Biometrics 74, 1311–1319
December 2018

DOI: 10.1111/biom.12888

# Order Selection and Sparsity in Latent Variable Models via the Ordered Factor LASSO

**Francis K. C. Hui**[1,*] **Emi Tanaka**[2] **and David I. Warton**[3]

[1]Mathematical Sciences Institute, The Australian National University, Acton, ACT 2601, Australia
[2]School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia
[3]School of Mathematics and Statistics, and the Evolution & Ecology Research Centre, UNSW Sydney, NSW 2052, Australia
[*]*email:* francis.hui@anu.edu.au

Summary. Generalized linear latent variable models (GLLVMs) offer a general framework for flexibly analyzing data involving multiple responses. When fitting such models, two of the major challenges are selecting the order, that is, the number of factors, and an appropriate structure for the loading matrix, typically a sparse structure. Motivated by the application of GLLVMs to study marine species assemblages in the Southern Ocean, we propose the Ordered Factor LASSO or OFAL penalty for order selection and achieving sparsity in GLLVMs. The OFAL penalty is the first penalty developed specifically for order selection in latent variable models, and achieves this by using a hierarchically structured group LASSO type penalty to shrink entire columns of the loading matrix to zero, while ensuring that non-zero loadings are concentrated on the lower-order factors. Simultaneously, individual element sparsity is achieved through the use of an adaptive LASSO. In conjunction with using an information criterion which promotes aggressive shrinkage, simulation shows that the OFAL penalty performs strongly compared with standard methods and penalties for order selection, achieving sparsity, and prediction in GLLVMs. Applying the OFAL penalty to the Southern Ocean marine species dataset suggests the available environmental predictors explain roughly half of the total covariation between species, thus leading to a smaller number of latent variables and increased sparsity in the loading matrix compared to a model without any covariates.

Key words: Dimension reduction; Factor analysis; Generalized linear latent variable models; LASSO; Loadings; Penalized likelihood; Regularization.

## 1. Introduction

Generalized linear latent variable models (GLLVMs) offer a general and attractive framework for model-based dimension reduction and parsimoniously modeling correlations between multiple responses (Skrondal and Rabe-Hesketh, 2004). This article is motivated by the analysis of multivariate abundance data in ecology, collected as part of the Southern ocean continuous plankton recorder (SO-CPR) survey on marine assemblages using vessels traversing the Indian sector of the Southern Ocean (Hosie et al., 2003). We use GLLVMs to study the correlations between species in the marine assemblage, for example, unobserved covariates, biotic interactions, phylogeny, and to understand the influence of environmental variables driving the species community while accounting for these correlations (see Warton et al., 2015, 2016, for other recent applications of GLLVMs in ecology).

One of the major challenges when fitting and performing inference with GLLVMs is selecting the order of the model, that is, the number of factors or latent variables. The number of latent variables is rarely known a priori, and instead a data-driven approach is required to select the appropriate order. A considerable amount of research has been devoted to this problem, ranging from heuristic methods such as selecting the minimum order which ensures the overall proportion of variance explained by the factors exceeds some arbitrary threshold (Smith et al., 2015), to statistically driven approaches such as hypothesis tests and information criteria (Bai and Ng, 2002). Approaches such as information criteria are well-grounded, but a potential weakness is that they treat the order selection problem as a discrete process: a range of orders are tried and the one optimizing the proposed criterion is selected. Like subset selection in regression then, this discreteness may lead to high variability and instability in the selection (Hastie et al., 2009).

Aside from selecting the order, sparsity in the loading matrix is also often desirable for reasons of interpretation, for example, if the factors are regarded as unobserved environmental covariates, then each species may only be influenced by a subset of these unobserved variables. Traditionally, sparsity in the loading matrix is obtained after estimation as part of a two stage approach. The order of the GLLVM is first selected and the associated, unstructured loading matrix is estimated using maximum likelihood, for instance. Afterward, some rotation is applied to the matrix to achieve structure and sparsity, with the most popular being the varimax rotation (Kaiser, 1958). In the context of maximum likelihood estimation, which we focus on in this article, any orthogonal rotation does not change the value of the likelihood for the fitted model, and thus the choice of rotation matrix is, in many ways, arbitrary. To overcome this problem, and motivated

by the rising popularity of methods such as the least absolute selection and shrinkage operator (LASSO, Tibshirani, 1996) for regression modeling, there has been growing interest in using penalized likelihood methods to achieve sparsity in the loading matrix (e.g., Choi et al., 2010; Hirose and Konishi, 2012; Hirose and Yamamoto, 2014, 2015). Such an approach is computationally and conceptually attractive: efficient coordinate-wise optimization algorithms can be used to construct the regularization path, while the inclusion of a penalty leads to a single stage method where estimation and sparsity is achieved simultaneously. Empirical studies have also shown that penalized likelihood tends to outperform rotation methods at recovering the true underlying structure of the loading matrix (Hirose and Yamamoto, 2015). On the other hand, while penalties has been proposed for achieving sparsity, none have explicitly addressed the critical issue of order selection. That is, while some penalties can achieve dimension selection as a by-product, for example, if all the individual coefficients corresponding to a factor are shrunk to zero (Choi et al., 2010), no penalty has been developed so far which directly performs order selection in GLLVMs.

In this article, we propose the Ordered FActor LASSO (OFAL) penalty for order selection and achieving sparsity in GLLVMs. The proposed approach is novel in two important ways: it is the first penalty developed specifically for order selection in GLLVMs. This is achieved by utilizing a group sparsity approach which encourages all the coefficients corresponding to a particular latent variable to be shrunk to zero simultaneously. To deal with the added complication of the arbitrary ordering of the factors, the OFAL penalty further imposes a hierarchical structure such that a particular column of the loading matrix is entirely shrunk to zero only if all the so-called higher order loadings have already been set to zero. This hierarchical structure in OFAL ensures the fitted GLLVM is concentrated on the lower order loadings, that is, first few factors. Similar to Choi et al. (2010) and Hirose and Yamamoto (2015), we also incorporate a component in the OFAL penalty which encourages individual coefficient sparsity in the loading matrix, in order to facilitate interpretation and improve prediction. A second key contribution, motivated by our application, is that we develop our penalty in a broader context to cover non-normal responses. Previous research, including those reviewed above, have exclusively assumed normally distributed responses. We first develop a result which allows us to reparameterize the OFAL penalty into a elastic-net type regularization problem (Zou and Hastie, 2005), thus facilitating the use of coordinate-ascent algorithms. We then combine this result with an Expectation–Maximization (EM) algorithm (Rubin and Thayer, 1982), where for dealing with overdispersed species counts assumed to be negative binomially distributed, we develop novel minorizing functions which lead to closed form updates of the loading matrix and regression coefficients.

Regarding the critical choice of the tuning parameter in the OFAL penalty, we propose using the Extended Regularized Information Criterion (ERIC, Hui et al., 2015b), which employs a dynamic model complexity penalty to promote "aggressive" shrinkage of the latent variable coefficients. This is appropriate because, for order selection, we anticipate most of the elements of the loading matrix to be shrunk to zero and thus a relatively high degree of sparsity. Simulation studies show that the OFAL penalty, in conjunction with ERIC, performs strongly compared with traditional methods and other penalties for order selection, sparsity, and prediction. Applying the OFAL penalty to the SO-CPR survey reveals that environmental predictors explain roughly half of the total covariation between species (as based on the change in the trace of estimated residual covariance matrix), thus leading to a smaller number of latent variables and further sparsity in the loading matrix compared to a model without any covariates.

To summarize, the main contributions of this article are as follows: 1) we propose the first penalized likelihood method for simultaneous order selection and achieving sparsity in GLLVMs. The OFAL penalty utilizes both group and hierarchical sparsity to shrink entire columns of the loading matrix to zero, while concentrating the non-zero elements on the lower order loadings; 2) we propose an estimation procedure based on reformulating the OFAL penalty as a elastic-net type regularization problem ; 3) we develop computationally efficient updates for the specific case of negative binomial responses; 4) simulations demonstrate that, when combined with an aggressive information criterion for choosing the tuning parameter, OFAL performs strongly compared to other currently available methods for order selection and/or achieving sparsity in GLLVMs. We provide `R` code for fitting GLLVMs with the OFAL penalty in Web Appendix F.

## 2. Latent Variable Models

We first introduce the basic GLLVM before discussing the specific application to overdispersed species counts in ecology. Consider a set of $n$ observational units $\{(\boldsymbol{x}_i, \boldsymbol{y}_i); i = 1, \ldots, n\}$, where $\boldsymbol{y}_i$ denotes a $p$-vector of responses and $\boldsymbol{x}_i$ is a $q$-vector of covariates, including an intercept term as its first element. Conditional on a set of $d \ll p$ latent variables $\boldsymbol{u}_i$, the elements of $\boldsymbol{y}_i$, denoted here as $y_{ij}$, are assumed to be independent observations from the exponential family of distributions with mean $\mu_{ij}$ and dispersion parameter $\psi_j$. We assume the $p$ elements are of the same response type, although the developments below can be straightforwardly extended to the case of mixed responses, for example, combining multivariate abundance datasets involving presence–absence and species counts. The mean is regressed against the covariates and latent variables as $g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{u}_i^\top \boldsymbol{\lambda}_j$, where $g(\cdot)$ is a known link function, $\boldsymbol{\beta}_j$ is a $q$-vector of response-specific coefficients associated with the covariates $\boldsymbol{x}_i$, and $\boldsymbol{\lambda}_j$ is a $d$-vector of loadings for response $j$. The latent variables are assumed to come from a multivariate standard normal distribution, $\boldsymbol{u}_i \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{I}_d)$, where the zero mean vector and identity covariance matrix are used to fix the location and scale of the model. Importantly, the latent variables induce a correlation between the responses on the linear predictor scale: if we let $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1 \ldots \boldsymbol{\lambda}_p)^\top$ denote the $p \times d$ loading matrix and $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ip})$, then $\text{Cov}(\boldsymbol{\eta}_i | \boldsymbol{x}_i) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$. That is, the covariance between responses is modeled parsimoniously via rank reduction with rank $d$.

The standard factor analytic model is a special case of GLLVMs where the responses are assumed to be normally distributed. We write $y_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{u}_i^\top \boldsymbol{\lambda}_j + \epsilon_{ij}$, where

$\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ip}) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Psi})$ and $\boldsymbol{\Psi} = \text{Diag}(\boldsymbol{\psi})$ with $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)$. It follows that $\text{Cov}(\boldsymbol{y}_i|\boldsymbol{x}_i) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p, \text{vec}(\boldsymbol{\Lambda}^\top), \boldsymbol{\psi})$ denote the vector of all parameters. Based on the formulation above, the marginal log-likelihood for the GLLVM is then given by $\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log(\int \prod_{j=1}^p f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{x}_i, \boldsymbol{\theta}) f(\boldsymbol{u}_i)\, d\boldsymbol{u}_i)$. For estimation purposes (see Section 4), we will also make use of the complete log-likelihood, which, ignoring constants with respect to the parameters, is given by

$$\begin{aligned}\ell_c(\boldsymbol{\theta}) &= \sum_{i=1}^n \sum_{j=1}^p \log f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{x}_i, \boldsymbol{\theta}) + \sum_{i=1}^n \log f(\boldsymbol{u}_i) \\ &= \sum_{i=1}^n \sum_{j=1}^p \left\{ \frac{1}{\psi_j} \left\{ y_{ij}\kappa_{ij} - b(\kappa_{ij}) \right\} + c(y_{ij}, \kappa_{ij}) \right\} \\ &\quad -\frac{1}{2} \sum_{i=1}^n \boldsymbol{u}_i^\top \boldsymbol{u}_i, \end{aligned} \tag{1}$$

for known functions $b(\cdot)$ and $c(\cdot)$ and canonical parameter $\kappa_i$, such that $\text{E}(y_{ij}|\boldsymbol{u}_i, \boldsymbol{x}_i) = \mu_{ij} = b'(\kappa_{ij})$ and $\text{Var}(y_{ij}|\boldsymbol{u}_i, \boldsymbol{x}_i) = \psi_j b''(\kappa_{ij})$.

### 2.1. *Multivariate Abundance Data*

Due to most species being rarely observed, multivariate abundance data in ecology are characterized by discrete responses with lots of zeros. For example, the SO-CPR survey comprises records of 27 marine species, with 16 of those species found at less than 20% of the 54 sites. Let the $n$ observational units be the sites visited and the $p$ responses correspond to the species recorded. Also, let $\boldsymbol{x}_i$ represent a vector of environmental covariates which we want to perform inference on while accounting for additional correlation between the $p$ species. To apply GLLVMs to multivariate abundance data, we can adopt the general formulation in Section 2 with species-specific regression coefficients $\boldsymbol{\beta}_j$ for $j = 1, \ldots, p$. Furthermore, for the case of overdispersed counts in the SO-CPR survey, we assume a negative binomial distribution with a log link function, such that $\text{Var}(y_{ij}|\boldsymbol{u}_i, \boldsymbol{x}_i) = \mu_{ij} + \psi_j \mu_{ij}^2$ and $\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{j=1}^p \{y_{ij} \log(\psi_j \mu_{ij}) - (y_{ij} + \psi_j^{-1}) \log(1 + \psi_j \mu_{ij}) + \log \Gamma(y_{ij} + \psi_j^{-1}) - \log \Gamma(\psi_j^{-1})\} - 2^{-1} \sum_{i=1}^n \boldsymbol{u}_i^\top \boldsymbol{u}_i$, ignoring constants, where $\log \Gamma(\cdot)$ is the log-Gamma function and $\mu_i = \exp(\eta_i)$.

## 3. The Ordered Factor LASSO Penalty

For the developments below, we focus on the general form of the GLLVM defined in Section 2. The extension to the case of negative binomial responses is straightforward and only requires additional tedious algebra. Let $\lambda_{jk}$ denote element $(j, k)$ in the loading matrix $\boldsymbol{\Lambda}$, which corresponds to the coefficient relating response $j = 1, \ldots, p$ to latent variable $k = 1, \ldots, d$. To simultaneously select the dimension of the model and achieve sparsity in the loading matrix, we propose the following penalized likelihood approach.

DEFINITION 1. *Consider the latent variable model defined in Section 2. Then for a single tuning parameter $s > 0$, the Ordered FActor LASSO (OFAL) estimator is defined as*

$$\begin{aligned}\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}}\ \ell(\boldsymbol{\theta}) &- ns \sum_{l=1}^d w_{1l} \left( \sum_{k=l}^d \sum_{j=1}^p \lambda_{jk}^2 \right)^{1/2} \\ &- ns \sum_{j=1}^p \sum_{k=1}^d w_{2jk} |\lambda_{jk}|. \end{aligned}$$

*where $\ell(\boldsymbol{\theta})$ is the marginal likelihood, with the corresponding complete likelihood given by (1), and $\{w_{1l}; l = 1, \ldots, d\}$ and $\{w_{2jk}; j = 1, \ldots, p; k = 1, \ldots, d\}$ are two sets of positive adaptive weights constructed a priori.*

The OFAL penalty is additive in design, and consists of two components designed to achieve dimension and sparsity in the loadings, respectively. Note these two goals can be achieved using other types of penalty designs, for example, single bilevel penalties (e.g., Huang et al., 2012; Hui et al., 2015a). However, we choose to work with an additive structure and two convex penalties as this lends itself to a computationally efficient estimation algorithm; see Section 4.

The first component of the OFAL penalty aims to achieve order selection by exploiting the group structure of the loadings in the columns in $\boldsymbol{\Lambda}$. That is, by encouraging all the coefficients in a particular column $k$ of the loading matrix to be shrunk to zero simultaneously, it amounts to removing factor $k$ from the model. However, there is an additional complication brought about by the fact that the GLLVM is invariant with respect to the order of the columns in $\boldsymbol{\Lambda}$ (and correspondingly, the order of the elements in $\boldsymbol{u}_k$). To overcome this, the OFAL penalty imposes an additional hierarchical structure which ensures that elements in column $k$ of $\boldsymbol{\Lambda}$ are simultaneously shrunk to zero *only if* all the elements in columns $k' = k + 1, \ldots, d$, that is, the so-called higher order loadings, are shrunk to zero as well. A simple illustration of this can be seen if we consider the case of $d = 2$. Ignoring the tuning parameter, the first component of OFAL is given by $w_{11} \left( \sum_{j=1}^p \lambda_{j1}^2 + \lambda_{j2}^2 \right)^{1/2} + w_{12} \left( \sum_{j=1}^p \lambda_{j2}^2 \right)^{1/2}$. From this, we can see that either the second factor is removed from the model, that is, $\lambda_{j2} = 0$ for all $j = 1, \ldots, p$ , or both factors are removed from the model, that is, $\lambda_{j1} = \lambda_{j2} = 0$ for all $j = 1, \ldots, p$. That is, by design the penalty does not permit the first factor to be removed while retaining the second factor in the model. This hierarchical design of OFAL guarantees the resulting latent variable model is concentrated on the first few or lower order loadings. For practical applications, we may regard $d$ as the maximum number of latent variables we want in the model, and set it to some sufficiently large integer based on scientific and/or interpretation grounds. The estimated GLLVM from Definition 1 will then select $d' \leq d$ factors. For example, with the SO-CPR dataset containing $p = 27$ species, it is believed that species responses are heavily driven by the few available environmental predictors, and so we expect $d = 8$ is more than sufficient to model any residual covariation. We point out that choosing a very high value of $d$ can be problematic as the estimates from the (close to) unpenalized GLLVM fit may be unstable due to overfitting.

The second component of the OFAL penalty applies a weighted $L_1$ norm on the loading matrix $\mathbf{\Lambda}$ in order to achieve sparsity. Such individual coefficient sparsity has been proposed for factor analysis before by Hirose and Yamamoto (2015) among others, and reflects the idea that each response may only be informed by a subset of latent factors, that is, the true loading matrix is sparse. It is also an attractive way to produce a unique solution in the model, as the $L_1$ penalty and hence $\hat{\boldsymbol{\theta}}$ is not invariant under an orthogonal rotation of the loadings (Choi et al., 2010).

The inclusion of adaptive weights $w_{1l}$ and $w_{2jk}$ serves two main purposes. First, it allows us to employ only a single tuning parameter in constructing the OFAL estimator, which simplifies computation when building the regularization path compared to other potential penalties that will require two or more tuning parameters. That is, if weights are not included (or equivalently $w_{1l} = w_{2jk} = 1$ for all $j, k, l$), then separate tuning parameters will be required for the two component penalties to ensure differential regularization (analogous to the original elastic-net penalty, Zou and Hastie, 2005). Second, and related to the first point, is that the inclusion of adaptive weights facilitates order and sparsity selection consistency, although we reserve any formal asymptotic development for future research (although we expect the proofs and results to be similar to Zou, 2006, for instance). In particular, we design the adaptive weights to exhibit differing asymptotic behavior for truly zero versus non-zero loadings and loading elements; see Section 4 for further details on their construction.

Along the same lines as above, we acknowledge that it is possible to include an additional, "mixing" tuning parameter in the OFAL penalty to balance the ratio between the two penalty components. That is, analogous to the implementation of the elastic-net in the `glmnet` package (Friedman et al., 2010), we can extend Definition 1 to be of the form $ns\alpha \sum_{l=1}^{d} w_{1l} \left( \sum_{k=l}^{d} \sum_{j=1}^{p} \lambda_{jk}^2 \right)^{1/2} - ns(1-\alpha) \sum_{j=1}^{p} \sum_{k=1}^{d} w_{2jk} |\lambda_{jk}|$ where $\alpha \in [0, 1]$. Note $\alpha = 1$ corresponds then to pure order selection penalty, while $\alpha = 0$ corresponds to an adaptive LASSO penalty. However, we have chosen not to include $\alpha$ (or equivalently, fix $\alpha = 2^{-1}$) for two reasons. First, similar to why adaptive weights are included, the computational burden of having to perform a two or higher dimensional grid search to find the optimal set of tuning parameters may be impractical for some applications. Second, we point out that the $\alpha$ parameter has a specific role in the elastic-net penalty, namely to balance out sparse (LASSO) and non-sparse (ridge) penalties acting on the same coefficient. By contrast, the OFAL penalty is an additive combination of two sparse penalties which are acting on groups of and individual coefficients, respectively. Therefore, there is arguably less motivation for its inclusion.

## 4. Estimation

To facilitate estimation of latent variable models with the OFAL penalty, we make use of the following result, whose proof may be found in the Web Appendix A.

LEMMA 1. *The OFAL estimates given in Definition 1 can be obtained by solving the equivalent optimization problem*

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\tau}}) = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\tau} \geq 0} \ \ell(\boldsymbol{\theta}) - \frac{(ns)^2}{4} \sum_{l=1}^{d} \frac{w_{1l}}{\tau_l} \sum_{k=l}^{d} \sum_{j=1}^{p} \lambda_{jk}^2 - ns \sum_{j=1}^{p} \sum_{k=1}^{d} w_{2jk} |\lambda_{jk}| - \sum_{l=1}^{d} w_{1l} \tau_l,$$

*where* $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_d)$. *That is, if* $\hat{\boldsymbol{\theta}}$ *is a local maximizer of Definition 1 then there exists a local maximizer* $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\tau}})$ *of the above such that* $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$. *Similarly, if* $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\tau}})$ *is a local maximizer of the above then* $\hat{\boldsymbol{\theta}}$ *is also a local maximizer of Definition 1.*

The above result reformulates OFAL into an elastic-net type regularization problem, and means that conditional on $\boldsymbol{\tau}$, we can employ coordinate-wise optimization to obtain sparse loading estimates. Specifically, we combine Lemma 1 with an Expectation–Maximization (EM) algorithm as follows. Define the penalized complete log-likelihood as $\ell_{c,\text{pen}}(\boldsymbol{\theta}) = \ell_c(\boldsymbol{\theta}) - 4^{-1}(ns)^2 \sum_{l=1}^{d} \tau_l^{-1} w_{1l} \sum_{k=l}^{d} \sum_{j=1}^{p} \lambda_{jk}^2 - ns \sum_{j=1}^{p} \sum_{k=1}^{d} w_{2jk} |\lambda_{jk}| - \sum_{l=1}^{d} w_{1l} \tau_l$, where $\ell_c(\boldsymbol{\theta})$ is given by equation (1). Note that if we let $\nu_k = \sum_{l=1}^{k} w_{1l} \tau_l^{-1}$, then we can rewrite the second term in the above as $4^{-1}(ns)^2 \sum_{k=1}^{d} \sum_{j=1}^{p} \nu_k \lambda_{jk}^2$. This form makes explicit the hierarchical nature of the penalty: the quantities $\nu_k$ satisfy $0 < \nu_1 < \nu_2 \ldots < \nu_d$, ensuring that the higher order factors are penalized more heavily than their lower order counterparts.

The proposed EM algorithm then involves iterating between two steps until convergence, as summarized in Algorithm 1. We provide details of the E- and M-steps in Web Appendix B, focusing on negative binomial GLLVMs. Specifically, to perform the E-step we use Monte-Carlo integration, while in CM-steps 1 and 3, in order to overcome the non-linear nature of $\ell_c(\boldsymbol{\theta})$ with respect to the $\boldsymbol{\beta}_j$'s and $\lambda_{jk}$'s, respectively, we propose minorizing functions which allow computationally efficient, tractable updates based only on the posterior expectation of relatively simple quantities.

We now discuss one approach to constructing the adaptive weights $w_{1l}$ and $w_{2jk}$. As discussed in Section 3, the weights both help to simplify computation (we require only one tuning parameter), and facilitate potential selection consistency properties (the weights are designed to exhibit differing large sample behavior for truly zero versus non-zero loadings). Unlike standard adaptive lasso regression, it is not immediately obvious that we can build these weights based on the full model with $d$ candidate factors, since without any constraints or penalization the loading matrix $\mathbf{\Lambda}$ is unidentifiable due to rotational invariance. Therefore, weights constructed from the estimated $\mathbf{\Lambda}$ may conflict with the hierarchical nature of the OFAL penalty because the columns of $\mathbf{\Lambda}$ from the full model are not ordered in any way, and so (for example) they may be arranged such that the dense loadings are concentrated on the higher order columns instead of the lower order columns. This is inconsistent with the design of the OFAL penalty, where we want the dense (sparse) loadings to be concentrated on the lower (higher) order columns. To overcome this problem, we propose the following approach to building

**Algorithm 1** EM algorithm for estimation GLLVMs with the OFAL penalty.

**repeat**

**E-step** Calculate the Q-function $Q(\boldsymbol{\theta}) = \mathrm{E}(\ell_c(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{x}, \hat{\boldsymbol{\theta}}^{(r)})$, where the expectation is with respect to the posterior distribution $f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{x}, \hat{\boldsymbol{\theta}}^{(r)})$.

**M-step** Obtain updated estimates $\hat{\boldsymbol{\theta}}^{(r+1)}$ based on maximizing

$$Q(\boldsymbol{\theta}) - \frac{(ns)^2}{4}\sum_{l=1}^{d}\frac{w_{1l}}{\tau_l}\sum_{k=l}^{d}\sum_{j=1}^{p}\lambda_{jk}^2 - ns\sum_{j=1}^{p}\sum_{k=1}^{d}w_{2jk}|\lambda_{jk}| - \sum_{l=1}^{d}w_{1l}\tau_l.$$

This can be achieved using the following conditional maximization steps, for instance.

**CM-step 1** Update the dispersion parameters $\hat{\boldsymbol{\psi}}^{(r+1)}$ and coefficients $\hat{\boldsymbol{\beta}}_j^{(r+1)}$ based on maximizing $Q(\boldsymbol{\theta})$.

**CM-step 2** For $l = 1, \ldots, d$, update $\hat{\tau}_l^{(r+1)} = 2^{-1}ns\left(\sum_{k=l}^{d}\sum_{j=1}^{p}\hat{\lambda}_{jk}^{(r)2}\right)^{1/2}$. Then for $k = 1, \ldots, d$, calculate $\hat{v}_k^{(r+1)} = \sum_{l=1}^{k} w_{1l}(\hat{\tau}_l^{(r+1)})^{-1}$.

**CM-step 3** For $j = 1, \ldots, p$ and $k = 1, \ldots, d$, update the loadings $\hat{\lambda}_{jk}^{(r+1)}$ coordinate-wise based on solving an elastic-net type optimization problem, $Q(\boldsymbol{\theta}) - 4^{-1}(ns)^2\hat{v}_k^{(r+1)}\lambda_{jk}^2 - nsw_{2jk}|\lambda_{jk}|$.

**until** Convergence criterion met e.g., $\|\hat{\boldsymbol{\theta}}^{(r+1)} - \hat{\boldsymbol{\theta}}^{(r)}\| < \epsilon$ for some small $\epsilon > 0$.

the adaptive weights. First, we fit the full GLLVM and obtain the estimated loading matrix, denoted here as $\check{\boldsymbol{\Lambda}}$. Next, we perform an eigendecomposition on the associated covariance matrix $\check{\boldsymbol{\Lambda}}\check{\boldsymbol{\Lambda}}^\top = \check{\boldsymbol{Q}}\check{\boldsymbol{D}}\check{\boldsymbol{Q}}^\top$. By definition, only the first $d$ eigenvalues of the diagonal matrix $\check{\boldsymbol{D}}$ will be positive (and arranged in descending order), while the remaining $p - d$ eigenvalues will be zero. We thus calculate a new loading matrix $\boldsymbol{\Lambda}^*$ by taking only the first $d$ columns of $\check{\boldsymbol{Q}}\check{\boldsymbol{D}}^{1/2}$. Let $\lambda_{jk}^*$ denote element $(j, k)$ of $\boldsymbol{\Lambda}^*$. Then, we construct the adaptive weights as

$$w_{1l} = \left(\sum_{k=l}^{d}\sum_{j=1}^{p}(\lambda_{jk}^*)^2\right)^{-1/2} = \left(\sum_{k=l}^{d}\check{D}_{kk}\right)^{-1/2}; l = 1, \ldots, d$$

$$w_{2jk} = |\lambda_{jk}^*|^{-1}; j = 1, \ldots, p; k = 1, \ldots, d,$$

where $\check{D}_{kk}$ is the $k$-th eigenvalue in $\check{D}$. Note it is possible for the weights to also depend on an additional power parameter $\gamma > 0$, for example, $w_{2jk} = |\lambda_{jk}^*|^{-\gamma}$. However based on empirical testing, we found the above forms for the adaptive weights worked quite well, but acknowledge the possibility of this extension in future research (although this will increase computational burden due to heaving to search over multiple tuning parameters). More generally, future research could also examine other approaches to constructing the adaptive weights, such as choosing $d$ in a data driven manner and applying a varimax rotation to $\check{\boldsymbol{\Lambda}}$ to induce some prior structure. More importantly, by performing an eigendecomposition before constructing the adaptive weights, it ensures that the weights are consistent with the hierarchical nature of the OFAL penalty, for example, it is straightforward to show that $0 < w_{11} < \ldots < w_{1d}$.

To build the regularization path, we first determine a value $s_{\max}$ for which the OFAL penalty shrinks all the loadings $\lambda_{jk}$ to zero. Then, we construct a decreasing sequence of values on the log scale on the interval from $s_{\max}$ to $0.001s_{\max}$ (in the simulations and applications we used a grid of 1000 values), and calculate the OFAL estimates for each value on the sequence. The most appropriate value of $s$, and hence the best fitted GLLVM is then determined using the information criterion detailed in Section 4.1.

### 4.1. *Tuning Parameter Selection*

Given both the large number of parameters available for selection ($pd$) and the aim being to select only a small number of latent variables, that is, we anticipate most the $d$ columns of $\boldsymbol{\Lambda}$ will be shrunk to zero, then we advocate using an information criterion for choosing the tuning parameter $s$ which facilitates aggressive shrinkage and aims for a higher degree of sparsity. Specifically, we propose using the Extended Regularized Information Criterion for choosing $s$, which takes the form

$$\mathrm{ERIC}(s) = -2\ell(\hat{\boldsymbol{\theta}}) - \log(s)\sum_{j=1}^{p}\sum_{k=1}^{d}\mathbb{1}(\hat{\lambda}_{jk} \neq 0), \qquad (2)$$

where $\ell(\hat{\boldsymbol{\theta}})$ is the marginal log-likelihood of the GLLVM evaluated at the OFAL estimates in Definition 1, and $\mathbb{1}(\hat{\lambda}_{jk} \neq 0)$ equals one if the estimate loading $\hat{\lambda}_{jk}$ is not shrunk zero, and zero otherwise. Note the original form of ERIC proposed by Hui et al. (2015b) included an additional parameter in the model complexity term to increase the severity of penalization for high-dimensional regression. However we have chosen to fix this additional parameter (equal to 0.5), but acknowledge that future research should explore the potential inclusion of this parameter. We provide a justification for the use of ERIC in the case of the OFAL penalty in Web Appendix C.

The central feature of ERIC is the *dynamic* model complexity penalty which depends on the tuning parameter itself. This contrasts to the static complexity penalties in the Akaike and Bayesian Information Criteria (AIC and BIC) as well as modifications of these such as the Extended BIC (see Hirose and Yamamoto, 2014, 2015, e.g., of information criteria used previously in penalized sparse factor analysis). For a given dataset, all these criteria penalize a fixed amount for every non-zero parameter in the model. By contrast, the degree of penalization induced by ERIC differs depending on how complex the model is already, as captured by the value of $s$. The penalty $-\log(s)$ becomes more severe the smaller $s$ is, and since small values of $s$ correspond to larger models, this implies ERIC tends to produce more aggressive shrinkage and generally sparser models. As discussed above, such aggressive shrinkage is advantageous given we expect most of the elements $\lambda_{jk}$ to equal zero. Empirically, based on extensive simulations we found that the aggressive shrinkage flavor of ERIC produced substantially better performance in terms of achieving sparsity in GLLVMs compared to other information criteria for selecting $s$.

## 5. Simulation Study

We considered two simulation designs to study the performance of OFAL. The first design is a factor analytic model where the performance of OFAL can be compared against other publicly available penalized likelihood software. The second design involves negative binomial GLLVMs based on our application to the SO-CPR survey. For brevity, we only present results from the first design. Results from the second simulation design (see Web Appendix D) exhibited similar trends, namely that for negative binomial GLLVMs, the OFAL approach performed best overall in terms of estimation and selection of the loadings, latent scores predictions, and estimation of the linear predictor overall, as compared to unpenalized fits and using an information criterion plus varimax rotation approach.

### 5.1. *Setting 1: Factor Analysis*

We adapted the standard factor analytic model simulation design of Hirose and Yamamoto (2015), and compared the performance of OFAL against the penalized likelihood method proposed in that article. The true model is assumed to consist of two factors. We first simulated the latent scores $\boldsymbol{u}_i$ based on three groups, with each group comprising $n/3$ observational units. The scores for the three groups were generated from a bivariate normal distributions with respective means $(-2, 1)$, $(0, 1)$, and $(2, -1)$, and with all the covariances set equal to an identity matrix. In doing so the true latent variables thus display a location pattern, being clustered into three groups. Next, we constructed a $30 \times 2$ loading matrix $\boldsymbol{\Lambda}$ with form

$$\boldsymbol{\Lambda} = \begin{pmatrix} 0.95\mathbf{1}_5 & 0.9\mathbf{1}_5 & 0.8\mathbf{1}_5 & 0\mathbf{1}_{15} \\ 0\mathbf{1}_{15} & 0.8\mathbf{1}_5 & 0.75\mathbf{1}_5 & 0.7\mathbf{1}_5 \end{pmatrix}^\top,$$

where $\mathbf{1}_k$ denotes a $k$-vector of ones. Finally, we set $\boldsymbol{\Psi} = \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$, and then for $i = 1, \ldots, n$ simulated responses as $\boldsymbol{y}_i \sim \mathcal{N}_{30}(\boldsymbol{\lambda}\boldsymbol{u}_i, \boldsymbol{\Psi})$. We considered the number of observational units $n = 50, 100, 200$, and for each value of $n$ simulated 500 datasets.

We compared OFAL, in conjunction with ERIC for choosing the tuning parameter, to: 1) three versions of the penalized likelihood approach proposed in Hirose and Yamamoto (2015), which is implemented in the `R` package `fanc` (Hirose et al., 2016). We considered the MC+ penalty and used AIC, BIC, and EBIC to select the tuning parameter, as these tended to be the combinations which worked best in Hirose and Yamamoto (2015); 2) a popular approach where the number of factors is chosen using information criteria and then a varimax rotation is applied to the estimated loading matrix. For all the methods considered, and with given $p = 30$ responses, we set an upper limit of $d = 8$ factors, remembering that only the first two loadings are truly non-zero. We also conducted additional testing with a higher value of $d$, but found that it made little difference to the simulation results.

We assessed performance in two ways. For estimation performance, we considered the Procrustes error (PE) between the estimated and true $\boldsymbol{\Lambda}$, and between the true and predicted latent scores. For selection, we considered the percentage of datasets where the correct number of factors (two) is chosen, the mean false positive rate or FPR (the proportion of truly zero loadings that are estimated as non-zero), and the mean false negative rate or FNR (the proportion of truly non-zero loadings that are estimated as zero). The Procrustes error was used since it can be regarded as a mean squared error between two matrices after accounting for differences in rotation, sign, and scaling, and is a commonly used measure in ecology to assess ordination accuracy (Legendre and Legendre, 2012).

In terms of point estimation, OFAL almost always had the lowest Procrustes error for both the loadings and prediction of the latent scores (Table 1). Both OFAL and the information criterion plus varimax rotation method performed strongly at selecting the correct number of latent variables, while the penalized likelihood method of Hirose and Yamamoto (2015) using BIC performed poorly at $n = 50$ but much better at the two larger sample sizes. Using AIC either to select the tuning parameter in the penalized likelihood method, or to directly select the number of latent variables, often led to overfitting. This occurred because many of the non-true loadings that were meant to be in the second column of $\boldsymbol{\Lambda}$ were allocated to high-order loadings, resulting in a comparably high false negative rate. Furthermore, the overfitting worsened with increasing sample size. Using the other two information criteria (BIC, EBIC) led to similar strong performance to OFAL in terms of identifying both truly zero and non-zero loadings.

## 6. Application to SO-CPR Survey

We illustrate the application of the OFAL penalty to data collected as part of the Southern Ocean continuous plankton recorder (SO-CPR) survey. As introduced in Section 1, the overall goal of the SO-CPR survey is to better understand the relative influence of environmental, biotic, and human-induced factors driving marine assemblages in the Southern ocean, which in turn has important implications for (say) biodiversity and fisheries management (Hosie et al., 2003). We consider a subset of the survey containing observations of $p = 27$ species collected at $n = 54$ sites in 2016 by the vessel *Umitaka Maru*. Along with the species responses, we also have information on a variety of environmental covariates including water salinity and temperature, fluorescence, and phytoplankton color index. An exploratory analysis based on scatterplot matrices and correlations revealed many of these covariates were strongly correlated.

We first fitted a pure negative binomial GLLVM, outlined in Section 2.1, without any covariates. The goal of fitting this model was to ordinate the sites on a low-dimensional space representing species relative abundance. We set $d = 8$ as the maximum number of latent variables available for order selection, with additional testing using higher values of up to $d = 12$ led to the same number of latent variables being selected. The OFAL penalty selected a sparse, three factor model, with the (first) three columns of the resulting loading matrix $\hat{\boldsymbol{\Lambda}}$ containing 16, 17, and 9 non-zero coefficients, respectively (see Web Appendix E). From the resulting ordination plots, we observe that: 1) latent variables 1 and 2 display a location pattern for sites corresponding to a gradient change in fluorescence and phytoplankton color index. There is also a dispersion effect between sites with different values of color index, 2) latent variable 2 displays a strong

**Table 1**

*Simulation results for setting 1 with a standard factor analytic model. The methods compared include: OFAL, three versions of the penalized likelihood approach in Hirose and Yamamoto (2015), referred as fanc:AIC, fanc:BIC, fanc:EBIC, and choosing the number of factors using information criteria followed by a varimax rotation on the estimated loading matrix (AIC, BIC, EBIC). Performance was assessed in terms of Procrustes errors of loadings (PE-Loadings) and latent scores (PE-scores), percentage of datasets where the correct number of factors is chosen, FPR, and FNR.*

| $n$ | | OFAL | fanc:AIC | fanc:BIC | fanc:EBIC | AIC | BIC | EBIC |
|---|---|---|---|---|---|---|---|---|
| | PE-Loadings* | 0.068 | 0.828 | 0.231 | 0.070 | 0.140 | 0.095 | 0.095 |
| | PE-scores* | 0.208 | 7.028 | 4.070 | 0.151 | 0.326 | 0.143 | 0.143 |
| 50 | % correct LVs | 96.8 | 1 | 16.6 | 100 | 90.0 | 100 | 100 |
| | FPR | 0.035 | 0.069 | 0.046 | 0.059 | 0.027 | 0.016 | 0.016 |
| | FNR | 0 | 0.398 | 0.317 | 0 | 0.034 | 0 | 0 |
| | PE-Loadings* | 0.023 | 0.581 | 0.024 | 0.034 | 0.089 | 0.049 | 0.049 |
| | PE-scores* | 0.127 | 6.142 | 0.198 | 0.150 | 0.379 | 0.144 | 0.141 |
| 100 | % correct LVs | 99.6 | 8.2 | 97.6 | 99.4 | 84.4 | 99.8 | 100 |
| | FPR | 0.027 | 0.059 | 0.008 | 0.063 | 0.024 | 0.008 | 0.008 |
| | FNR | 0 | 0.396 | 0.007 | 0.002 | 0.066 | 0.002 | 0 |
| | PE-Loadings* | 0.012 | 0.389 | 0.017 | 0.020 | 0.108 | 0.025 | 0.025 |
| | PE-scores* | 0.121 | 5.032 | 0.132 | 0.217 | 0.761 | 0.135 | 0.135 |
| 200 | % correct LVs | 100 | 20.4 | 99.8 | 99.8 | 60 | 100 | 100 |
| | FPR | 0.028 | 0.048 | 0.002 | 0.057 | 0.046 | 0.001 | 0.001 |
| | FNR | 0 | 0.344 | 0.001 | 0.001 | 0.211 | 0 | 0 |

*All Procrustes errors are multiplied by 10 for ease of visual presentation.

site pattern reflecting a water temperature gradient, 3) latent variable 3 does not display any pattern associated with the available covariates (Figure 1). As mentioned above though, many of the covariates are strongly collinear. A plot of the resulting correlation matrix constructed from the model, that is, $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^\top$, showed a predominance of positive correlations, suggesting many of the species share a similar response to the latent variables and thus likely the environmental covariates (see Web Appendix E).

Given the above results and the strong correlation between the covariates, we decided to fit a second negative binomial GLLVM with temperature and fluorescence included as linear terms in the model. Both covariates were standardized to have mean zero and unit variance prior to their inclusion in the model. This goal of this model was to infer which species responses were significantly influenced by these two predictors, while accounting for any residual correlation between species due to other missing covariates as well as biotic factors.

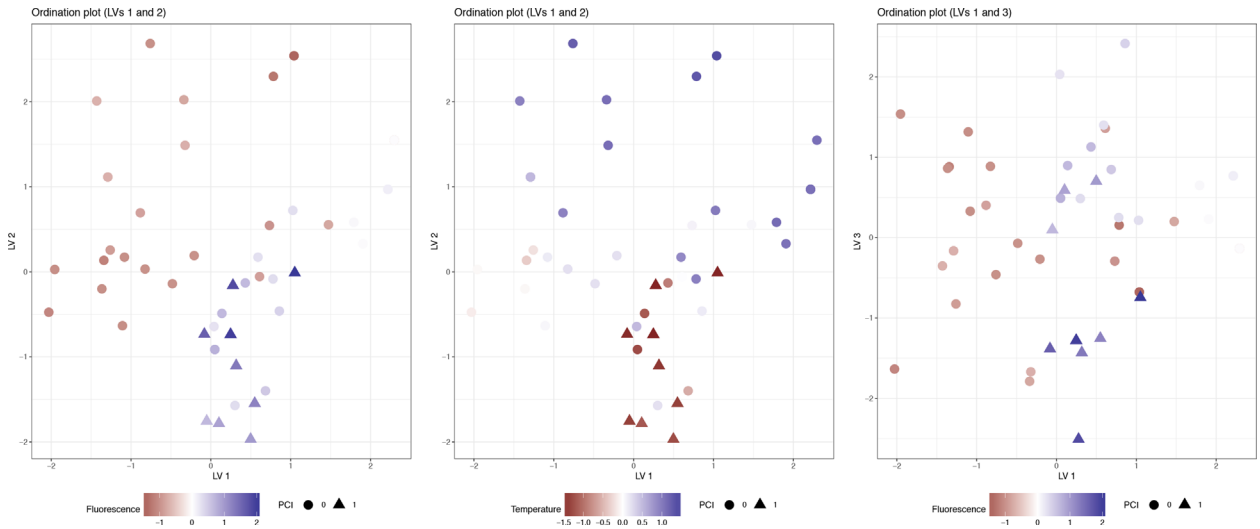


**Figure 1.** Ordination plots from the negative binomial GLLVM fitted to the SO-CPR survey, with no covariates included in the model. We observe that: 1) latent variables 1 and 2 display a location pattern for sites corresponding to a gradient change in fluorescence and phytoplankton color index (PCI). There is also a dispersion difference between sites with different values of color index (left), 2) latent variable 2 also displays a strong site pattern reflecting a water temperature gradient (center), 3) latent variable 3 does not display any pattern which is associated with any of the available covariates (right). This figure appears in color in the electronic version of this article.

With temperature and fluorescence included, OFAL selected a two factor model (see Web Appendix E for the estimated loading matrix). The resulting ordination plots of the two latent variables (not shown) no longer exhibited any site patterns corresponding to fluorescence, temperature, and color index. To construct standard errors for the fitted species-specific regression coefficients $\hat{\boldsymbol{\beta}}_j$, we used numerical differentiation to first calculate the observed Fisher information matrix based on the marginal likelihood of the fitted model, $\boldsymbol{I}(\hat{\boldsymbol{\theta}}) = \nabla^2 \ell(\hat{\boldsymbol{\theta}})$, where the differentiation was done only with respect to the parameters not shrunk to zero, that is, the $\hat{\boldsymbol{\beta}}_j$'s, the non-zero elements of $\hat{\boldsymbol{\Lambda}}$, and $\hat{\boldsymbol{\psi}}$. Standard errors were then constructed based on the diagonal elements of $\boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}$. We acknowledge this approach is ad hoc in that the construction of standard errors does not take into account the variable selection process, but emphasize that the field of post-model selection inference remains an active area of research (e.g., Lee et al., 2016). The resulting caterpillar plots showed six and seven species, respectively, had coefficients to temperature and fluorescence that were significantly less than zero, implying a preference for cooler temperature waters and waters with low levels of chlorophyll, respectively. Four (*Oithona.s*, *Limacina*, *Ctenocalanus*, and *Calanoida*) had significant negative responses to both predictors (Figure 2), thus providing some idea of the specific environmental niche occupied by these species. Given small copepods such as *Oithona.s* and *Ctenocalanus* often dominate the grazing impact of many zooplankton assemblages, a better understanding of the niche of such species is useful for assessing their long term environmental impact on the wider marine community. When compared to the model without covariates, the residual correlation matrix from the GLLVM with the two covariates included exhibited considerably less correlations, reflecting the fact that a considerable proportion of the co-occurrences can be explained by shared environmental responses to temperature and fluorescence (see Web Appendix E). Indeed, comparing the trace terms, trace($\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^\top$), between the two models, we find that 52% of the covariation between species can be explained by temperature and fluorescence alone, suggesting that water temperature and chlorophyll levels together are key environmental filters for zooplankton assemblages.

## 7. Discussion

There are many avenues of future research which can be undertaken in relation to the OFAL penalty. For instance, the motivating dataset considered in this article was not particularly high dimensional, although other applications of GLLVMs may involve this (e.g., Bai and Li, 2012). Therefore, it will be important to study the performance of OFAL for high-dimensional GLLVMs with non-normal responses, where both $n$ and $p$ may be in their hundreds or thousands. In particular, the OFAL penalty and ERIC (or any other information criterion used) may need to be modified to penalize even more heavily on the higher-order loadings to reflect this greater degree of sparsity. Furthermore, the construction of adaptive weights for high-dimensional GLLVMs also presents a challenge, as the initial, unpenalized model used to build these weights may be unstable due to overfitting; see also the discussion in Section 3 on the choice of $d$. Future research can thus look into alternative methods for constructing $w_{1l}$ and $w_{2jk}$, or even modifying the OFAL penalty such that weights are not required (see the review of Huang et al., 2012, for example).

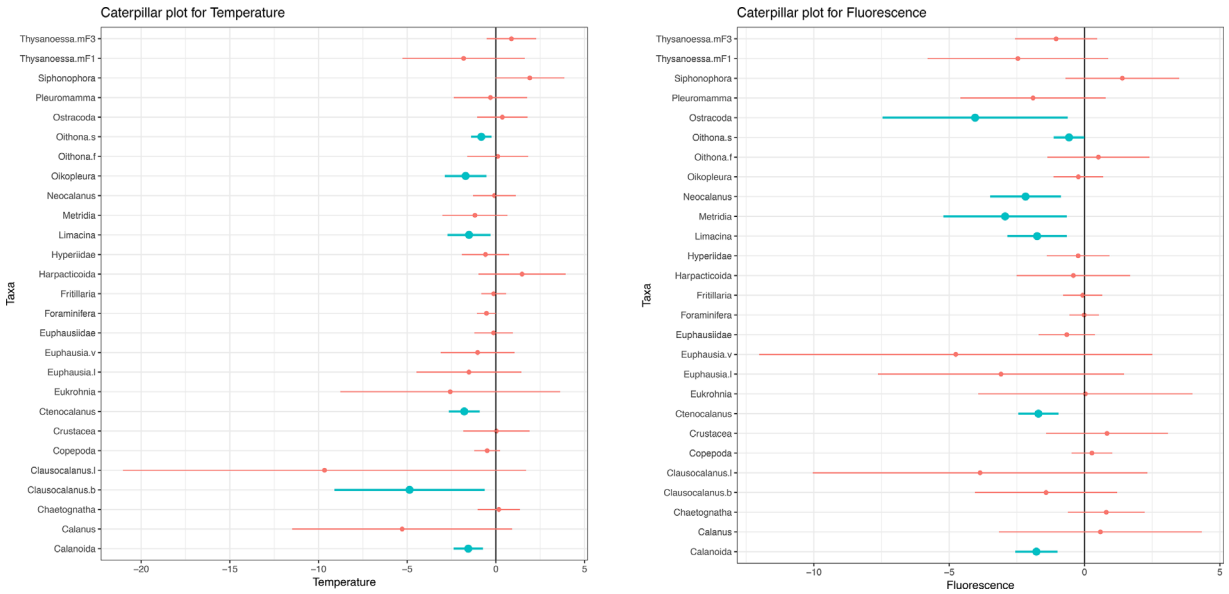


**Figure 2.** Caterpillar plots (point estimates and 95% confidence intervals) of the species-specific coefficients $\hat{\boldsymbol{\beta}}_j$ for temperature (left) and fluorescence (right) from the negative binomial GLLVM fitted to the SO-CPR survey, with these two covariates included. The bolded and thicker intervals are those who confidence intervals do not contain zero. Six and seven species, respectively, had coefficients to temperature and fluorescence significantly less than zero, with *Oithona.s*, *Limacina*, *Ctenocalanus*, and *Calanoida* having strong negative responses to both predictors. This figure appears in color in the electronic version of this article.

Additionally, the computational methods proposed in this article may not be feasible in such settings, and instead we may need to combine OFAL with faster, approximate likelihood-based estimation methods (Hui et al., 2017; Niku et al., 2017). Such research into more computationally efficient approaches will also be beneficial if we wanted to include more tuning parameters in the OFAL penalty, for example, the mixing parameter $\alpha$ discussed in Section 3. Yet another major avenue of research is modifying the OFAL penalty for specific types of GLLVM applications, for example, microarray studies when we want to shrink entire rows as well as columns of the loading matrix to zero to select only "informative" genes on the latent space (e.g., Hirose and Konishi, 2012), and multi-response data where the objective is to simultaneously achieve sparsity in both the regression coefficients and loading matrix, in which case OFAL can be combined with penalties on the $\boldsymbol{\beta}_j$'s such as the adaptive LASSO and group LASSO penalties (Yuan and Lin, 2006).

## 8. Supplementary Material

Web Appendices referenced in Sections 1 and 4–6 are available with this article at the *Biometrics* website on Wiley Online Library. The `R` code for fitting GLLVMs with the OFAL penalty is also provided as part of the Web Appendices. The SO-CPR survey data are freely available upon request from https://data.aad.gov.au/aadc/cpr/.

## Acknowledgements

Thanks to Nicole Hill and John Kitchener for useful discussions related to the SO-CPR survey data.

## References

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics* **40**, 436–465.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.

Choi, J., Oehlert, G., and Zou, H. (2010). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and Its Interface* **3**, 429–436.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The Elements of Statistical Learning*, volume 2. New York: Springer-Verlag.

Hirose, K. and Konishi, S. (2012). Variable selection via the weighted group lasso for factor analysis models. *Canadian Journal of Statistics* **40**, 345–361.

Hirose, K. and Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis* **79**, 120–132.

Hirose, K. and Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing* **25**, 863–875.

Hirose, K., Yamamoto, M., and Nagata, H. (2016). *fanc: Penalized Likelihood Factor Analysis via Nonconvex Penalty*. R package version 2.2.

Hosie, G., Fukuchi, M., and Kawaguchi, S. (2003). Development of the Southern Ocean continuous plankton recorder survey. *Progress in Oceanography* **58**, 263–283.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science* **27**, 481–499.

Hui, F. K. C., Warton, D. I., and Foster, S. D. (2015a). Multi-species distribution modeling using penalized mixture of regressions. *The Annals of Applied Statistics* **9**, 866–882.

Hui, F. K. C., Warton, D. I., and Foster, S. D. (2015b). Tuning parameter selection for the adaptive lasso using ERIC. *Journal of the American Statistical Association* **110**, 262–269.

Hui, F. K. C., Warton, D. I., Ormerod, J. T., Haapaniemi, V., and Taskinen, S. (2017). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics* **26**, 35–43.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187–200.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* **44**, 907–927.

Legendre, P. and Legendre, L. (2012). *Numerical Ecology, Third Edition*, volume 20. Oxford: Elsevier.

Niku, J., Warton, D. I., Hui, F. K. C., and Taskinen, S. (2017). Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology. *Journal of Agricultural, Biological and Environmental Statistics* **22**, 498–522.

Rubin, D. B. and Thayer, D. (1982). EM Algorithms For ML Factor Analysis. *Psychometrika* **47**, 69–76.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Florida: Chapman & Hall/CRC Press.

Smith, A. B., Ganesalingam, A., Kuchel, H., and Cullis, B. R. (2015). Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theoretical and Applied Genetics* **128**, 55–72.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)* **58**, 267–288.

Warton, D. I., Blanchet, F. G., OHara, R., Ovaskainen, O., Taskinen, S., Walker, S. C., et al. (2016). Extending joint models in community ecology: A response to Beissinger et al. *Trends in Ecology & Evolution* **31**, 737–738.

Warton, D. I., Blanchet, F. G., OHara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., et al. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution* **30**, 766–779.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **68**, 49–67.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **67**, 301–320.

*Received December* 2017. *Revised February* 2018. *Accepted March* 2018.