boldsymbol{Y}_i; \boldsymbol{\mu}, \boldsymbol{B}) \geq J_q(\tilde{p}_i, \boldsymbol{\mu}, \boldsymbol{B})$$
$$:= \mathbb{E}_{\tilde{p}_i}[\log p(\boldsymbol{W}_i) + \log p(\boldsymbol{Y}_i|\boldsymbol{W}_i; \boldsymbol{\mu}, \boldsymbol{B}) - \log \tilde{p}_i(\boldsymbol{W}_i)].$$

In our variational approximation, we choose here the set $\mathcal{Q}$ of product distribution of q-dimensional multivariate Gaussian with diagonal covariance matrices:

(7)
$$\mathcal{Q} = \left\{ \tilde{p} := \tilde{p}_{\boldsymbol{M}, \boldsymbol{S}}; \ \tilde{p}(\boldsymbol{w}) = \prod_{i=1}^{n} \tilde{p}_i(\boldsymbol{w}_i) \right\},$$
$$\text{where } \tilde{p}_i = \mathcal{N}(\boldsymbol{m}_i, \text{diag}(\boldsymbol{s}_i \odot \boldsymbol{s}_i)), \quad (\boldsymbol{m}_i, \boldsymbol{s}_i) \in \mathbb{R}^q \times \mathbb{R}_+^q.$$

The $n \times q$ matrices $\boldsymbol{M}$ and $\boldsymbol{S}$ are obtained by respectively stacking $\boldsymbol{m}_i^\mathsf{T}$ and $\boldsymbol{s}_i^\mathsf{T}$. Note that by construction $p(\boldsymbol{W}|\boldsymbol{Y})$ is a product distribution and that the approximation only stems from the functional form of each $\tilde{p}_i$, *i.e.* multivariate normal with diagonal variance-covariance matrix. For such $\tilde{p} = \tilde{p}_{\boldsymbol{M}, \boldsymbol{S}}$, results on first and second order moments of multivariate Gaussian show that

$$J_q(\boldsymbol{\mu}, \boldsymbol{B}, \boldsymbol{m}_i, \boldsymbol{s}_i) := J_q(\tilde{p}_i, \boldsymbol{\mu}, \boldsymbol{B})$$
$$= \boldsymbol{Y}_i^\mathsf{T}(\boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{m}_i) - \frac{1}{2}[\|\boldsymbol{m_i}\|_2^2 + \|\boldsymbol{s}_i\|_2^2] + \frac{1}{2}(\boldsymbol{2}_q^\mathsf{T}\log(\boldsymbol{s}_i) + q)$$
$$- \boldsymbol{1}_p^\mathsf{T}\mathbb{E}_{\tilde{p}_i}[b(\boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{W}_i)] - K(\boldsymbol{Y}).$$

Therefore,

$$
(8) \quad J_q(\boldsymbol{\mu}, \boldsymbol{B}, \boldsymbol{M}, \boldsymbol{S}) := J_q(\tilde{p}_{\boldsymbol{M},\boldsymbol{S}}, \boldsymbol{\mu}, \boldsymbol{B}) = \sum_{i=1}^{n} J_q(\boldsymbol{\mu}, \boldsymbol{B}, \boldsymbol{m}_i, \boldsymbol{s}_i)
$$

$$
= \mathbf{1}_n^\intercal \left[\boldsymbol{Y} \odot (\mathbf{1}_n \boldsymbol{\mu}^\intercal + \boldsymbol{M}\boldsymbol{B}^\intercal) - \mathbb{E}_{\tilde{p}}[b\left(\mathbf{1}_n^\intercal \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{B}^\intercal\right)]\right] \mathbf{1}_p
$$

$$
- \frac{1}{2}\mathbf{1}_n^\intercal \left[\boldsymbol{M} \odot \boldsymbol{M} + \boldsymbol{S} \odot \boldsymbol{S} - 2\log(\boldsymbol{S}) - \mathbf{1}_{n,q}\right] \mathbf{1}_q - K(\boldsymbol{Y}).
$$

Depending on the natural exponential family and thus the exact value of b in (8), we may have a fully explicit variational bound for the complete likelihood which paves the way for efficient optimization. In particular, this is the case with the Poisson distribution that we investigate in further details in Section 5.

Before moving on to actual inference, we show that the framework that we introduced above can be extended to account for covariates and offsets.

3. Accounting for covariates and offsets. Multivariate analyses typically aim at deciphering dependencies between variables. Variations induced by the effect of covariates may be confounded with these dependencies. Therefore, it is desirable to account for such effects to focus on the residual dependencies. The rational of our approach is to postulate the existence of a model similar to linear regression in the *parameter* space. We consider the general framework of linear regression with multivariate outputs, which encompasses multivariate analysis of variance.

3.1. *Model and likelihood.* Suppose that each observation i is associated to a known d-dimensional covariate vector $\boldsymbol{X}_i$. We assume that the covariates act linearly in the *parameter* space through a $p \times d$ regression matrix $\boldsymbol{\Theta}$, *i.e.* $\boldsymbol{X}_i$ is linearly related to $\boldsymbol{Z}_i$. It can be also useful to add an offset to model different sampling efforts and/or exposures. There is usually one known offset parameter O_{ij} per observation Y_{ij} and this offset can be readily incorporated in our framework. Thus, a natural generalization of (4) accounting for covariates and offsets is

(9)
$$
\boldsymbol{W}_i \quad \text{i.i.d.} \quad \boldsymbol{W}_i \sim \mathcal{N}(\boldsymbol{0}_q, \boldsymbol{I}_q)
$$
$$
\boldsymbol{Z}_i | \boldsymbol{W}_i = \boldsymbol{O}_i + \boldsymbol{\Theta}\boldsymbol{X}_i + \boldsymbol{B}\boldsymbol{W}_i \quad \boldsymbol{Z}_i \sim \mathcal{N}(\boldsymbol{O}_i + \boldsymbol{\Theta}\boldsymbol{X}_i, \boldsymbol{B}\boldsymbol{B}^\intercal)
$$
$$
Y_{ij} | Z_{ij} \quad \text{indep.} \quad Y_{ij} | Z_{ij} \sim \exp\left(Y_{ij}Z_{ij} - b(Z_{ij}) - a(Y_{ij})\right)
$$

where a column of ones can be added to the data matrix $\mathbf{X}$ to get an intercept in the model. The log-likelihood can be computed from (9) like before to get

$$(10) \quad \log p(\boldsymbol{Y}, \boldsymbol{W}; \boldsymbol{B}, \boldsymbol{\Theta}, \boldsymbol{O})$$
$$= \mathbf{1}_n^\intercal \left[\boldsymbol{Y} \odot (\boldsymbol{O} + \boldsymbol{X}\boldsymbol{\Theta}^\intercal + \boldsymbol{W}\boldsymbol{B}^\intercal) - b(\boldsymbol{O} + \boldsymbol{X}\boldsymbol{\Theta}^\intercal + \boldsymbol{W}\boldsymbol{B}^\intercal) \right] \mathbf{1}_p$$
$$- \frac{\|\boldsymbol{W}\|_F^2}{2} - \frac{nq}{2} \log(2\pi) - K(\boldsymbol{Y}),$$

where the focus of inference is on $\boldsymbol{B}$ and $\boldsymbol{\Theta}$ while $\boldsymbol{O}$ is known.

3.2. *Variational bound of the likelihood.* We can use the variational class $\mathcal{Q}$ previously defined in (7) to lower bound the likelihood from Eq. (10). We first introduce the instrumental matrix $\boldsymbol{A}$, which appears in many equations.

$$(11) \quad \begin{aligned} \boldsymbol{A} &= \mathbb{E}_{\tilde{p}}[b(\boldsymbol{O} + \boldsymbol{X}\boldsymbol{\Theta}^\intercal + \boldsymbol{W}\boldsymbol{B}^\intercal)] \\ &= \mathbb{E}[b(\boldsymbol{O} + \boldsymbol{X}\boldsymbol{\Theta}^\intercal + (\boldsymbol{M} + \boldsymbol{S}\odot\boldsymbol{U})\boldsymbol{B}^\intercal)] \quad = \mathbb{E}[b(\boldsymbol{Z})], \end{aligned}$$

where $\boldsymbol{Z} = (\boldsymbol{O} + \boldsymbol{X}\boldsymbol{\Theta}^\intercal + (\boldsymbol{M} + \boldsymbol{S}\odot\boldsymbol{U})\boldsymbol{B}^\intercal)$ and $\boldsymbol{U}$ is a $n \times q$ matrix with unit variance independent Gaussian components.

Since $\boldsymbol{O}$ is known, we drop it from the arguments of J_q and obtain the following lower bound, which extends the bound from Eq. (8):

$$(12) \quad J_q(\boldsymbol{\Theta}, \boldsymbol{B}, \boldsymbol{M}, \boldsymbol{S}) = \mathbf{1}_n^\intercal \left(\boldsymbol{Y} \odot (\boldsymbol{O} + \boldsymbol{X}\boldsymbol{\Theta}^\intercal + \boldsymbol{M}\boldsymbol{B}^\intercal) - \boldsymbol{A} \right) \mathbf{1}_p$$
$$- \frac{1}{2}\mathbf{1}_n^\intercal \left[\boldsymbol{M}\odot\boldsymbol{M} + \boldsymbol{S}\odot\boldsymbol{S} - 2\log(\boldsymbol{S}) - \mathbf{1}_{n,q} \right] \mathbf{1}_q - K(\boldsymbol{Y}).$$

4. Inference. As usual in the variational framework, we aim to maximize the lower bound J_q which we call the objective function in an optimization perspective. The optimization shall be performed on $\boldsymbol{\Theta}, \boldsymbol{B}, \boldsymbol{M}, \boldsymbol{S}$. We only give results in the most general case (12) with covariates and offsets. All other case are deduced by setting $\boldsymbol{O} = \boldsymbol{0}_{n\times p}$ and/or $\boldsymbol{X} = \mathbf{1}_n$ hereafter.

4.1. *Inference strategy.* We first highlight the biconcavity of the objective function J_q. The major part of the proof is postponed to Appendix A.

PROPOSITION 1. *The variational objective function $J_q(\boldsymbol{\Theta}, \boldsymbol{B}, \boldsymbol{M}, \boldsymbol{S})$ is concave in $(\boldsymbol{\Theta}, \boldsymbol{B})$ for $(\boldsymbol{M}, \boldsymbol{S})$ fixed and vice-versa.*

PROOF. Fix $(\boldsymbol{M}, \boldsymbol{S})$ in (12). The non explicit part of J_q, that is to say $-\mathbf{1}_n^\intercal \boldsymbol{A}\mathbf{1}_p$, is concave in $(\boldsymbol{\Theta}, \boldsymbol{B})$ thanks to Lemma 2 (see Appendix A). By inspection, the explicit part of J_q involves linear, quadratic and concave functions of $(\boldsymbol{\Theta}, \boldsymbol{B})$ and is also concave. The objective J_q is therefore concave in $(\boldsymbol{\Theta}, \boldsymbol{B})$. The same is true for $(\boldsymbol{M}, \boldsymbol{S})$ when fixing $(\boldsymbol{\Theta}, \boldsymbol{B})$. □

A standard approach for maximizing biconcave functions is block coordinate descents, of which the Expectation-Maximization (EM) algorithm is a popular representative in the latent variable setting. It is especially powerful when we have access to closed formula for both the optimal $(\boldsymbol{M}, \boldsymbol{S})$ given $(\boldsymbol{\Theta}, \boldsymbol{B})$ (E-step) and the optimal $(\boldsymbol{\Theta}, \boldsymbol{B})$ given $(\boldsymbol{M}, \boldsymbol{S})$ (M-step). However, the non-linear nature of $\mathbb{E}_{\widetilde{p}}[b\left(\boldsymbol{O} + \boldsymbol{X}\boldsymbol{\Theta}^{\mathsf{T}} + \boldsymbol{W}\boldsymbol{B}^{\mathsf{T}}\right)]$ combined with careful inspection of the objective function J_q shows that setting the derivatives of J_q to 0, even after fixing the variational or model parameters, does not lead to closed formula neither for $(\boldsymbol{M}, \boldsymbol{S})$ nor $(\boldsymbol{B}, \boldsymbol{\Theta})$. Nevertheless, since we may derive convenient expressions for the gradient $\boldsymbol{\nabla} J_q$ (see next Section 4.2), we propose to rely on Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton updates with box constraints and limited memory (a.k.a L-BFGS-B) to maximize J_q (see, e.g. Press et al., 1989). In the general case (12), the total number of parameters to optimize $J_q(\boldsymbol{\Theta}, \boldsymbol{B}, \boldsymbol{M}, \boldsymbol{S})$ is $p(d+q)+2nq$. The box constraints concern the variational parameters $\boldsymbol{S}$, standing for standard deviations in (7) and thus only defined on $\mathbb{R}_+^q$. The starting point is chosen according to the exact value of b.

4.2. *Blockwise gradients of J_q.* The blockwise gradient of $J_q(\boldsymbol{\Theta}, \boldsymbol{B}, \boldsymbol{M}, \boldsymbol{S})$ can be expressed compactly in matrix notations. We skip the tedious but straightforward derivation and present only the resulting partial gradients. We introduce $\boldsymbol{A}' = \mathbb{E}[b'\left(\boldsymbol{Z}\right)]$, the natural counterpart to matrix $\boldsymbol{A}$ given in (11). Intuitively, A'_{ij} is the conditional expectation of Y_{ij} under $\widetilde{p}_i$. On top of that, we need two other matrices denoted $\boldsymbol{A}'_1$ and $\boldsymbol{A}'_2$, defined as follows:

$$\boldsymbol{A}'_1 = \mathbb{E}[b'\left(\boldsymbol{Z}\right)^{\mathsf{T}}\left(\boldsymbol{S} \odot \boldsymbol{U}\right)], \qquad \boldsymbol{A}'_2 = \mathbb{E}[(b'\left(\boldsymbol{Z}\right)\boldsymbol{B}) \odot \boldsymbol{U}].$$

With those matrices the derivatives of J_q can be expressed compactly as

(13)
$$
\begin{aligned}
\frac{\partial J_q}{\partial \boldsymbol{\Theta}} &= (\boldsymbol{Y} - \boldsymbol{A}')^{\mathsf{T}}\boldsymbol{X}, & \frac{\partial J_q}{\partial \boldsymbol{B}} &= (\boldsymbol{Y} - \boldsymbol{A}')^{\mathsf{T}}\boldsymbol{M} - \boldsymbol{A}'_1, \\
\frac{\partial J_q}{\partial \boldsymbol{M}} &= (\boldsymbol{Y} - \boldsymbol{A}')\boldsymbol{B} - \boldsymbol{M}, & \frac{\partial J_q}{\partial \boldsymbol{S}} &= \left[\boldsymbol{S}^{\oslash} - 2\boldsymbol{A}'_2 - 2\boldsymbol{S}\right],
\end{aligned}
$$

where the $n \times q$ matrix $\boldsymbol{S}^{\oslash}$ is the elementwise inverse of $\boldsymbol{S}$, i.e. $S_{ij}^{\oslash} = S_{ij}^{-1}$ for all $i = 1, \ldots, n$, $q = 1, \ldots, Q$.

4.3. *About missing data.* In the presence of missing data, note $\Omega \subset \{1, \ldots, n\} \times \{1, \ldots, p\}$ the set of observed data and $\boldsymbol{\Omega}$ the matrix where $\Omega_{ij} = 1$ if $(i,j) \in \Omega$ and 0 otherwise. The likelihood can be adapted from

Eq. (10) as follows:

$$
\begin{aligned}
\log p(\boldsymbol{Y}, \boldsymbol{W}; \boldsymbol{B}, \boldsymbol{\Theta}, \boldsymbol{O}) \\
= \boldsymbol{1}_n^\intercal \Big(\big(\boldsymbol{Y} \odot (\boldsymbol{O} + \boldsymbol{X}\boldsymbol{\Theta}^\intercal + \boldsymbol{W}\boldsymbol{B}^\intercal) - b(\boldsymbol{O} + \boldsymbol{X}\boldsymbol{\Theta}^\intercal + \boldsymbol{W}\boldsymbol{B}^\intercal) \big) \odot \boldsymbol{\Omega} \Big) \boldsymbol{1}_p \\
- \frac{\|\boldsymbol{W}\|_F^2}{2} - \frac{nq}{2}\log(2\pi) - \operatorname{tr}(\boldsymbol{\Omega}^\intercal a(\boldsymbol{Y})).
\end{aligned}
$$

The corresponding variational bound J_q and its partial derivatives are then simple adaptations from Equations (12) and (13) where $\boldsymbol{Y}$ (resp. $\boldsymbol{A}$, $\boldsymbol{A'}$) is replaced with $\boldsymbol{Y} \odot \boldsymbol{\Omega}$ (resp. $\boldsymbol{A} \odot \boldsymbol{\Omega}$, $\boldsymbol{A'} \odot \boldsymbol{\Omega}$).

Note that it is strictly equivalent for the quasi-Newton method to use $(\boldsymbol{Y} - \boldsymbol{A'}) \odot \boldsymbol{\Omega}$ or to impute missing Y_{ij} with A'_{ij} before using Eq. (13). Since $\boldsymbol{A'}$ is computed as part of the gradient computation at each step, imputation of missing data is essentially a free by-product of the quasi-Newton method. Finally, note that $A'_{ij} = \mathbb{E}_{\widetilde{p}_i}[Y_{ij}]$ so that the imputation makes intuitive sense: we're imputing Y_{ij} with its conditional expectation under the current variational parameters.

4.4. *Model selection.* The dimension q of the latent space itself needs to be estimated. To this aim, we adopt a penalized-likelihood approach, replacing the log-likelihood by its lower bound J_q. We considered two classical criteria: BIC (Schwarz, 1978) and ICL (Biernacki et al., 2000). We remind that ICL uses the conditional entropy of the latent variables given the observations as an additional penalty with respect to BIC. Because the true conditional distribution $p(\boldsymbol{W}|\boldsymbol{Y})$ is intractable, we replace it with its variational approximation $\widetilde{p}(\boldsymbol{W})$ to evaluate this entropy. The difference between BIC and ICL measures the uncertainty of the representation of the observations in the latent space.

Because the number of parameters in our model is $p(q+d)$ and the entropy of each W_i under $\widetilde{p}_i$ is $q\log(2\pi e)/2 + \boldsymbol{1}_q^\intercal \log(\boldsymbol{s}_i)]$, we obtain the following pseudo-BIC and pseudo-ICL criteria:

$$
\begin{aligned}
(14) \qquad & BIC(q) = J_q - p(d+q)\log(n) \\
& ICL(q) = J_q - p(d+q)\log(n) - \frac{nq}{2}\log(2\pi e) - \boldsymbol{1}_n^\intercal \log(\boldsymbol{S})\boldsymbol{1}_q
\end{aligned}
$$

5. Poisson Family. Each term of the expectation matrix $\boldsymbol{A}$ in (11) can be reduced to computing expectations of the form $\mathbb{E}[b(a + cU)]$ for a convex analytic function b, a standard Gaussian $U \sim \mathcal{N}(0,1)$ and arbitrary scalars $(a, c) \in \mathbb{R}^2$. It can therefore be computed numerically efficiently using Gauss-Hermite quadrature (see, e.g., Press et al., 1989). However in

the special case of Poisson-distributed observations, $b(x) = e^x$ and most of the expectations can be computed analytically leading to explicit formulas for Equations (11), (12) and (13).

5.1. *Explicit form of $\boldsymbol{A}$, J_q, and $\boldsymbol{\nabla} J_q$.* In the Poisson-case, the variational expectation of the non-linear part involving b – the matrix of conditional expectations $\boldsymbol{A}$ – is equal to $\boldsymbol{A'}$ and can be expressed as

$$\boldsymbol{A} = \boldsymbol{A'} = \exp\left(\boldsymbol{O} + \boldsymbol{X}\boldsymbol{\Theta}^\intercal + \boldsymbol{M}\boldsymbol{B}^\intercal + \frac{1}{2}(\boldsymbol{S} \odot \boldsymbol{S})(\boldsymbol{B} \odot \boldsymbol{B})^\intercal\right).$$

The lower bound J_q and matrices $\boldsymbol{A'}_1$, $\boldsymbol{A'}_2$ appearing in (13) can be expressed simply from $\boldsymbol{A}$ as

$$\boldsymbol{A}'_1 = [\boldsymbol{A}^\intercal(\boldsymbol{S} \odot \boldsymbol{S})] \odot \boldsymbol{B}, \qquad \boldsymbol{A'}_2 = 2[\boldsymbol{A}(\boldsymbol{B} \odot \boldsymbol{B})] \odot \boldsymbol{S}.$$

5.2. *Implementation details.* We implemented our inference algorithm for the Poisson family in the R package **PLNmodels**, the last version of which is available on github https://github.com/jchiquet/PLNmodels. Maximization of variational bound J_q is done using the L-BFGS-B implementation of Byrd et al. (1995) available from the R optim function carefully tuned (R Development Core Team, 2008). All graphics are produced using the **ggplot2** package (Wickham, 2009).

The choice of a good starting value is crucial in iterative procedures as it helps the algorithm start in the attractor field of a good local maximum and can substantially speed-up convergence. Here we initialize $(\boldsymbol{\Theta}, \boldsymbol{B})$ by fitting a GLM-Poisson to $\boldsymbol{Y}$, then extracting the regression coefficients $\boldsymbol{\Theta}_{GLM}$ and the variance-covariance matrix $\boldsymbol{\Sigma}_{GLM}$ of the Pearson residuals. We set $\boldsymbol{\Theta}_0 = \boldsymbol{\Theta}_{GLM}$ and $\boldsymbol{B}_0 = (\boldsymbol{\Sigma}_{GLM}^{(q)})^{1/2}$ where $\boldsymbol{\Sigma}_{GLM}^{(q)}$ is the best rank q approximation of $\boldsymbol{\Sigma}_{GLM}$, as given by keeping the first q-dimensions of a SVD of $\boldsymbol{\Sigma}_{GLM}$. We set the other starting values as $\boldsymbol{M}_0 = \boldsymbol{S}_0 = \boldsymbol{0}_{n \times q}$.

6. Visualization.

6.1. *Specific issues in non-Gaussian PCA.* This section is dedicated to the visualization of the results of the proposed modeling. Although this problem has many similarities with visualization in standard PCA, two important differences exist that make adaptations of the usual procedures necessary.

(i) In the general case, the parameter space defined in (4) is different from the observation space, as opposed to the special case of Gaussian PCA.

(ii) The optimal subspaces of dimension $q = 1, 2, \ldots$ are not nested. As a consequence, in a model with $q > 1$ latent dimensions, there is no such thing as a 'first axis'. The genuine first axis corresponds to the optimal subspace of dimension 1, which is provided by the model with $q = 1$ latent dimension.

To address point (i), we choose to provide representations in the parameter space as it sets us in the Gaussian (and accompanying Euclidean) space setting practitioners are most familiar with. We thus focus on the positions $\boldsymbol{Z}_i$ and more specifically on their inferred version $\boldsymbol{O}_i + \widehat{\boldsymbol{\Theta}} \boldsymbol{X}_i + \widehat{\boldsymbol{B}} \widetilde{\boldsymbol{m}}_i$, gathered in the $n \times p$ matrix $\widetilde{\boldsymbol{Z}} := \boldsymbol{O} + \boldsymbol{X} \widehat{\boldsymbol{\Theta}}^{\mathsf{T}} + \widehat{\boldsymbol{B}} \widetilde{\boldsymbol{M}}$. Note that $\widetilde{\boldsymbol{Z}}$ is most useful to assess goodness of fit. For visualization of the latent structure remaining after correction for the offset and the covariates, we consider instead positions $\widetilde{\boldsymbol{P}} = \widehat{\boldsymbol{B}} \widetilde{\boldsymbol{M}}$ in the latent subspace of dimension q.

6.2. *Quality of the dimension reduction.* A first important criterion in PCA is the amount of information that is preserved by the q-dimensional reduction. To evaluate this criterion we define the matrix $\boldsymbol{\Lambda}^{(q)} = [\lambda_{ij}^{(q)}]$ where we use entry $\lambda_{ij}^{(q)} := \exp(\widetilde{Z}_{ij})$ as an estimate of the canonical parameter λ_{ij} of the distribution of Y_{ij} given in (3). Thus, we can define the log-likelihood ℓ_q of the observed data with these estimates as

$$\ell_q = \sum_{i=1}^{n} \sum_{j=1}^{p} [Y_{ij} g(\lambda_{ij}^{(q)}) - Y_{ij}] - K(\boldsymbol{Y}).$$

We can compare the log-likelihood of the saturated model $\ell_{\max}$ (replacing $\lambda_{ij}^{(q)}$ with $\lambda_{ij}^{\max} := Y_{ij}$) and the log-likelihood $\ell_{\min}$ of a GLM for Poisson regression with no latent structure, which plays the role of the null model here (using $\lambda_{ij}^{\min} := o_{ij} + \widehat{\boldsymbol{\Theta}} \boldsymbol{X}_i$, with $\widehat{\boldsymbol{\Theta}}$ estimated using a standard GLM). Then, similarly to the deviance criterion used in GLMs, we define the following measure of fit:

(15) $$R_q^2 = (\ell_q - \ell_{\min}) / (\ell_{\max} - \ell_{\min}) .$$

Note that, unlike standard PCA, there is no guarantee the R_q^2 is nondecreasing in q as ℓ_q is not the objective function of the variational inference algorithm of Section 4.

6.3. *Graphical outputs.* Because no orthogonality constraint is applied in the inference step, the (centered) columns of $\widetilde{\boldsymbol{P}}$ are not orthogonal in general, as opposed to standard PCA. Furthermore, because of point (ii),

the order of the columns of $\boldsymbol{B}$ have no specific meaning as only the subspace spanned by $\widetilde{\boldsymbol{P}}$ is identifiable.

To provide a representation of the observations in a series of 2D-plots, we simply apply PCA to $\widetilde{\boldsymbol{P}}$. As in regular PCA, we use the associated relative squared singular value $d_j^2 / \left(\sum_{k=1}^q d_k^2 \right)$ to measure the contribution of component j. Note that this measures the contribution of the component to the q-dimensional representation of the data and that it needs to be combined with the global measure of fit provided by R_q^2. The variational variances $\widetilde{\boldsymbol{s}}_i$ can be used to draw confidence ellipsoids. Following the same line, we may plot the correlations between the columns of $\widetilde{\boldsymbol{P}}$ and the components arising from its PCA, to help with the interpretation of these components.

## 7. Illustrations.

7.1. *Oak powdery mildew pathobiome.*

Description of the experiment. We considered the metagenomic dataset introduced in Jakuschkin et al. (2016), which consists in abundance measures of 66 bacterial species and 48 fungal species ($p = 114$) collected on the surface of $n = 116$ oak leaves. One aim of this experiment is to understand the association between the abundance of the fungal pathogenic species *E. alphitoides*, responsible for the oak powdery mildew, and the other species. The leaves were collected on three different trees and the species abundances were measured via metabarcoding.

Importance of the offset. The abundances Y_{ij} (where i denotes the leaf and j the species) were measured separately for fungi and bacteria resulting in different sampling efforts for the two types of species: the median total abundance were respectively 668 for bacteria and 2166 for fungi. To account for this we define an offset o_{ij} term as the log-total count of each species type (fungal or bacteria) for each leaf.

Model selection. The three trees from which the leafs where collected were respectively susceptible, intermediately resistant (hereafter "intermediate") and resistant to mildew. We first fitted a null lognormal-Poisson model M_0 as defined in (9) only with an offset term. Alternatively, we considered model M_1 involving two covariates: the tree from which each leaf was collected from, and the orientation (0=south-east, 1=north-west) of its branch.

Figure 1a displays the lower bound J, the BIC and the ICL for model M_0 (left) and M_1 (right) as a function of the number of axes q considered. We observe that the J_q is always increasing and that the BIC and ICL criteria behave similarly. According to the ICL criterion, we selected $\widehat{q}_0 = 24$ ($ICL =$
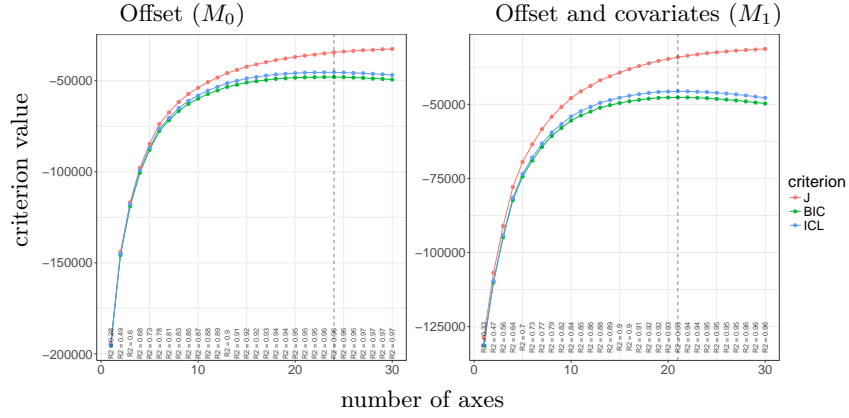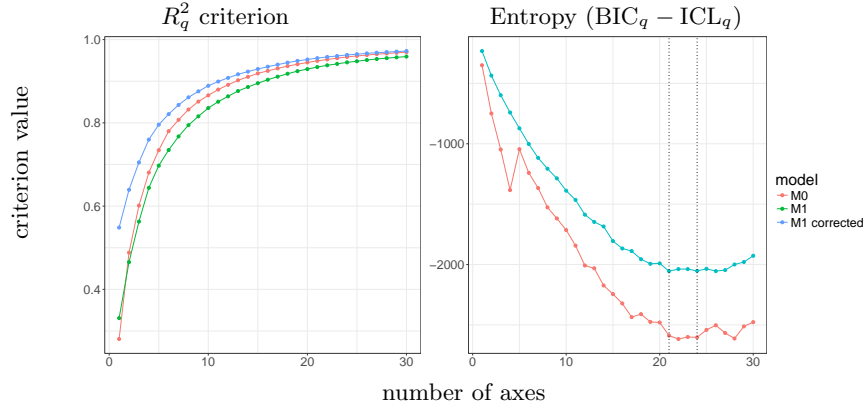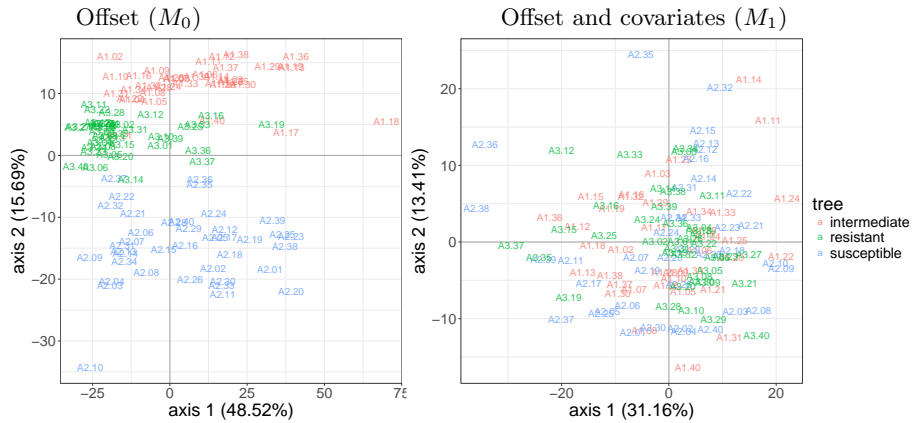
Fig 1: Dataset from Jakuschkin et al. (2016). (*a*) model selection criteria J_q, BIC_q and ICL_q for model M_0 (left) and M_1 (right); (*b*) R_q^2 criterion and entropy of $\widetilde{p}(\boldsymbol{W})$; (*c*) scatter plot of the leaves on the first two principal components (left: M_0, right: M_1).

-45377) latent dimensions for model M_0 and $\widehat{q} = 21$ ($ICL = -45430$) for model M_1. This suggest that the two models (with their respective optimal dimension) provide a very similar fit.

We looked at the approximate posterior entropy in panel left of Figure 1b: we observed that it is minimal near to the respective optimum in terms of model selection. This indicates that the selected dimensions are also optimal in terms of uncertainty on the latent variables.

Effect of the covariates. The choice between model M_0 and M_1 is mostly a matter of the type of dependency we analyze with each of them, as the former does not account for the covariates whereas the later does. This is illustrated in Figure 1c, when plotting the first principal plane. In model M_0 (left), the leafs collected on each tree are clearly separated. As expected, taking the tree as a covariate (right) removes the tree effect from the principal plane.

The effect of covariates on the abundance of *E. alphitoides* were also consistent: the estimates parameters θ_{ij} associated with the intermediate and resistant trees were -4.53 and -8.83, respectively, taking the susceptible tree as a reference.

We compared the respective estimates of $\boldsymbol{\Sigma}$ under M_0 (denoted $\widehat{\boldsymbol{\Sigma}}_0$) and under M_1 ($\widehat{\boldsymbol{\Sigma}}_1$) focusing on the correlation between *E. alphitoides* and the other species. $\widehat{\boldsymbol{\Sigma}}_0$ contains correlations between species, that are either due to marginal co-variations between them or to the effects of the covariates, whereas as the correlations in $\widehat{\boldsymbol{\Sigma}}_1$ are corrected from the effects of covariates. We first observed a reduction of the variances (mean=8.96, median=2.73 in $\widehat{\boldsymbol{\Sigma}}_0$; mean=3.72, median=1.53 in $\widehat{\boldsymbol{\Sigma}}_1$), which proves the strong effect of the covariates on the abundance of the different species. Then we considered the correlations and found a low similarity between their rankings under model M_0 and M_1 (Kendall's $\tau = .40$), showing that the covariates drastically change the apparent relationship between species abundances.

Percentage of variance. We now comment on use of the R_q^2 criterion defined in Section 6 to evaluate the proportion of variability captured by a model with q latent dimensions. R_q^2 compares the pseudo-likelihood ℓ_q^m obtained with q latent dimensions under model M_m ($m = 0, 1$) with the likelihoods $\ell_{\min}^m$ and $\ell_{\max}^m$. We know that $\ell_{\max}^0 = \ell_{\max}^1$ whereas $\ell_{\min}^0 < \ell_{\min}^1$ because $\ell_{\min}^0$ only relies on the offsets whereas $\ell_{\min}^1$ accounts for both the offsets and the covariates. As a consequence, R_q^2 tends to be higher under M_0 than under M_1 for a given q. Right panel of Figure 1b compares the genuine R_q^2 under models M_0 and M_1 and the corrected version of R_q^2 under model M_1 using $\ell_{\min}^0$ in place of $\ell_{\min}^1$. As expected, the corrected version of R_q^2 is always higher under M_1 than under M_0. We also observe that, for both

models, the proportion of variability captured by the latent space is quite high: $R^2_{24} = 95.8\%$ for M_0 and $R^2_{21} = 95.5\%$ for M_1. We remind that $\widehat{q}_0 = 24$ and $\widehat{q}_1 = 21$ should both be compared with $p = 114$.

Variance of the variational posterior.   We remind that S_{ij} is the approximate conditional variance of W_{ij} given the data. This parameter measures the precision of the location of individual i along the j-th latent dimension. We can derive from them the approximate conditional variance of each Z_{ij} as $[\boldsymbol{B}\,\mathrm{diag}(\boldsymbol{s}_i \odot \boldsymbol{s}_i)\boldsymbol{B}^{\mathsf{T}}]_{jj}$. Figure 2, shows that this variance is much higher when the corresponding abundance Y_{ij} is low. Indeed, any large negative negative values of Z_{ij} yields in a Poisson parameter close to zero and, so, to a null Y_{ij}. As a consequence, large negative Z_{ij} can not be predicted accurately.
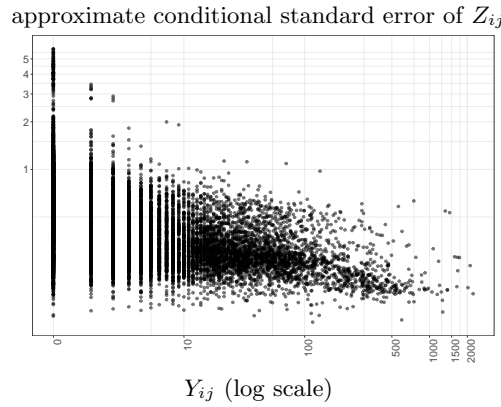

approximate conditional standard error of Z_{ij}

Y_{ij} (log scale)

Fig 2: Variational approximate conditional standard of the Z_{ij} (y axis) as a function of the abundance Y_{ij} (x axis).

7.2. *Impact of weaning on piglet microbiome.*

Description of the experiment.   We considered the metagenomic dataset introduced in Mach et al. (2015), which consists in abundance measures of $p = 4031$ bacterial species collected from the feces of 31 piglets at 5 times points after birth ($n = 155$). One aim of this experiment is to understand the impact of weaning on gut microbiota. The species abundances were measured via metabarcoding (see Mach et al. (2015) for details). We mostly use this example to illustrate of the the proposed methodology behaves when the number of variable p increases.

Numerical Experiments.   To test the impact of the number of variables on the dimension of the latent subspace, we inferred q on nested subsets of the

count table. We selected only the 3000, 2000, 1000, 500 and 100 most abundant species and fitted a model with appropriate offset to each subset. The offsets were chosen as log-total count of each sample. For context, the 2500 least abundant species each have total abundance lower than 5, and more than half (1287) are even seen only once in one sample. As expected, Figure 4 shows that removing low-counts and low-information species increases the dimension of the latent subspace ($\hat{q}$ going up from 4 to 23) and the R^2 (up from 53% to 93%). We are thus able to explain a larger fraction of the variability when considering only the most abundant species rather than all of them. Figure 3 shows that running times increase sublinearly with q and linearly with p.
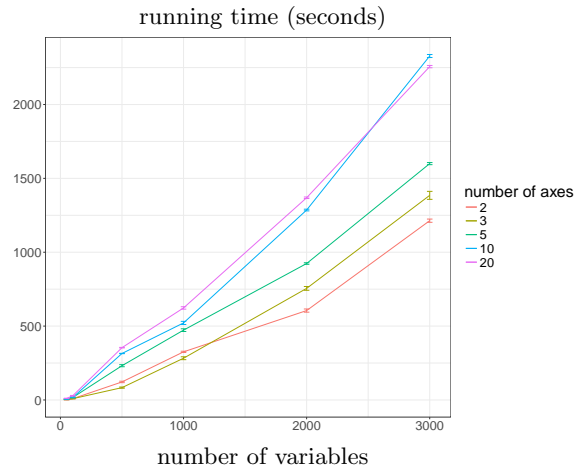

Fig 3: Dataset from Mach et al. (2015). Running times averaged over 4 replicates of the `PLNPCA` function in `R` **PLNModels** package. Single core Intel i7-4600U CPU 2.33GHz, `R` 3.3.3, Linux Ubuntu 16.04.

Impact of Weaning.   We focus on results obtained on the 500 most abundant species, which account for 90.3% of the total counts. The ICL criteria on this subset selects $\hat{q} = 19$ ($R^2 = 86\%$). The main structure present in the latent subspace is the strong and systematic impact of weaning (Fig. 5, left), almost entirely captured by Axis 1. The variable factor map highlights species from two specific bacterial families: Lactobacillaceae (red) and Prevotellaceae (blue). The former are typically found in dairy products and thought to be transmitted to the piglets via breast milk. As expected, they are enriched in suckling piglets and negatively correlated with Axis 1. The latter produce enzymes that are essential to degrade cereals introduced in
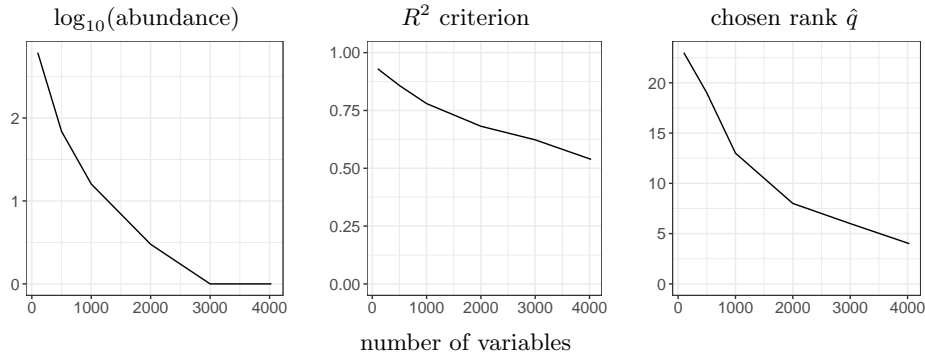
Fig 4: Dataset from Mach et al. (2015). The minimum overall abundance of included species (left panel), quality of approximation R_q^2 (central panel) and selected value $\hat{q}$ (right panel) decrease when species with low abundance are added to the dataset.

the diet after weaning. As reported in Mach et al. (2015), they are enriched after weaning and positively correlated with Axis 1. There is no systematic effect of weaning on Axes 3 and 4 and neither Lactobacillaceae nor Prevotellaceae are highly correlated with those axes.

References.

A. Acharya, J. Ghosh, and M. Zhou. Nonparametric bayesian factor analysis for dynamic count matrices. In *AISTATS*, 2015.

J. Aitchison and C. H Ho. The multivariate Poisson-log normal distribution. *Biometrika*, 76(4):643–653, 1989.

T. W. Anderson. *An introduction to multivariate statistical analysis.* Series in Probability and Statistics. Wiley, 3 edition, 2003.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Machine Intel.*, 22(7): 719–25, 2000.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5): 1190–1208, 1995.

Y. Cao and Y. Xie. Poisson matrix completion. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1841–1845. IEEE, 2015.

M. Collins, S. Dasgupta, and R. E Schapire. A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems*, pages 617–624, 2001.
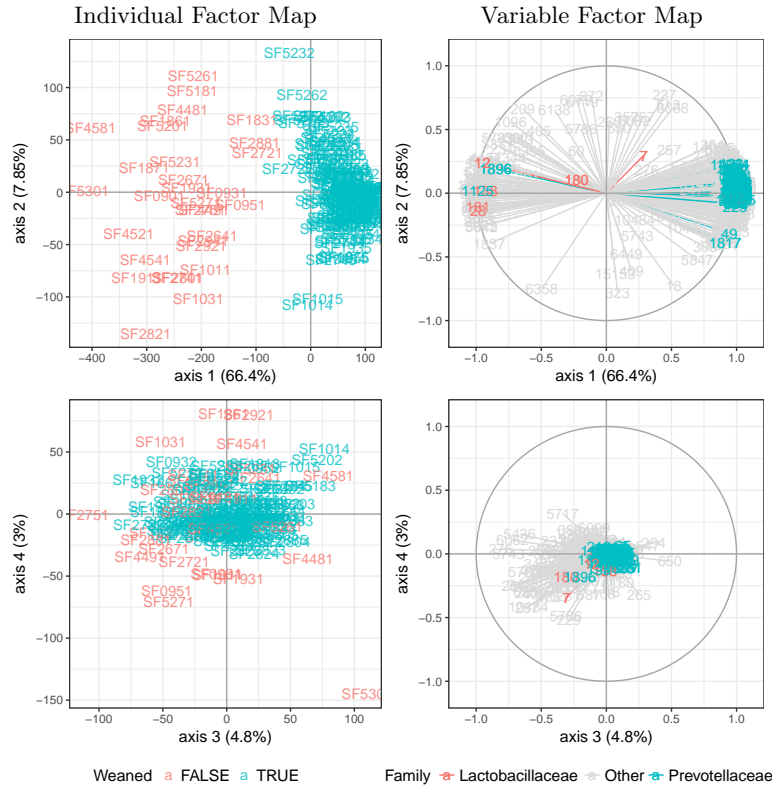
Fig 5: Individual (left) and variable (right) maps corresponding to the first (top, Axes 1 and 2) and second (bottom, Axes 3 and 4) principal plane. Weaning has a strong and systematic effect on gut microbiota composition, well captured by axis 1. Bacterial families Prevotellaceae (red) and the Lactobacillaceae (blue) are two families well known to be affected by weaning and have a high correlation with Axis 1. The latent structure seen in Axes 2, 3 and 4 does not correspond to weaning.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39:1–38, 1977.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936a.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936b. ISSN 1860-0980. . URL http://dx.doi.org/10.1007/BF02288367.

P. Hall, J. T Ormerod, and MP Wand. Theory of gaussian variational approximation for a Poisson mixed model. *Statistica Sinica*, pages 369–389, 2011.

R. Izsák. Maximum likelihood fitting of the Poisson log-normal distribution. *Environmental and Ecological Statistics*, 15(2):143–156, 2008.

T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

B. Jakuschkin, V. Fievet, L. Schwaller, T. Fort, C. Robin, and C. Vacher. Deciphering the pathobiome: Intra-and interkingdom interactions involving the pathogen Erysiphe alphitoides. *Microbial ecology*, pages 1–11, 2016.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete multivariate distributions*, volume 165. Wiley New York, 1997.

D. Karlis. EM algorithm for mixed Poisson and other discrete distributions. *Astin bulletin*, 35(01):3–24, 2005.

J. Lafond. Low rank matrix completion with exponential family noise. *arXiv preprint arXiv:1502.06919*, 2015.

A. J Landgraf and Y. Lee. Dimensionality reduction for binary data through the projection of natural parameters. *arXiv preprint 1510.06112*, 2015.

D. D Lee and H S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

J. Li and D. Tao. Simple exponential family PCA. In *AISTATS*, pages 453–460, 2010.

L. T. Liu, E. Dobriban, and A. Singer. *e*PCA: High Dimensional Exponential Family PCA. *ArXiv e-prints arXiv:1611.05550*, 2016.

N. Mach, M. Berri, J. Estellé, F. Levenez, G. Lemonnier, C. Denis, J.-J. Leplat, C. Chevaleyre, Y. Billon, J. Dor, and et al. Early-life establishment of the swine gut microbiome and impact on host phenotypes. *Environmental Microbiology Reports*, 7(3):554–569, May 2015. ISSN 1758-2229. . URL http://dx.doi.org/10.1111/1758-2229.12285.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate analysis*. Academic press, 1979.

T. P Minka. Automatic choice of dimensionality for PCA. In *NIPS*, volume 13, pages 598–604, 2000.

S. Mohamed, Z. Ghahramani, and K. A Heller. Bayesian exponential family PCA. In *Advances in neural information processing systems*, pages 1089–1096, 2009.

J. F Nelson. Multivariate gamma-poisson models. *Journal of the American Statistical Association*, 80(392):828–834, 1985.

William H Press, Brian P Flannery, Saul A Teukolsky, William T Vetterling, et al. *Numerical recipes*. cambridge University Press, cambridge, third edition edition, 1989.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0.

J A. Royle and C. K Wikle. Efficient statistical mapping of avian count data. *Environmental and Ecological Statistics*, 12(2):225–243, 2005.

J. Salmon, Z. Harmany, C.-A. Deledalle, and R. Willett. Poisson noise reduction with non-local PCA. *Journal of mathematical imaging and vision*, 48(2):279–294, 2014.

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6:461–4, 1978.

S. Srivastava and L. Chen. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic acids research*, 38(17):e170–e170, 2010.

M. E Tipping and C. M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, 2008.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org.

D. M Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

M. Zhou. Nonparametric bayesian negative binomial factor analysis. *arXiv preprint arXiv:1604.07464*, 2016.

M. Zhou, L. Hannah, D. B Dunson, and L. Carin. Beta-negative binomial process and poisson factor analysis. In *AISTATS*, volume 22, pages 1462–1471, 2012.

APPENDIX A: CONVEXITY LEMMAS

LEMMA 1. *For any vectors $\boldsymbol{\theta}$, $\boldsymbol{x}$, $\boldsymbol{m}$, $\boldsymbol{s}$ and $\boldsymbol{b}$ (with matching dimensions) and convex function f, if $\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{I})$ and $\boldsymbol{w} = \boldsymbol{m} + \boldsymbol{s} \odot \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{m}, \mathrm{diag}(\boldsymbol{s} \odot \boldsymbol{s}))$, then the map $g : (\boldsymbol{\theta}, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{b}) \mapsto \mathbb{E}[f(\boldsymbol{\theta}^\mathsf{T}\boldsymbol{x} + \boldsymbol{b}^\mathsf{T}\boldsymbol{w})]$ is convex in $(\boldsymbol{\theta}, \boldsymbol{b})$ for $(\boldsymbol{m}, \boldsymbol{s})$ fixed and vice-versa.*

PROOF. Note $Z = \boldsymbol{\theta}^\mathsf{T}\boldsymbol{x} + \boldsymbol{b}^\mathsf{T}\boldsymbol{w} = (\boldsymbol{\theta}^\mathsf{T}\boldsymbol{x} + \boldsymbol{b}^\mathsf{T}\boldsymbol{m}) + \boldsymbol{b}^\mathsf{T}(\boldsymbol{s} \odot \boldsymbol{u})$. The first order derivative of g is

$$\boldsymbol{\nabla}(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{m}, \boldsymbol{s}) = \mathbb{E}\left[f'(Z) \begin{bmatrix} \boldsymbol{x} & \boldsymbol{m} + \boldsymbol{s} \odot \boldsymbol{u} & \boldsymbol{b} & \boldsymbol{b} \odot \boldsymbol{u} \end{bmatrix}^\mathsf{T} \right].$$

The second order partial derivatives of g are:

$$\boldsymbol{\Psi}_1(\boldsymbol{\theta}, \boldsymbol{b}) = \mathbb{E}\left[f''(Z) \begin{bmatrix} \boldsymbol{x}\boldsymbol{x}^\mathsf{T} & \boldsymbol{x}(\boldsymbol{m} + \boldsymbol{u} \odot \boldsymbol{s})^\mathsf{T} \\ (\boldsymbol{m} + \boldsymbol{s} \odot \boldsymbol{u})\boldsymbol{x}^\mathsf{T} & (\boldsymbol{m} + \boldsymbol{s} \odot \boldsymbol{u})(\boldsymbol{m} + \boldsymbol{s} \odot \boldsymbol{u})^\mathsf{T} \end{bmatrix} \right]$$

$$\boldsymbol{\Psi}_2(\boldsymbol{m}, \boldsymbol{s}) = \mathbb{E}\left[f''(Z) \begin{bmatrix} \boldsymbol{b}\boldsymbol{b}^\mathsf{T} & \boldsymbol{b}(\boldsymbol{b} \odot \boldsymbol{u})^\mathsf{T} \\ (\boldsymbol{b} \odot \boldsymbol{u})\boldsymbol{b}^\mathsf{T} & (\boldsymbol{b} \odot \boldsymbol{u})(\boldsymbol{b} \odot \boldsymbol{u})^\mathsf{T} \end{bmatrix} \right]$$

And the associated quadratic form $\Phi_1(\boldsymbol{v}, \boldsymbol{w}) = (\boldsymbol{v}, \boldsymbol{w})^\mathsf{T}\boldsymbol{\Psi}_1(\boldsymbol{\theta}, \boldsymbol{b})(\boldsymbol{v}, \boldsymbol{w})$ and $\Phi_2(\boldsymbol{v}, \boldsymbol{w}) = (\boldsymbol{v}, \boldsymbol{w})^\mathsf{T}\boldsymbol{\Psi}_2(\boldsymbol{m}, \boldsymbol{s})(\boldsymbol{v}, \boldsymbol{w})$ can be simplified to

$$\Phi_1(\boldsymbol{v}, \boldsymbol{w}) = \mathbb{E}[f''(Z)(\boldsymbol{x}^\mathsf{T}\boldsymbol{v} + (\boldsymbol{m} + \boldsymbol{s} \odot \boldsymbol{u})^\mathsf{T}\boldsymbol{w})^2] \geq 0$$

$$\Phi_2(\boldsymbol{v}, \boldsymbol{w}) = \mathbb{E}[f''(Z)(\boldsymbol{b}^\mathsf{T}\boldsymbol{v} + (\boldsymbol{b} \odot \boldsymbol{u})^\mathsf{T}\boldsymbol{w})^2] \geq 0$$

The Hessians $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ are thus semidefinite positive, which ends the proof. $\square$

LEMMA 2.   *For any matrices $\boldsymbol{\Theta}$, $\boldsymbol{X}$, $\boldsymbol{M}$, $\boldsymbol{S}$ and $\boldsymbol{B}$ (with matching dimensions) and convex function f, if $\boldsymbol{U} = [\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n]^\intercal$ where the $\boldsymbol{U}_i$ are i.i.d and $\boldsymbol{U}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{W} = \boldsymbol{M} + \boldsymbol{S} \odot \boldsymbol{U}$. The map $g : (\boldsymbol{\Theta}, \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{B}) \mapsto \boldsymbol{1}_n^\intercal \mathbb{E}[f(\boldsymbol{X}\boldsymbol{\Theta}^\intercal + \boldsymbol{W}\boldsymbol{B}^\intercal)]\boldsymbol{1}_p$ is convex in $(\boldsymbol{\Theta}, \boldsymbol{B})$ for $(\boldsymbol{M}, \boldsymbol{S})$ fixed and vice-versa.*

PROOF.   The function g is a sum of functions of the form $g_{ij} : (\boldsymbol{\Theta}, \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{B}) \mapsto \mathbb{E}[f(\boldsymbol{X}_i^\intercal \boldsymbol{\Theta}_j + \boldsymbol{B}_j^\intercal(\boldsymbol{M}_i + \boldsymbol{S}_i \odot \boldsymbol{U})]$. The result follows from Lemma 1. $\square$

MIA-Paris
UMR 518 AgroParisTech / INRA
AgroParisTech
16, rue Claude Bernard
75231 Paris CEDEX 05, France
E-mail: julien.chiquet@inra.fr; stephane.robin@inra.fr

INRA Unité MaIAGE
Bt. 233 et 210
Domaine de Vilvert
78352 JOUY-EN-JOSAS CEDEX, France
E-mail: mahendra.mariadassou@inra.fr