Régression linéaire généralisée sur composantes supervisées pour la modélisation jointe des réponses

Julien Gibaud

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biostatistique

École doctorale Information, Structures et Systèmes (I2S)

Unité mixte de recherche 5149 - Institut Montpelliérain Alexander Grothendieck (IMAG)

Régression linéaire généralisée sur composantes supervisées pour la modélisation jointe des réponses

Présentée par Julien GIBAUD
Le 9 décembre 2022

Sous la direction de Catherine TROTTIER et Xavier BRY

Devant le jury composé de

Jean Noël BACRO	Professeur	Université de Montpellier	Président du jury
Xavier BRY	Maître de conférences	Université de Montpellier	Co-encadrant de thèse
Marie CHAVENT	Professeure	Université de Bordeaux	Examinatrice
Fabien LAROCHE	Ingénieur de recherche	INRAE de Toulouse	Examinateur
Jérôme SARACCO	Professeur	Institut Polytechnique de Bordeaux	Rapporteur
Catherine TROTTIER	Maîtresse de conférences	Université Paul Valéry - Montpellier 3	Directrice de thèse
David I. WARTON	Professeur	University of New South Wales - Sydney	Rapporteur

UNIVERSITÉ DE MONTPELLIER

REMERCIEMENTS

Comme il est coutume de le faire, mes premiers remerciements iront à mes encadrant·es de thèse. Catherine, Xavier, je vous remercie pour ces trois années. J'imagine que choisir un étudiant en thèse n'est pas chose évidente, merci pour cela. Vos connaissances m'ont permis de découvrir la beauté de l'analyse des données tandis que votre pédagogie me permettait d'en appréhender les subtilités. Cependant, un doctorat ce n'est pas seulement que les pratiques des sciences, c'est aussi, malheureusement trop souvent mis de coté, un ensemble d'interactions humaines qui permettent une bonne entente. Toutes ces réunions, qu'elles soient au bureau ou en visio, ont permis cela. Merci Catherine pour le thé. Merci Xavier pour le chocolat.

Je remercie Jean-Noël Bacro, Marie Chavent et Fabien Laroche pour avoir accepté de participer à ma soutenance de thèse. Plus particulièrement, merci Jérôme Saracco et David Iain Warton pour avoir rapporté ce manuscrit[1].

Une Université n'est pas seulement constituée que d'enseignant·es chercheurs·ses, c'est aussi un grand nombre de personnels des bibliothèques, ingénieurs, administratifs, techniques, sociaux et de santé (BIATSS) ainsi que des personnes présentes pour entretenir la propreté des locaux. On a tendance à penser que ces personnes ne participent pas à la production des connaissances. Ceci est faux, car sans ces travailleurs et travailleuses le monde de l'Enseignement Supérieur et de la Recherche ne pourrait simplement pas tourner. Baptiste, Brigitte, Carmela, Catherine, Céline, François-David, Ghislain, Jean-Baptiste, Laurence, Nathalie, Samira, Stéphanie et Sophie, je vous remercie pour cela.

Pour des raisons de légalité, ce paragraphe est écrit en vérité inversée. Je tient à ne surtout pas remercier Alexandra Elbakyan pour la création du site web Sci-Hub qui permet de "supprimer toutes les barrières pour un accès aux sciences"[2]. Je n'incite en aucune

[1] I really thank Jérôme Saracco and David Iain Warton for being the referees of this manuscript.
[2] "To remove all barriers in the way of science"

façon les chercheurs et chercheuses du monde entier à s'emparer de cet outil contournant les paywalls de l'édition scientifique. Je ne préciserais pas non plus que Sci-Hub est en collaboration avec le moteur de recherche Library Genesis (LibGen) permettant aussi la récupération de livres scientifiques.

Ces trois années passées au sein de l'Institut Montpelliérain Alexander Grothendieck (IMAG) seront principalement marquées par le plaisir d'avoir partagé mes journées avec les (post-)doctorant·es et ATER du laboratoire. C'est, au moins en partie, l'ambiance régnant au premier et deuxième étages du bâtiment 9 qui m'a permis d'aller au bout de cette aventure. Merci Aurélio pour ta gentillesse et tes pâtes à la carbonara (sans crème !). Merci Morgane et Bart pour votre passion des randonnées, de la "vraie" musique et du tofu. Merci Nathan pour tes discussions sur la didactique et l'épistémologie que je trouve passionnantes. Merci Thiziri pour m'avoir accueilli si gentiment dans le bureau et supporté pendant tout ce temps. Merci Tom pour ta passion de One Piece et ta bonne humeur contagieuse. Merci Victor pour la découverte de l'escrime médiévale. Merci Hermès, Pablo et Tanguy pour avoir accepté d'être désignés volontaires pour la cogestion du séminaire. Merci aussi Alain, Alan, Amandine, Amélie, Antonia, Cassio, Corentin, Elena, Emmanuel, Florent, Guillaume, Inés, Ivan, Juliette, Marien, Meriem, Mireille, Ozy, Pascal, Radia, Raphaël, Salomé, Steven, Thibault, Tiffany et Zaineb pour avoir été présent·es.

L'amitié ne se limitant pas à l'enceinte de l'Université, je remercie toutes les personnes que j'ai apprécié dans les différentes villes que j'ai visité. Merci Coline, Marie, Romane et Vincent pour les vacances partagées et pour continuer à garder le contact après autant d'années. Même si nous avons été dispatché·es aux quatre coins de la France, c'est toujours un immense plaisir de vous retrouver. Merci Brieuc pour m'avoir accompagné pendant trois ans de licence, pour la rédaction des devoirs maisons ainsi que pour la piscine et les barbecues. Je le dis derechef, Isabelle et toi avez toutes mes félicitations pour la venue au monde d'Anouk ! Maëllie, merci pour ces huit années passées ensemble. Même si désormais nos chemins se séparent, je ne regrette rien.

Même si je ne suis pas la personne la plus attachée à la valeur "famille", je n'hésiterais pas à reconnaitre que vous avez toujours été là pour moi. Bastian, Maman, Papa, merci pour votre amour et votre présence. Du haut de mon grand âge (27 ans tout de même), j'affirme la chance que j'ai de vous avoir à mes côtés. Évidement, les remerciements des membres de la famille ne sauraient être complets sans citer Guillaume, Léa, Maminou, Marie, Papi Bernard, Papi Jean-Paul, Sandrine et toutes celles et ceux qui se sont éloigné·es ou ne sont plus.

RÉSUMÉ DU MANUSCRIT

Comme explicité par le rapport du Groupe d'experts Intergouvernemental sur l'Évolution du Climat (GIEC, 2022), le changement climatique produit de nombreux déséquilibres de l'écosystème qui entrainent une large extinction d'espèces animales et végétales. Dans ce contexte, le développement de modèles qui permettent de prédire le futur de la biodiversité est devenu un problème majeur. Plusieurs avancées ont été faites en ce sens, en particulier en étendant les modèles de distribution d'espèces (Species Distribution Models, SDM, Guisan and Thuiller, 2005), qui traitent les espèces séparément, aux modèles joints de distribution d'espèces (Joint Species Distribution Models, JSDM, Pollock et al., 2014). Les JSDM permettent de formaliser l'interdépendance entre les espèces et de comprendre son impact sur la composition des communautés. De plus, modéliser les variables réponses (dans notre cas, l'abondance d'espèces) peut nécessiter de prendre en compte un grand nombre de covariables explicatives possiblement fortement corrélées, ce qui est le cas pour les variables climatiques. Ainsi, SDM et JSDM demandent à être régularisés. La régularisation peut être effectuée par exemple à travers une réduction de dimension fondée sur des modèles à composantes. Ceci consiste à supposer qu'il existe un petit nombre de dimensions latentes explicatives que nous cherchons à capturer au travers de combinaisons linéaires de variables explicatives que nous appelons "composantes". Dans cette thèse, nous voulons construire des composantes qui peuvent être interprétées comme de nouvelles et pertinentes variables synthétiques.

La régression linéaire généralisée sur composantes supervisées (Supervised Component-based Generalized Linear Regression, SCGLR, Bry et al., 2013) et son extension thématique, THEME-SCGLR (Bry et al., 2020b), sont des approches couplant l'estimation des modèles linéaires généralisés (Generalized Linear Model, GLM, McCullagh and Nelder, 1989) multivariés avec la recherche de composantes explicatives. Cependant, SCGLR a deux limitations majeures. Premièrement, cette méthode suppose que toutes les réponses sont expliquées par les mêmes dimensions latentes. Dans de nombreux contextes, cela peut ne pas être le cas : les réponses peuvent être très différentes et sont donc susceptibles d'être modélisées à partir de dimensions explicatives qui sont,

dans une certaine mesure, spécifiques. Comme deuxième limite, SCGLR suppose que toutes les réponses sont indépendantes conditionnellement aux variables explicatives. Néanmoins, dans un contexte d'analyse multivariée, les dépendances mutuelles entre les réponses doivent, en toute rigueur, être prises en compte. L'objectif principal de cette thèse est de surmonter ces limitations des versions précédentes de SCGLR.

Nous proposons d'abord d'étendre SCGLR de manière à trouver des groupes de variables réponses modélisés par des dimensions explicatives spécifiques. Les méthodes de classification ou les techniques classiquement utilisées dans la littérature d'écologie statistique pour identifier des groupes ne considèrent pas les données d'occurrence ou d'abondance comme des réponses à des variables explicatives (Dufrêne and Legendre, 1997; De Cáceres et al., 2010). Dans l'objectif de prendre en compte la modélisation des variables réponses dans la classification, nous proposons de combiner la méthode SCGLR avec un modèle de mélange fini (Finite Mixture Model, FMM, McLachlan and Peel, 2004) sur les réponses, chaque classe étant caractérisée par des composantes explicatives propres. Dans un second travail, comme pour THEME-SCGLR, la matrice réponse est modélisée par un partitionnement thématique des variables explicatives, nommés "thèmes". Ainsi, la régularisation est effectuée afin de chercher, dans chacun des thèmes, un nombre approprié de composantes qui contribuent à la fois à la prédiction de la matrice réponse et à la capture d'informations pertinentes dans chacun des thèmes. De plus, nous relâchons l'hypothèse d'indépendance conditionnelle. Dans un contexte écologique par exemple, les co-occurrences d'espèces qui ne sont pas expliquées par les variables environnementales demandent à être modélisées. Avec cet objectif en tête, nous modélisons la matrice de variance-covariance conditionnelle des réponses au moyen d'un ensemble de variables aléatoires latentes appelées "facteurs".

Maintenant que les principaux objectifs de la thèse ont été exposés, le manuscrit propose dans un deuxième chapitre un état de l'art qui ne se veut pas exhaustif mais plutôt introductif aux différentes méthodologies dont nous avons besoin pour développer nos approches.

Les modèles linéaires généralisés (GLM, Nelder and Wedderburn, 1972) sont introduits dans un contexte où la distribution gaussienne est inappropriée au type de variables étudiées, comme les données qualitatives ou discrètes. Les GLM couvrent l'ensemble de ces situations en permettant aux observations d'être issues de variables aléatoires ayant une distribution appartenant à la famille exponentielle (Binomiale, Gamma, Normale, Poisson...). Contrairement au modèle linéaire classique, l'espérance de la variable aléatoire n'est pas définie directement par une combinaison linéaire des variables explicatives, mais au moyen d'une fonction la reliant à ces dernières. Pour plus de détails, McCullagh and Nelder (1989) proposent un aperçu complet de ce sujet et Fahrmeir (1994) étend les GLM au cas de l'analyse multivariée.

Élaboré à partir des moindres carrés partiels repondérés itérativement (Iteratively Reweighted Partial Least Squares, IRPLS, Marx, 1996), SCGLR permet de construire des composantes dans un contexte de GLM multivariés. Contrairement aux méthodes comme les moindres carrés partiels (Partial Least Squares, PLS, Wold et al., 1984) ou au modèle linéaire généralisé à vecteur de rang réduit (Reduced Rank Vector Generalized Linear Model, RRVGLM, Yee and Hastie, 2003), SCGLR optimise un critère de compromis entre la qualité d'ajustement (Goodness-of-Fit, GoF) du modèle et la pertinence structurelle (Structural Relevance, SR, Bry and Verron, 2015) des directions dans les variables explicatives. Cette méthodologie permet à la fois de trouver des directions explicatives fortes et interprétables, et de produire des prédicteurs régularisés dans le cadre de grande dimension. Par la suite, SCGLR est raffinée de manière à modéliser différents types de données (Chauvet et al., 2019; Bry et al., 2020a,b).

Les modèles de mélanges finis (FMM) se caractérisent par un cadre paramétrique dans lequel l'objectif est de modéliser une distribution de probabilité inconnue par une somme finie de distributions. Les FMM fournissent une approche mathématique à la modélisation statistique d'une grande variété de phénomènes aléatoires. En raison de leur utilité et d'une méthode de modélisation extrêmement flexible, les FMM continuent de recevoir une attention croissante, tant d'un point de vue pratique que théorique (voir McLachlan and Peel (2004) pour un livre de référence). L'étendue et le potentiel d'application des FMM se sont considérablement élargis. En effet, de nombreux champs de recherche dans lesquels les FMM sont impliqués peuvent être cités : astronomie (Lee et al., 2012), écologie (Pledger and Phillpot, 2008) ou les approches quantitatives de la psychologie (Colder et al., 2002) et de la sociologie (Jones et al., 2001).

Les modèles à facteurs sont introduits par Spearman (1904); Thomson (1916); Thurstone (1931) pour des données issues de la psychologie. Depuis, cette méthode a été largement diversifiée comme le montre les travaux de Bartholomew (1995); Saidane (2006) et Tami (2016). Les facteurs sont des variables aléatoires latentes non-corrélées résumant un ensemble de variables observées corrélées. Les variables observées sont décrites par une combinaison linéaire des facteurs, un paramètre de moyenne et une erreur de mesure. Les facteurs n'étant pas observés directement, ils doivent être prédits en même temps que l'estimation des paramètres du modèle. Les méthodes fréquentistes proposées pour estimer les paramètres sont soit basées sur le maximum de vraisemblance d'un échantillon de matrice de variance-covariance (Jöreskog, 1967, 1969) soit sur l'algorithme Espérance-Maximisation (EM, Dempster et al., 1977; Rubin and Thayer, 1982; Jamshidian, 1997). Afin d'identifier le modèle, toutes les méthodes existantes doivent imposer des contraintes sur les paramètres.

Ensuite, dans un troisième chapitre, le manuscrit présente le cadre et les objectifs de modélisation pour lesquels nous proposons de combiner SCGLR avec un FMM. Nous en profitons pour décrire l'algorithme global qui combine l'algorithme EM pour estimer les

paramètres du modèle de mélange et l'algorithme du gradient normé projeté itéré (Projected Iterated Normed Gradient, PING) qui permet de trouver les composantes. Dans ce travail, nous utilisons une modélisation inspirée de Dunstan et al. (2011, 2013) qui suppose que les espèces peuvent être classifiées dans un petit nombre de groupes en fonction de leur réponse à un gradient environnemental. Dans le modèle proposé par Dunstan et al. (2011, 2013), les réponses à l'intérieur d'un groupe partagent les mêmes paramètres de régression mais possèdent un paramètre de moyenne spécifique à chaque espèce. Contrairement à cela, nous proposons d'autoriser les réponses à posséder leurs propres paramètres de régression. Nous définissons un groupe comme un ensemble de réponses dépendant de dimensions explicatives communes. Pour y parvenir, le critère compromis de SCGLR est amélioré de manière à empêcher les groupes de réponses de dépendre de sous-espaces explicatifs trop proches.

Deux schémas de simulation sont mis en œuvre pour évaluer les performances de la méthode implémentée. Le premier porte sur l'identification de groupes de réponses dans un cas de fortes corrélations entre variables latentes engendrant les espaces explicatifs. Dans cette simulation, nous cherchons quelle combinaison d'hyper-paramètres permet de retrouver les vrais groupes de réponses. Nous utilisons l'index de Rand (Rand Index, RI, Rand, 1971) et l'index de Rand ajusté (Adjusted Rand Index, ARI, Hubert and Arabie, 1985) pour évaluer la qualité de la classification. Les hyper-paramètres étant très nombreux, nous choisissons une heuristique permettant d'approcher une bonne combinaison grâce au critère d'information Bayésien (Bayesian Information Criterion, BIC, Schwarz, 1978). Afin de comparer les performances de notre approche, nous calculons les temps de calcul ainsi que les RI et ARI des partitions trouvées par d'autres méthodes de la littérature. Dans la deuxième expérience, nous étudions la détection du vrai nombre de composantes dans un contexte de faible corrélation entre les variables latentes. Un autre objectif de ces simulations est de traiter conjointement plusieurs types de réponses modélisées par de nombreuses variables explicatives présentant à la fois des ensembles de variables hautement redondantes et des variables isolées. Une telle structure de données est souvent rencontrée en pratique lorsqu'aucune présélection de variables explicatives n'a été effectuée, et entraîne des difficultés de modélisation et d'estimation.

Nous appliquons ensuite notre approche à un jeu de données constitué d'abondances d'espèces végétales se trouvant dans les forêts humides du bassin du Congo. Pour prédire ces abondances, des variables climatiques ainsi que quelques variables caractérisant la localisation ont été mesurées. L'application de notre méthode permet de détecter trois groupes d'espèces. Le premier contient les abondances qui seraient liées à un gradient de température et qui seraient sensibles à un déficit en eau. Le deuxième contient les espèces liées à un gradient régional opposant des zones à saison sèche fraîche et pauvre en lumière (les côtes du Gabon) et des zones à fort taux d'évapotranspiration (limite nord des forêts d'Afrique centrale). Enfin, il apparait que les espèces qui ne sont connectées à aucun gradient en particulier mais seulement à une combinaison quelconque de variables clima-

tiques composent le troisième groupe. En effet, comme détaillé par Réjou-Méchain et al. (2021), dans les données récoltées, une part importante de la liaison entre les abondances d'espèces et les variables climatiques est due au hasard. De plus, nous montrons, même si l'amélioration est modeste, que les groupes trouvés permettent une meilleure prédiction des abondances. Cependant, il faut préciser que la prédiction des abondances à partir des seules variables climatiques est connue pour être mauvaise (Beale et al., 2008).

Le quatrième chapitre est constitué d'une deuxième approche que nous proposons où l'extension thématique de SCGLR et les modèles à facteurs sont combinés. Nous présentons aussi une nouvelle manière de fusionner les algorithmes PING et EM. En effet, bien qu'il existe des travaux proposant l'estimation des GLM dans un contexte de modèle à facteurs (Generalized Linear Latent Variable Model, GLLVM, Skrondal and Rabe-Hesketh, 2004), aucun consensus ne permet de définir la bonne manière d'effectuer cette estimation. Ainsi, contrairement aux méthodes fréquentistes se basant sur une approximation de la log-vraisemblance (Hui et al., 2017; Niku et al., 2017; Korhonen et al., 2023), nous présentons une méthode inspirée de Saidane et al. (2013) estimant les paramètres après la linéarisation du modèle. Comme dans le travail précédent, nous cherchons des groupes de réponses. En effet, des réponses partageant de fortes corrélations conditionnelles positives ou négatives demanderaient à être groupées. Ainsi, grâce au positionnement multidimensionnel (Multidimensional Scaling, MDS, Cox and Cox, 2008) et à une distance fondée sur la matrice de corrélation conditionnelle issue de la matrice de variance-covariance conditionnelle modélisée par le modèle à facteurs, nous proposons d'identifier ces groupes.

Différentes expériences numériques sont réalisées pour tester les performances de cette approche. La première consiste encore une fois à s'assurer que la bonne combinaison de composantes et de facteurs est retrouvée à l'aide du BIC. La deuxième a pour objectif de déterminer quelle combinaison d'hyper-paramètres permet d'identifier la meilleure partition, au travers du RI et de l'ARI, dans un cas où les groupes sont simulés de manière plus ou moins séparés. De la même manière que dans les simulations précédentes, ces expériences permettent de modéliser une matrice réponse incluant plusieurs distributions. La nouveauté réside dans le partitionnement des variables explicatives (encore nombreuses et redondantes) en deux thèmes distincts. Dans un contexte de variables réponses binaires, nous comparons ensuite le temps de calcul, le RI et l'ARI de notre méthode avec la librairie **gllvm** (Niku et al., 2019b) sur des jeux de données simulées de différentes tailles. Cependant, dans cette troisième expérience, dans un objectif de comparaison, nous nous limitons à un petit nombre de variables explicatives accompagnées d'une covariable additionnelle catégorielle.

Cette nouvelle approche a ensuite été testée sur un jeu de données constitué de mesures (abondance, richesse, diversité) réalisées sur des communautés de carabidés et de plantes vasculaires dans des champs de céréales des Vallées et Coteaux de Gascogne. Pour prédire cette biodiversité agricole, des variables explicatives réparties en quatre thèmes

ainsi qu'une variable binaire représentant l'année d'observation (2016 ou 2017) ont été récoltées. Le potentiel de prédation, l'intensité fermière et l'hétérogénéité paysagère liée aux couverts semi-naturels et à la mosaïque des cultures sont les thèmes incorporés dans la modélisation. Après utilisation de la méthode, il apparait que seul le thème d'intensité fermière est pertinent dans la prédiction de l'agrobiodiversité. Plus particulièrement, se sont les variables explicatives représentant le traitement par herbicides, le nombre d'opérations effectuées par les fermiers, la profondeur du labourage et la quantité d'azote qui sont le plus impliquées dans cette prédiction. Les groupes de réponses identifiés grâce à la matrice de corrélation conditionnelle sont au nombre de quatre. Le premier regroupe des mesures faites sur les carabidés possédant de très fortes corrélations conditionnelles tandis que les autres groupes sont composés d'un mélange entre plantes et carabidés.

La présentation et le développement de nouvelles méthodes étant faites, nous décrivons dans un dernier chapitre les travaux en cours ainsi que les perspectives pour l'avenir des composantes supervisées. Nous expliquons d'abord pourquoi nous pensons que les composantes construites par SCGLR seraient plus intéressantes que des composantes classiques (ACP, PLS ...) sur des données issues de l'écologie. Ensuite, nous proposons de nouvelles approches permettant de combiner les versions précédentes de SCGLR à celles détaillées dans ce manuscrit. En effet, Chauvet et al. (2019) ayant étendu SCGLR au modèle mixte, nous supputons que l'incorporation de ces travaux dans des contextes de mélanges sur les réponses ou de modèles à facteurs latents permettrait une meilleure exploitation de l'information contenue dans les données. Finalement, nous soumettons un moyen de construire des composantes pour lesquelles les vecteurs de coefficients associés possèdent des zéros pour les variables explicatives non pertinentes dans la prédiction de la matrice réponse. Dans l'esprit de Simon et al. (2013), ce moyen pourrait aussi être envisagé pour l'identification de thèmes pertinents.

Enfin, nous détaillons en annexe le matériel supplémentaire dont nous avons eu besoin dans ce manuscrit. Nous présentons les considérations théoriques de l'algorithme PING ainsi que les expressions analytiques des sous-critères utilisés pour définir SCGLR.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

CHAPTER 1

INTRODUCTION

Contents

This chapter exposes the context and the main objectives of this work.

1.1 Context of the work

As highlighted by the report of the Intergovernmental Panel on Climate Change (IPCC, 2022), the climate change produces many ecosystem imbalances which might involve large extinctions of animal or plant taxa. In this context, the development of models which allow to predict the future of the biodiversity has become a crucial issue. A number of advances have been made, in particular by extending Species Distribution Models (SDMs, Guisan and Thuiller, 2005), which treat the taxa separately, to Joint Species Distribution Models (JSDMs, Pollock et al., 2014). Both SDMs and JSDMs are based on a Generalized Linear Models (GLM, McCullagh and Nelder, 1989) structure. JSDMs allow to formalize the interdependence between taxa, and to understand its impact on the composition of communities. Besides, modeling responses (here, the abundances of taxa) requires taking into account a large set of possibly highly correlated explanatory covariates, which is the case of climatic variables, so SDMs as JSDMs demand regularization. This can be carried out by means of component-based dimension reduction. It consists in assuming that there is a small number of common latent explanatory dimensions, which we aim to capture through as many linear combinations of the explanatory variables, named components. Moreover, the case where the explanatory variables outnumber the observations (referred to as "high dimensional") is likely to become a new standard (Warton et al., 2015). In this thesis, we aim to build components which can be interpreted as new and relevant synthetic climatic variables.

The Supervised Component-based Generalized Linear Regression (SCGLR), introduced by Bry et al. (2013), bridges the multivariate GLM estimation, with the component-based dimension reduction of the explanatory space. More formally, a response matrix $\boldsymbol{Y}$ is assumed to depend on a set $\boldsymbol{X}$ of explanatory variables, and a set $\boldsymbol{A}$ of additional covariates. Explanatory variables are supposed many and redundant, thus demanding dimension reduction and regularization. By contrast, additional covariates contain few selected variables which are forced into the regression model, no regularization being carried out with respect to them. Originally, SCGLR was designed to extract from the explanatory variables a sequence of components $\boldsymbol{f^h} = \boldsymbol{X}\boldsymbol{u^h}$, where $\boldsymbol{u^h}$ is a loading vector. Denoting $\boldsymbol{\eta}_k$ the kth linear predictor associated with response $\boldsymbol{y}_k$, it is then assumed that

$$\boldsymbol{\eta}_k = \sum_h \left(\boldsymbol{X}\boldsymbol{u^h}\right)\gamma_k^h + \boldsymbol{A}\boldsymbol{\delta}_k,$$

where γ_k^h and $\boldsymbol{\delta}_k$ are regression parameters. However, SCGLR still has two major limitations. First, SCGLR assumes that all the responses are explained by the same latent dimension. In several contexts, this might well not be the case: the responses could be very different, and are thus likely to be modeled from explanatory dimensions which are, to some extent, specific. As a second limitation, SCGLR assumes that all the responses are independent conditional on the explanatory variables. Nevertheless, in a framework of multivariate analysis, the mutual dependencies between the responses need to be taken

into account. The main objective of this thesis is to overcome these limitations of former versions of SCGLR.

We first propose to extend SCGLR so as to find groups of response variables being modeled by the same specific explanatory dimensions. The clustering models or techniques classically used in statistical literature to identify groups do not consider the presence or abundance data as responses to explanatory variables (Dufrêne and Legendre, 1997; De Cáceres et al., 2010). In order to take the modeling of outcomes into account within the clustering, we propose to combine the SCGLR model with a Finite Mixture Model (FMM) of responses (see McLachlan and Peel (2004) for a reference book), leading to a method of response mixture SCGLR we name rmSCGLR. In our work, we use a modeling approach based on Dunstan et al. (2011, 2013), which assumes that all outcomes can be clustered into a small number of groups with respect to their responses to environmental gradients. In their model, the outcomes within a group share the same regression parameters with an intercept specific to each outcome. By contrast, we propose to entitle responses to their own regression parameters, and to define the gth group as a set of responses depending on the same common explanatory components

$$\boldsymbol{\eta}_{kg} = \sum_h \left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{g}}^{\boldsymbol{h}}\right)\gamma_{kg}^h + \boldsymbol{A}\boldsymbol{\delta}_{kg},$$

where $\boldsymbol{u}_{\boldsymbol{g}}^{\boldsymbol{h}}$ is the loading vector of the hth specific component of group g. To estimate the model parameters, we propose a criterion extending that of SCGLR, and develop an algorithm combining component-based model and Expectation Maximization (EM, Dempster et al., 1977) estimation.

Then, in a second work, we propose to relax the conditional independence assumption. In an ecological context for instance, the species co-occurrences that are not explained by the environmental variables demand to be modeled. With this aim in mind, we model the conditional variance-covariance matrix of the responses by means of a set $\boldsymbol{G}$ of random latent variables called factors. However, even though a strong conditional covariance between responses may hint at a biological interaction between species (Pollock et al., 2014), Poggiato et al. (2021) argue the conditional correlations cannot distinguish the biotic from the abiotic effects. Moreover, we henceforth assume that the response matrix is modeled by a thematic partitioning of the explanatory $\boldsymbol{X_1}, \ldots, \boldsymbol{X_R}$, named "themes". We thus search for components in each theme to model $\boldsymbol{Y}$. The linear predictor writes as a linear combination of deterministic latent variables (the components), stochastic latent variables (the factors) and additional covariates

$$\boldsymbol{\eta_k} = \sum_h^{H_1} \left(\boldsymbol{X_1}\boldsymbol{u_1^h}\right)\gamma_{k1}^h + \cdots + \sum_h^{H_R} \left(\boldsymbol{X_R}\boldsymbol{u_R^h}\right)\gamma_{kR}^h + \boldsymbol{A}\boldsymbol{\delta_k} + \boldsymbol{G}\boldsymbol{b_k},$$

where $\boldsymbol{b_k}$ is a vector of parameters. An approach, named Generalized Linear Latent Variable Model (GLLVM, Skrondal and Rabe-Hesketh, 2004), has been proposed to combine

GLM with random latent variables. Unfortunately, in the particular case of factors, the log-likelihood derived from GLLVM cannot be solved analytically. In absence of consensus about the maximization of the log-likelihood, we propose to use a modeling approach based on Saidane et al. (2013), which assumes that this maximization should be performed through the EM algorithm iteratively performed on a linearization of the model. To estimate the model parameters, we present an algorithm encapsulating thematic component-based model estimation and factor model estimation. We name the methodology resulting from this development: Factor SCGLR (F-SCGLR).

Across the manuscript, many simulations schemes are presented with the aim to illustrate the interest of our developments and the good performances of the proposed algorithms. To challenge rmSCGLR, we propose simulation studies with specific explanatory dimensions. The main objective of rmSCGLR in these situations is to detect the true groups of responses. The performances of F-SCGLR are tested by simulating the covariance between the responses so as to get blocks in the conditional variance-covariance matrix. F-SCGLR has to identify these blocks of responses sharing mutual dependencies. All the simulations highlight the importance of a relevant selection of the many hyper-parameters involved in the model. Moreover, rmSCGLR and F-SCGLR are respectively compared to ecology-oriented R packages **ecomix** (Dunstan et al., 2011, 2013) and **gllvm** (Niku et al., 2019b). Finally, in order to give relevant tools for applied statistical modelers as biologists or ecologists, our methods are used to analyze ecology datasets.

1.2 Preliminary notations

The manuscript contains mathematical developments which use notations listed hereafter.

- Let $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N$ be vectors and $\boldsymbol{W} \in \mathbb{R}^{N\times N}$ be a symmetric positive definite matrix. The Euclidean scalar product between $\boldsymbol{a}$ and $\boldsymbol{b}$ with respect to metric $\boldsymbol{W}$ is given by $\langle \boldsymbol{a}, \boldsymbol{b}\rangle_{\boldsymbol{W}} = \boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{b}$. Likewise, $\cos_{\boldsymbol{W}}(\boldsymbol{a}, \boldsymbol{b}) = \dfrac{\langle \boldsymbol{a}, \boldsymbol{b}\rangle_{\boldsymbol{W}}}{\|\boldsymbol{a}\|_{\boldsymbol{W}} \|\boldsymbol{b}\|_{\boldsymbol{W}}}$ denotes the cosine of the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$ with respect to metric $\boldsymbol{W}$.
- If $\boldsymbol{a}$ and $\boldsymbol{b}$ are centred and $\boldsymbol{W} = \frac{1}{N}\boldsymbol{I_N}$, the cosine defines the linear correlation coefficient, denoted ρ. In this paper, unless otherwise stated, the correlation refers to this coefficient.
- $\boldsymbol{A} = [\boldsymbol{a_1}, \ldots, \boldsymbol{a_P}] \in \mathbb{R}^{N\times P}$ and $\boldsymbol{B} = [\boldsymbol{b_1}, \ldots, \boldsymbol{b_Q}] \in \mathbb{R}^{N\times Q}$ being matrices, the space spanned by their column-vectors is denoted by $\text{span}[\boldsymbol{A}, \boldsymbol{B}]$.
- Let ω_n be the weight of unit n, and $\boldsymbol{W} = \text{diag}(\omega_n)_{n=1,\ldots,N}$. Let $\mathbb{R}^N$ be endowed with metric $\boldsymbol{W}$, and let $\boldsymbol{A} \in \mathbb{R}^{N\times P}$ be a matrix. The $\boldsymbol{W}$-orthogonal projector onto $\text{span}[\boldsymbol{A}]$ is given by $\boldsymbol{\Pi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{A}]} = \boldsymbol{A}\left(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{W}$. Thus, the cosine of the

angle between a vector $\boldsymbol{b} \in \mathbb{R}^N$ and $\text{span}[\boldsymbol{A}]$ with respect to metric $\boldsymbol{W}$ is given by $\cos_{\boldsymbol{W}}(\boldsymbol{b}, \text{span}[\boldsymbol{A}]) = \cos_{\boldsymbol{W}}\left(\boldsymbol{b}, \boldsymbol{\Pi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{A}]}\boldsymbol{b}\right)$.

- Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{N\times P}$ be two real matrices. The Frobenius product is computed as $\langle \boldsymbol{A}, \boldsymbol{B}\rangle_{\text{Frob}} = \text{Tr}\left(\boldsymbol{A}^*\boldsymbol{B}\right)$, where Tr denotes the trace of a matrix and $\boldsymbol{A}^* = \boldsymbol{W}^{-1}\boldsymbol{A}^T\boldsymbol{W}$ the adjoint of $\boldsymbol{A}$.
- The unit orthogonal projector with respect to the Frobenius norm is given by $\boldsymbol{\varpi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{A}]} = \boldsymbol{\Pi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{A}]}/\sqrt{\text{rank}\left(\boldsymbol{A}\right)}$.

1.3 Outline

The manuscript is organized as follows. Chapter 2 presents the state-of-the-art of the statistical methodologies we need to develop our approaches. An introduction of the GLM is given in Section 2.1, the original component-based models and SCGLR are recalled in Section 2.2, while Section 2.3 and Section 2.4 respectively focus on the presentation of the usual finite mixture and factor models. Chapter 3 is dedicated to the response mixture extension of SCGLR we propose. As for the factor model extension, Chapter 4 details how we relax the independence assumption between the responses. The supplementary materials we need in this work are given in Chapter 6.

CHAPTER 2

STATE-OF-THE-ART

Contents

The main objective of this chapter is to present a non-exhaustive state-of-the-art of the statistical research fields encountered in the thesis.

2.1 Generalized Linear Models

The Generalized Linear Models (GLM) are introduced by Nelder and Wedderburn (1972) in a context where the Gaussian distribution assumption is inappropriate, as for qualitative or discrete data. GLMs cover the modelisation of all these types of data by allowing the random response variables to have any distribution from the exponential family. Contrary to the original linear model, the expected value of the random response variable is not directly equal to the linear predictor defined by a linear combination of the explanatory variables, but as a function which links the response variable and the explanatory variables. For further details, McCullagh and Nelder (1989) propose a complete overview of this subject, and Fahrmeir (1994) extend this overview to multivariate data analysis.

2.1.1 Definition of Generalized Linear Models

Let $\{y_n, n = 1, \ldots, N\}$ be a sample of the random variable $\boldsymbol{y}$ having one of the distributions of the exponential family. The y_n's are assumed independent and their distribution can be expressed in the form

$$\begin{aligned} L_n(y_n;\theta_n) &= \exp\left(\frac{y_n\theta_n - b(\theta_n)}{a_n(\phi)} + c(y_n,\phi)\right) \\ &= \exp\left(y_n\frac{\theta_n}{a_n(\phi)}\right)\exp\left(-\frac{b(\theta_n)}{a_n(\phi)}\right)\exp\left(c(y_n,\phi)\right), \end{aligned} \tag{2.1}$$

where a_n, b and c are known functions depending on the type of the distribution, θ_n is a canonical parameter and ϕ is the dispersion parameter. The function a_n writes $a_n = \phi/\omega_n$, where w_n is the weight associated with the nth statistical unit. For all the distributions belonging to the exponential family, the expectation and the variance can be expressed thanks to the functions a_n and b. Let

$$l\left(\boldsymbol{\Theta};\boldsymbol{y}\right) = \sum_{n=1}^{N} l_n\left(\theta_n;y_n\right) = \sum_{n=1}^{N} \ln\left(L_n\left(y_n;\theta_n\right)\right)$$

be the sample log-likelihood. The general conditions given by Kendall and Stuart (1961, pages 8–9) yield

$$\begin{aligned} \mathbb{E}\left[\frac{\partial l_n}{\partial\theta_n}\right] = 0 &\Leftrightarrow \mathbb{E}\left[\frac{y_n - b'(\theta_n)}{a_n(\phi)}\right] = 0 \\ &\Leftrightarrow \mathbb{E}\left[y_n\right] = b'(\theta_n) \end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial^2 l_n}{\partial \theta_n^2}\right] + \mathbb{E}\left[\left(\frac{\partial l_n}{\partial \theta_n}\right)^2\right] = 0 \qquad (2.2)\\
&\Leftrightarrow \mathbb{E}\left[-\frac{b''(\theta_n)}{a_n(\phi)}\right] + \mathbb{E}\left[\left(\frac{y_n - b'(\theta_n)}{a_n(\phi)}\right)^2\right] = 0\\
&\Leftrightarrow \frac{\mathbb{E}\left[(y_n - \mathbb{E}[y_n])^2\right]}{a_n(\phi)^2} = \frac{b''(\theta_n)}{a_n(\phi)}\\
&\Leftrightarrow \mathbb{V}[y_n] = a_n(\phi) b''(\theta_n).
\end{aligned}$$

We can thus rewrite the variance as a function of the expectation of y_n

$$\mathbb{V}[y_n] = a_n(\phi)\, b'' \circ (b')^{-1} (\mathbb{E}[y_n]).$$

Denoting $\mu_n := \mathbb{E}[y_n]$ and $v := b'' \circ (b')^{-1}$, the independence of the y_n's leads to

$$\mathbb{V}[\boldsymbol{y}] = \text{diag}\left(a_n(\phi) v(\mu_n)\right)_{n=1,\dots,N}.$$

As in the classic linear model, we may introduce a set $\boldsymbol{X} \in \mathbb{R}^{N\times P}$ of explanatory variables linearly involved in the model through the linear predictor $\boldsymbol{\eta}$ expressed as

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta},$$

where $\boldsymbol{\beta} \in \mathbb{R}^P$ is the regression parameter vector. Moreover, contrary to linear models, the linear predictor is connected to the expected value through a strictly monotonic and twice-differentiable link function h such that, for all n

$$\eta_n = h(\mu_n).$$

The link function which associates the expected value to the canonical parameter is called the canonical link function. In this case, we have

$$\eta_n = \theta_n.$$

2.1.2 Maximum likelihood estimation

Now, we present the process to maximize the likelihood of a GLM. We recall the log-likelihood function

$$l(\boldsymbol{\Theta}; \boldsymbol{y}) = \sum_{n=1}^{N} l_n(\theta_n; y_n),$$

where $\boldsymbol{\Theta} = \{\theta_n; n = 1, \dots, N\}$ is the vector of parameters and

$$\begin{aligned}
l_n(\theta_n; y_n) &= \ln\left(L_n(y_n; \theta_n)\right)\\
&= \frac{y_n\theta_n - b(\theta_n)}{a_n(\phi)} + c(y_n, \phi).
\end{aligned}$$

2.1.2.1 The maximum likelihood estimation from the chain derivative rule

The maximum likelihood estimation equations with respect to the parameter vector $\boldsymbol{\beta}$ are obtained from the chain derivative rule. For each $n = 1, \dots, N$ and for each $p = 1, \dots, P$, we then have

$$\begin{aligned}\frac{\partial l_n}{\partial \beta_p} &= \frac{\partial l_n}{\partial \theta_n}\frac{\partial \theta_n}{\partial \mu_n}\frac{\partial \mu_n}{\partial \eta_n}\frac{\partial \eta_n}{\partial \beta_p} \\ &= \frac{y_n - b'(\theta_n)}{a_n(\phi)}\frac{1}{b''(\theta_n)}\frac{1}{h'(\mu_n)}x_{np}.\end{aligned}$$

Thus,

$$\frac{\partial l}{\partial \beta_p} = \sum_{n=1}^{N} x_{np}\frac{1}{\mathbb{V}[y_n]h'(\mu_n)^2}h'(\mu_n)(y_n - \mu_n).$$

Finally, the maximum likelihood of the vector $\boldsymbol{\beta}$ is a solution of

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = 0 \Leftrightarrow \boldsymbol{X^T W}\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}(\boldsymbol{y} - \boldsymbol{\mu}) = 0, \tag{2.3}$$

where we denote

$$\boldsymbol{W} := \text{diag}\left(\frac{1}{\mathbb{V}[y_n]h'(\mu_n)^2}\right)_{n=1,\dots,N}$$

and

$$\begin{aligned}\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}} &:= \text{diag}\left(\frac{\partial \eta_n}{\partial \mu_n}\right)_{n=1,\dots,N} \\ &:= \text{diag}\left(h'(\mu_n)\right)_{n=1,\dots,N}.\end{aligned}$$

We may note that, since the matrix $\boldsymbol{W}$ and $\dfrac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}$, as well as the vector $\boldsymbol{\mu}$, depend on $\boldsymbol{\beta}$, Equation (2.3) is not linear in $\boldsymbol{\beta}$. So, we shall use an iterative process to estimate the parameters.

2.1.2.2 Two iterative estimation methods

Two classic iterative methods can thus be performed to maximize the likelihood: the Newton-Raphson (NR) method and the Fisher Scoring Algorithm (FSA). At iteration t, the estimation $\boldsymbol{\beta}^{(t+1)}$ is obtained from the previous estimate $\boldsymbol{\beta}^{(t)}$ by

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\mathbb{E}\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta^T}}\right]^{(t)}\right)^{-1}\left(\frac{\partial l}{\partial \boldsymbol{\beta}}\right)^{(t)} \quad \text{for the FSA,} \tag{2.4}$$

and

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\boldsymbol{T}}} \right)^{(t)} \right)^{-1} \left(\frac{\partial l}{\partial \boldsymbol{\beta}} \right)^{(t)} \quad \text{for the NR method.} \tag{2.5}$$

As mentioned by Osborne (1992), the FSA inherits most of the good properties of the NR method. Furthermore, in a context of distributions belonging to the exponential family, the calculation of the expectation of the Hessian matrix is possible. Indeed, the constraint given by Equation (2.2) gives for the general term of Equation (2.4)

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2 l}{\partial \beta_p \partial \beta_q} \right] &= \sum_{n=1}^{N} \mathbb{E}\left[\frac{\partial^2 l_n}{\partial \beta_p \partial \beta_q} \right] \\
&= -\sum_{n=1}^{N} \mathbb{E}\left[\left(\frac{\partial l_n}{\partial \beta_p} \right) \left(\frac{\partial l_n}{\partial \beta_q} \right) \right] \\
&= -\sum_{n=1}^{N} \frac{x_{np} x_{nq}}{\mathbb{V}[y_n] h'(\mu_n)^2} \mathbb{E}\left[\frac{h'(\mu_n)^2 (y_n - \mu_n)^2}{\mathbb{V}[y_n] h'(\mu_n)^2} \right] \\
&= -\sum_{n=1}^{N} \frac{x_{np} x_{nq}}{\mathbb{V}[y_n] h'(\mu_n)^2}.
\end{aligned}$$

As noticed by Nelder and Wedderburn (1972) and detailed by McCullagh and Nelder (1989), the FSA and the NR method are equivalent in the case of a canonical link. Indeed, the canonical link is defined by

$$\eta_n = \theta_n = h(\mu_n) = x_n^T \beta$$

or, in other words

$$h = (b')^{-1} \quad \text{and} \quad h' = \frac{1}{b'' \circ (b')^{-1}}.$$

In this case, we have

$$\frac{\partial \mu_n}{\partial \eta_n} = \frac{\partial \mu_n}{\partial \theta_n} = \frac{\partial b'(\theta_n)}{\partial \theta_n} = b''(\theta_n).$$

The chain derivative thus becomes

$$\begin{aligned}
\frac{\partial l_n}{\partial \beta_p} &= \frac{\partial l_n}{\partial \theta_n} \frac{\partial \theta_n}{\partial \mu_n} \frac{\partial \mu_n}{\partial \eta_n} \frac{\partial \eta_n}{\partial \beta_p} \\
&= \frac{y_n - b'(\theta_n)}{a_n(\phi)} \frac{1}{b''(\theta_n)} b''(\theta_n) x_{np} \\
&= x_{np} \frac{y_n - b'(\theta_n)}{a_n(\phi)}.
\end{aligned}$$

Finally, the terms involving the second derivatives in Equation (2.5) are calculated as

$$\begin{aligned}\frac{\partial^2 l_n}{\partial\beta_p\partial\beta_q} &= \frac{\partial}{\partial\beta_q}\left(x_{np}\frac{y_n-\mu_n}{a_n(\phi)}\right)\\ &= -\frac{x_{np}}{a_n(\phi)}\frac{\partial\mu_n}{\partial\beta_q}\\ &= -\frac{x_{np}}{a_n(\phi)}\frac{\partial\mu_n}{\partial\eta_n}\frac{\partial\eta_n}{\partial\beta_q}\\ &= -x_{np}x_{nq}\frac{b''(\theta_n)}{a_n(\phi)}\\ &= -\frac{x_{np}x_{nq}}{\mathbb{V}[y_n]h'(\mu_n)^2}.\end{aligned}$$

This proves the equivalence of the FSA and the NR method for the canonical link function. Thus, in both cases, the matrix part of Equation (2.4) and Equation (2.5) writes

$$\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\boldsymbol{T}}} = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X}.$$

2.1.2.3 The IRLS algorithm to estimate GLM parameters

In the following, we use the previous iteration in order to estimate the parameter vector $\boldsymbol{\beta}$. Then, step $t+1$ writes

$$\boldsymbol{\beta}^{(t+1)} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}^{(t)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}^{(t)}\boldsymbol{w}^{(t)} \tag{2.6}$$

where for all n, the working variable (or pseudo-response) w_n is calculated as the first order expansion of h at point $\mu_n^{(t)}$

$$\begin{aligned}w_n^{(t)} &= h\left(\mu_n^{(t)}\right) + h'\left(\mu_n^{(t)}\right)\left(y_n-\mu_n^{(t)}\right)\\ &= \boldsymbol{x_n^T}\boldsymbol{\beta}^{(t)} + h'\left(\mu_n^{(t)}\right)\left(y_n-\mu_n^{(t)}\right).\end{aligned}$$

Equation (2.6) can be seen as a least squares regression step of $\boldsymbol{w}^{(t)}$ on $\boldsymbol{X}$ weighted by $\boldsymbol{W}^{(t)}$ in the linearized model

$$\boldsymbol{w}^{(t)} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\zeta}^{(t)},$$

where $\mathbb{E}[\boldsymbol{\zeta}^{(t)}] = 0$ and $\mathbb{V}[\boldsymbol{\zeta}^{(t)}] = \boldsymbol{W^{(t)-1}}$. This development leads to the iterations of the Iteratively Re–weighted Least Squares (IRLS, Green, 1984) algorithm recalled in Algorithm 1.

In the particular case of the canonical link, the update of the weight matrix is simplified. We can recalculate $\boldsymbol{W}$ by

$$\boldsymbol{W}^{(t+1)} = \operatorname{diag}\left(\left[a_n(\phi)h'\left(\mu_n^{(t)}\right)\right]^{-1}\right)_{n=1,\dots,N}.$$

Algorithm 1: The IRLS algorithm

while *not convergence* **do**

$$\boldsymbol{w}^{(t+1)} = h\left(\boldsymbol{\mu}^{(t)}\right) + \text{diag}\left(h'\left(\boldsymbol{\mu}^{(t)}\right)\right)\left(\boldsymbol{y} - \boldsymbol{\mu}^{(t)}\right)$$

$$\boldsymbol{W}^{(t+1)} = \text{diag}\left(\left[a_n(\phi)v\left(\mu_n^{(t)}\right)h'\left(\mu_n^{(t)}\right)^2\right]^{-1}\right)_{n=1,\ldots,N}$$

$$\boldsymbol{\beta}^{(t+1)} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}^{(t+1)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}^{(t+1)}\boldsymbol{w}^{(t+1)}$$

$$\boldsymbol{\mu}^{(t+1)} = h^{-1}\left(\boldsymbol{X}\boldsymbol{\beta}^{(t+1)}\right)$$

$$t \leftarrow t+1$$

end

2.1.3 Special cases of GLMs

Three particular cases of GLM are used in this thesis and deserve mentioning

1. **The Gaussian distribution**
   A continuous random variable $y \sim \mathcal{N}(\mu, \sigma^2)$ has a probability density function of the form

$$\begin{aligned} L(y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left(y\frac{\mu}{\sigma^2}\right)\exp\left(-\frac{1}{2}\frac{\mu^2}{\sigma^2}\right)\exp\left(-\frac{1}{2}\frac{y^2}{\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right), \end{aligned}$$

   where $\theta = h(\mu) = \mu$, which characterizes the canonical link function as the identity function. The known functions are respectively defined by $a(\sigma^2) = \sigma^2$ and $b(\theta) = \frac{1}{2}\theta^2$. Thus, the expectation and variance are respectively given by

$$\begin{aligned} \mathbb{E}[y] &= b'(\theta) = \theta = \mu \\ \mathbb{V}[y] &= a(\sigma^2)b''(\theta) = \sigma^2. \end{aligned}$$

2. **The Poisson distribution**
   A discrete random variable $y \sim \mathcal{P}(\mu)$ has a mass function of the form

$$\begin{aligned} L(y;\mu) &= \frac{\mu^y\exp(-\mu)}{y!} \\ &= \exp\left(y\ln\left(\mu\right)\right)\exp(-\mu)\exp\left(-\ln\left(y!\right)\right), \end{aligned}$$

   where the canonical link function $\theta = h(\mu) = \ln(\mu)$ is defined by the logarithmic function. The known functions are respectively defined by $a(\phi) = 1$ and $b(\theta) =$

$\exp(\theta)$. Thus, the expectation and variance are respectively given by

$$\mathbb{E}[y] = b'(\theta) = \exp(\theta) = \exp(\ln(\mu)) = \mu$$
$$\mathbb{V}[y] = a(\phi)b''(\theta) = \exp(\theta) = \exp(\ln(\mu)) = \mu.$$

3. **The Bernoulli distribution**
A binary random variable $y \sim \mathcal{B}(\mu)$ has a mass function of the form

$$\begin{aligned} L(y;\mu) &= \mu^y(1-\mu)^{1-y} \\ &= \exp\left(y \ln\left(\frac{\mu}{1-\mu}\right)\right) \exp\left(-\ln\left(\frac{1}{1-\mu}\right)\right) \end{aligned}$$

where the canonical link function $\theta = h(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$ is defined by the logit function. The known functions are respectively defined by $a(\phi) = 1$, $b'(\theta) = \text{logit}^{-1}(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)}$ and $b''(\theta) = \frac{\exp(\theta)}{(1+\exp(\theta))^2}$. So, the expectation and the variance are respectively given by

$$\mathbb{E}[y] = b'(\theta) = \text{logit}^{-1}(\theta) = \text{logit}^{-1}(\text{logit}(\mu)) = \mu$$
$$\mathbb{V}[y] = a(\phi)b''(\theta) = \frac{\mu/(1-\mu)}{(1+\mu/(1-\mu))^2} = \frac{\mu/(1-\mu)}{1/(1-\mu)^2} = \mu(1-\mu).$$

For more examples about distributions from the exponential family, see Trottier (1998).

2.2 Component-based models

The main idea of the component–based models is to assume that the information contained in the explanatory variables $\boldsymbol{X}$ should be summarized into a small number of synthetic variables $\boldsymbol{f_1}, \ldots, \boldsymbol{f_H}$ called "components". The latter are defined as linear combinations of the original explanatory variables and need to be orthogonal with the aim to avoid the linear redundancy of the information. Each component then writes $\boldsymbol{f_h} = \boldsymbol{X}\boldsymbol{u_h}$, where $\boldsymbol{u_h}$ is a loading vector. In a regression context, the components are considered as new non correlated explanatory variables to replace $\boldsymbol{X}$. The different component-based approaches presented in this part only differ in the manner the components are built.

2.2.1 Reminder of PCR and PLSR

We consider the classical linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. We assume that the response is centered and that the P explanatory variables are normalized.

2.2.1.1 Principal Components Regression (PCR)

In this approach, the linear combinations of the explanatory variables $\boldsymbol{f_1}, \ldots, \boldsymbol{f_H}$ are the principal components. The first principal component $\boldsymbol{f_1} = \boldsymbol{X}\boldsymbol{u_1}$ is designed to capture as much empirical variance as possible in $\boldsymbol{X}$. So, $\boldsymbol{u_1}$ is the solution of the optimization program

$$\begin{aligned}\max_{\boldsymbol{u}^T\boldsymbol{u}=1} \mathbb{V}[\boldsymbol{X}\boldsymbol{u}] &\Leftrightarrow \max_{\boldsymbol{u}^T\boldsymbol{u}=1} \|\boldsymbol{X}\boldsymbol{u}\|_2^2 \\ &\Leftrightarrow \max_{\boldsymbol{u}^T\boldsymbol{u}=1} \boldsymbol{u}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{u}.\end{aligned}$$

Each subsequent component then captures as much as possible of $\boldsymbol{X}$'s variance not accounted for by former components, under the constraint that it is orthogonal to the former components.

Now, let us take the principal components as a set of non-redundant explanatory variables. Due to the orthogonality, the predictor of the random response variable $\boldsymbol{y}$ writes

$$\widehat{\boldsymbol{y}}_{\text{PCR}} = \sum_{h=1}^{H} \widehat{\gamma}_h \boldsymbol{f}_h, \tag{2.7}$$

where $\widehat{\gamma}_h$ is the coefficient of the classical regression of $\boldsymbol{y}$ on $\boldsymbol{f}_h$, that is

$$\widehat{\gamma}_h = \left(\boldsymbol{f}_h^T\boldsymbol{f}_h\right)^{-1}\boldsymbol{f}_h^T\boldsymbol{y} = \frac{\langle \boldsymbol{f}_h, \boldsymbol{y}\rangle}{\langle \boldsymbol{f}_h, \boldsymbol{f}_h\rangle}. \tag{2.8}$$

The predictor given by Equation (2.7) can then be expressed with respect to the original explanatory variables

$$\widehat{\boldsymbol{y}}_{\text{PCR}} = \sum_{h=1}^{H} \widehat{\gamma}_h \boldsymbol{X}\boldsymbol{u}_h = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{PCR}},$$

where $\widehat{\boldsymbol{\beta}}_{\text{PCR}} = \sum_{h=1}^{H} \widehat{\gamma}_h \boldsymbol{u}_h$.

An important disadvantage of PCR is that the principal components do not take into account the response variable $\boldsymbol{y}$ in their construction. Thus, in order to model the response $\boldsymbol{y}$, supervised components should be preferred. This is, for instance, what does the Partial Least Squares Regression (PLSR).

2.2.1.2 Partial Least Squares Regression (PLSR)

Originally introduced by Wold (1966), the PLS regression has become a standard in applied statistics, particularly in the field of chemometrics (Fonville et al., 2010). Contrary to PCR, the loading vector $\boldsymbol{u}_h$ maximizes the covariance between the component $\boldsymbol{f}_h$ and the response $\boldsymbol{y}$. The optimization program thus becomes

$$\boldsymbol{u}_h = \begin{cases} \underset{\boldsymbol{u}^T\boldsymbol{u}=1}{\text{argmax}}\ \text{cov}\,(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{y}) \\ \boldsymbol{X}\boldsymbol{u} \perp \boldsymbol{f_1}, \ldots, \boldsymbol{f_{h-1}} \end{cases} = \begin{cases} \underset{\boldsymbol{u}^T\boldsymbol{u}=1}{\text{argmax}}\ \langle \boldsymbol{X}\boldsymbol{u}, \boldsymbol{y}\rangle \\ \boldsymbol{X}\boldsymbol{u} \perp \boldsymbol{f_1}, \ldots, \boldsymbol{f_{h-1}} \end{cases}.$$

The orthogonality between the components can be ensured either by deflating the design matrix $\boldsymbol{X}$ at each step of the PLS algorithm or by adding an orthogonality constraint to the optimization program. Moreover, it is straightforward to show that the hth PLS loading vector $\boldsymbol{u_h}$ also solves

$$\boldsymbol{u_h} = \begin{cases} \underset{\boldsymbol{u}^T\boldsymbol{u}=1}{\operatorname{argmax}} \ \operatorname{cov}\left(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{y}\right)^2 \\ \boldsymbol{X}\boldsymbol{u} \perp \boldsymbol{f_1}, \ldots, \boldsymbol{f_{h-1}} \end{cases} = \begin{cases} \underset{\boldsymbol{u}^T\boldsymbol{u}=1}{\operatorname{argmax}} \ \rho\left(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{y}\right)^2 \mathbb{V}[\boldsymbol{X}\boldsymbol{u}] \\ \boldsymbol{X}\boldsymbol{u} \perp \boldsymbol{f_1}, \ldots, \boldsymbol{f_{h-1}} \end{cases},$$

indicating that the PLS components attempt a trade-off between capturing the highest variance in $\boldsymbol{X}$ (through $\mathbb{V}[\boldsymbol{X}\boldsymbol{u}]$) and modeling the response vector $\boldsymbol{y}$ (through $\rho\left(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{y}\right)^2$). As a result, there exists a vector of coefficients $\widehat{\boldsymbol{\beta}}_{\text{PLSR}}$, which can be expressed with respect to the sequence of PLS loading vectors, such that $\widehat{\boldsymbol{y}}_{\text{PLSR}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{PLSR}}$.

2.2.2 Extension to GLM

We henceforth present few extensions of PLSR that have been proposed in the literature to deal with the GLM framework.

2.2.2.1 Principal Component Generalized Linear Regression (PCGLR)

Extending PCR to PCGLR (Marx and Smith, 1990) is straightforward: one just has to use the principal components as the GLM's new explanatory variables. Indeed, in the IRLS algorithm, the regression parameter of the hth principal component writes

$$\widehat{\gamma_h} = \frac{\langle \boldsymbol{f_h}, \boldsymbol{w} \rangle}{\langle \boldsymbol{f_h}, \boldsymbol{f_h} \rangle},$$

where $\boldsymbol{w}$ is the working variable. However, PCGLR suffers from the same drawback as PCR: the principal components are not supervised by the response vector $\boldsymbol{y}$.

2.2.2.2 PLS Generalized Linear Regression

The approach proposed by Bastien et al. (2005) elaborates on the fact that the PLSR of a response $\boldsymbol{y}$ on $\boldsymbol{X} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_P}]$, where the explanatory variables are standardized, gives a component of the form

$$\boldsymbol{f_1} = \frac{\sum_{p=1}^{P} \operatorname{cov}(\boldsymbol{y}, \boldsymbol{x_p}) \boldsymbol{x_p}}{\sqrt{\sum_{p=1}^{P} \operatorname{cov}(\boldsymbol{y}, \boldsymbol{x_p})^2}},$$

where the covariance between $\boldsymbol{y}$ and the explanatory variable $\boldsymbol{x_p}$ represents the coefficient associated with the simple Ordinary Least Squares (OLS) regression of $\boldsymbol{y}$ on $\boldsymbol{x_p}$. The higher rank components can be produced in the same way, after replacing each explanatory variable $\boldsymbol{x_p}$ with its OLS regression residuals on $\boldsymbol{f_{h-1}}$. Hence it is proposed to substitute

the OLS regressions by the standardized sum of predictors given by Generalized Linear Regression (GLR) of $\boldsymbol{y}$ on each $\boldsymbol{x}_p$ to extend this method to GLM.

As highlighted by Bry et al. (2013) and Chauvet (2019), this extension may seem awkward due to the inconsistency of the weighting of the observations. Indeed, GLR of $\boldsymbol{y}$ on $\boldsymbol{x}_p$ alone implicitly uses a specific weighting matrix $\boldsymbol{W}_p$, which is different from the weighting matrix associated with GLR of $\boldsymbol{y}$ on components. Thus, the estimated variance structure of observations according to the model based on components is never used by this method.

2.2.2.3 Iteratively Reweighted Partial Least Squares

Another way to extend the PLSR to GLM is the Iteratively Re-weighted PLS (IRPLS) approach developed by Marx (1996). IRPLS might thus be viewed as an IRLS algorithm in which the step of weighted least squares regression used to update parameters is substituted by a weighted PLS regression. More formerly, let $\boldsymbol{w}^{(t)}$ and $\boldsymbol{W}^{(t)}$ respectively be the working variable and the weight matrix at the t-th iteration of the IRLS. Instead of the classical update for $\boldsymbol{\beta}$, also given by Equation (2.6), Marx (1996) rather suggests to take

$$\boldsymbol{\beta}^{(t+1)} = \text{PLSR}_{\boldsymbol{W}^{(t)}}\left(\boldsymbol{w}^{(t)}, \boldsymbol{X}\right),$$

where $\text{PLSR}_{\boldsymbol{W}^{(t)}}\left(\boldsymbol{w}^{(t)}, \boldsymbol{X}\right)$ refers to the PLS regression of $\boldsymbol{w}^{(t)}$ on $\boldsymbol{X}$, where the observations are weighted by $\boldsymbol{W}^{(t)}$. Thus, contrary to PLSGLR, the weighting matrix derived from the GLM's maximum likelihood estimation is taken into account in the PLS regression. This method has been extended to the multivariate case by Bry et al. (2013).

2.2.3 Supervised Component-based Generalized Linear Regression

Elaborating on the Iteratively Reweighted Partial Least Squares (IRPLS) developed by Marx (1996), Bry et al. (2013) proposed a methodology called Supervised Component-based Generalized Linear Regression (SCGLR) which combines the multivariate Generalized Linear Model (GLM) estimation with the component-based dimension reduction of the explanatory space. Unlike methods as Partial Least Squares (PLS, Wold et al., 1984) regression or Reduced Rank Vector Generalized Linear Model (RRVGLM, Yee and Hastie, 2003), SCGLR optimizes a general and flexible trade-off criterion between the Goodness-of-Fit (GoF) of the model and the Structural Relevance (SR, Bry and Verron, 2015) of directions with respect to the explanatory variables. This methodology allows both to find strong interpretable explanatory directions, and to produce regularized predictors in the high-dimensional framework. Later on, SCGLR has been extended and refined in several ways. First, Chauvet et al. (2019) proposed to combine SCGLR with Schall's algorithm to estimate a model with random effects. This extension aims at modeling responses with repeated measures or a group design on individuals. Later yet, SCGLR was extended to Cox's model for survival data (Bry et al., 2020a). More recently, Bry et al. (2020b)

extended it to a partitioning of the explanatory variables, following the approach of Bry and Verron (2015) called THEmatic Model Exploration (THEME). In this context, Generalized Linear Regression (GLR) demands dimension reduction and regularization with respect to each theme, i.e. variable group. An R package **SCGLR** (Mortier et al., 2016) is available at https://CRAN.R-project.org/package=SCGLR but we recommend to use the link https://github.com/SCnext/SCGLR to access the latest version.

2.2.3.1 The SCGLR context

In the framework of a multivariate GLM (Fahrmeir and Tutz, 2013), we consider K response-vectors encoded in a response matrix $\boldsymbol{Y} = [\boldsymbol{y_1}, \ldots, \boldsymbol{y_K}] \in \mathbb{R}^{N \times K}$, to be predicted through explanatory variables partitioned into two groups. The first one $\boldsymbol{A} = [\boldsymbol{a_1}, \ldots, \boldsymbol{a_R}] \in \mathbb{R}^{N \times R}$ is a group of covariates that are only few, and weakly or not redundant. These variables are *a priori* assumed to be interesting per se, and their marginal effects have to be taken into account explicitly in the model. The second group $\boldsymbol{X} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_P}] \in \mathbb{R}^{N \times P}$ is one of numerous and possibly highly redundant covariates, considered as proxies to latent dimensions, which must be found and interpreted. Thus, the matrix $\boldsymbol{X}$ demands dimension reduction and regularization. To achieve this, SCGLR searches for explanatory components in $\boldsymbol{X}$ jointly supervised by the response set. In this part, for simplicity's sake, we shall consider a single-component model. A component $\boldsymbol{f} \in \mathbb{R}^N$ writes $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$, where $\boldsymbol{u} \in \mathbb{R}^P$ is a loading vector. The linear predictor associated with response $\boldsymbol{y_k}$ is then given by

$$\boldsymbol{\eta_k} = (\boldsymbol{X}\boldsymbol{u})\,\boldsymbol{\gamma_k} + \boldsymbol{A}\boldsymbol{\delta_k},$$

where $\boldsymbol{\gamma_k}$ and $\boldsymbol{\delta_k}$ are regression parameters. The component $\boldsymbol{f}$ is common to all the responses, and for identification, we impose $\boldsymbol{u^T M^{-1} u} = 1$, where $\boldsymbol{M} \in \mathbb{R}^{P \times P}$ is a given symmetric positive definite matrix. It is assumed that the responses are independent conditional on the explanatory variables, and consequently on the component.

2.2.3.2 Measuring the Goodness-of-Fit

Given the component, the parameters of the GLM must be estimated, and we refer the reader to McCullagh and Nelder (1989) for a complete overview of GLM methodologies. Here, we make use of the Fisher Scoring Algorithm (FSA). Let $\boldsymbol{w_k}$ be the working variable associated with the response $\boldsymbol{y_k}$, and $\boldsymbol{W_k^{-1}}$ its variance matrix. In the spirit of Nelder and Wedderburn (1972), $\boldsymbol{w_k}$ can be viewed as the response in the linearized model

$$\boldsymbol{w_k} = (\boldsymbol{X}\boldsymbol{u})\,\boldsymbol{\gamma_k} + \boldsymbol{A}\boldsymbol{\delta_k} + \boldsymbol{\zeta_k},$$

with $\mathbb{E}[\boldsymbol{\zeta_k}] = 0$ and $\mathbb{V}[\boldsymbol{\zeta_k}] = \boldsymbol{W_k^{-1}}$. Due to the product $\boldsymbol{u\gamma_k}$, this linearized model derived from the FSA is not linear indeed, and must be estimated through an alternated weighted least squares process, estimating in turn $\{\boldsymbol{\gamma_k}, \boldsymbol{\delta_k}\}$ and $\boldsymbol{u}$.

Let $\mathbf{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{f},\boldsymbol{A}]}$ be the projection on span$[\boldsymbol{f}, \boldsymbol{A}]$ with respect to $\boldsymbol{W_k}$. The loading vector $\boldsymbol{u}$ solution of the least squares minimization may alternatively be viewed as the solution of the following equivalent optimization programs

$$
\begin{aligned}
&\min_{\boldsymbol{u},\boldsymbol{u^T M^{-1} u}=1} \sum_{k=1}^{K} \alpha_k \left\| \boldsymbol{w_k} - \mathbf{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{f},\boldsymbol{A}]} \boldsymbol{w_k} \right\|^2_{\boldsymbol{W_k}} \\
\Leftrightarrow &\min_{\boldsymbol{u},\boldsymbol{u^T M^{-1} u}=1} \sum_{k=1}^{K} \alpha_k \left\| \mathbf{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{f},\boldsymbol{A}]^\perp} \boldsymbol{w_k} \right\|^2_{\boldsymbol{W_k}} \\
\Leftrightarrow &\min_{\boldsymbol{u},\boldsymbol{u^T M^{-1} u}=1} \sum_{k=1}^{K} \alpha_k \left\| \boldsymbol{w_k} \right\|^2_{\boldsymbol{W_k}} \sin^2_{\boldsymbol{W_k}} \left(\boldsymbol{w_k}, \mathbf{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{f},\boldsymbol{A}]} \boldsymbol{w_k} \right) \\
\Leftrightarrow &\min_{\boldsymbol{u},\boldsymbol{u^T M^{-1} u}=1} \sum_{k=1}^{K} \alpha_k \left\| \boldsymbol{w_k} \right\|^2_{\boldsymbol{W_k}} \left[1 - \cos^2_{\boldsymbol{W_k}} \left(\boldsymbol{w_k}, \mathbf{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{f},\boldsymbol{A}]} \boldsymbol{w_k} \right) \right] \\
\Leftrightarrow &\max_{\boldsymbol{u},\boldsymbol{u^T M^{-1} u}=1} \psi_{\boldsymbol{A}}(\boldsymbol{u}),
\end{aligned}
$$

with $\psi_{\boldsymbol{A}}(\boldsymbol{u}) = \sum_{k=1}^{K} \alpha_k \left\| \boldsymbol{w_k} \right\|^2_{\boldsymbol{W_k}} \cos^2_{\boldsymbol{W_k}} \left(\boldsymbol{w_k}, \mathbf{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{f},\boldsymbol{A}]} \boldsymbol{w_k} \right)$, where $\{\alpha_1, \ldots, \alpha_K\}$ is a weighting system reflecting the *a priori* relative importance of working variables. Now, $\psi_{\boldsymbol{A}}$ is merely a Goodness-of-Fit (GoF) measure, and maximizing it does not lead to strong and interpretable components. The GoF measure must therefore be aptly combined with a measure of Structural Relevance (SR) to achieve both meaningful and predictive dimension reduction, together with regularization.

2.2.3.3 Measuring the Structural Relevance of components

Bry and Verron (2015) proposed the SR measure as a possible extension of the component's variance to measure the ability of a component to capture information in a set of variables containing latent structures such as variable-bundles. Informally, a bundle is a set of variables correlated "enough" to be viewed as produced by a common latent dimension. Let $\boldsymbol{\mathcal{N}} = \{\boldsymbol{N_1}, \ldots, \boldsymbol{N_J}\}$ be a set of J symmetric semi-definite positive matrices, $\boldsymbol{\Omega} = \{\omega_1, \ldots, \omega_J\}$ a set of weights and l a scalar such that $l \geq 1$. We call $\boldsymbol{W}$ the weight matrix reflecting the *a priori* relative importance of observations (typically, $\boldsymbol{W} = \frac{1}{N}\boldsymbol{I_N}$). Finally, consider component $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$, where $\boldsymbol{u}$ is constrained by $\boldsymbol{u^T M^{-1} u} = 1$. The associated SR measure ϕ is defined as the following generalized average of quadratic forms of $\boldsymbol{u}$

$$
\phi(\boldsymbol{u}) = \left(\sum_{j=1}^{J} \omega_j \left(\boldsymbol{u^T N_j u} \right)^l \right)^{1/l}.
$$

The matrices $\boldsymbol{N_j}$ are chosen such that the quadratic forms $\boldsymbol{u^T N_j u}$ measure the closeness of the loading vector $\boldsymbol{u}$, or equivalently the corresponding component, to some reference structures (variable-bundles or subspaces). Typically, if $\boldsymbol{M^{-1}} = \boldsymbol{I_P}$ and $\boldsymbol{N_j}$ is the orthog-

onal projector on a reference subspace $\boldsymbol{S_j}$, that is $\boldsymbol{N_j} = \boldsymbol{\Pi}_{S_j}$, then

$$\begin{aligned}\boldsymbol{u^T N_j u} &= \langle \boldsymbol{N_j u}, \boldsymbol{N_j u} \rangle \quad \text{since} \quad \boldsymbol{N_j} = \boldsymbol{N_j N_j} = \boldsymbol{N_j^* N_j} \\ &= \frac{\|\boldsymbol{N_j u}\|^2}{\|\boldsymbol{u}\|^2} \quad \text{due to} \quad \|\boldsymbol{u}\|^2 = 1 \\ &= \cos^2(\boldsymbol{u}, \boldsymbol{S_j}) .\end{aligned}$$

The locality of the bundles to be tracked by components is tuned through the hyper-parameter l. Components will line up with a more or less local bundle depending on whether l is greater or smaller, respectively. To illustrate this, let us pick up extreme values of l:

- If $l = 1$, the SR writes

$$\phi(\boldsymbol{u}) = \sum_{j=1}^{J} \omega_j \boldsymbol{u^T N_j u} = \boldsymbol{u^T} \left(\sum_{j=1}^{J} \omega_j \boldsymbol{N_j} \right) \boldsymbol{u}.$$

  The directions having the maximal structural relevance being the principal components, the maximization of this quantity leads to the first eigenvector of the corresponding PCA. In the particular case where $\boldsymbol{N_j}$ is the orthogonal projector on a reference subspace $\boldsymbol{S_j}$ and all ω_j's are equal, this PCA is the generalized canonical analysis (Kettenring, 1971) of the set of subspaces $\{\boldsymbol{S_j}, j = 1, \ldots, J\}$.

- We now consider the case where $l \to \infty$. We consider the quantity

$$\|\boldsymbol{N_j}\| := \sup_{\boldsymbol{u}, \boldsymbol{u^T M^{-1} u} = 1} \boldsymbol{u^T N_j u}.$$

  Then

  - If there exists $j^\star$ such that, for all $j \neq j^\star$, $\|\boldsymbol{N_j}\| < \|\boldsymbol{N_{j^\star}}\|$, we have

$$\operatorname*{argmax}_{\boldsymbol{u}, \boldsymbol{u^T M^{-1} u} = 1} \phi(\boldsymbol{u}) = \operatorname*{argmax}_{\boldsymbol{u}, \boldsymbol{u^T M^{-1} u} = 1} \boldsymbol{u^T N_{j^\star} u}.$$

    In this case, the loading vector $\boldsymbol{u}$ will be drawn to the subspace $\boldsymbol{S_{j^\star}}$ associated with $\boldsymbol{N_{j^\star}}$.

  - If there exists λ such that, for all $j = 1, \ldots, J$, $\|\boldsymbol{N_j}\| = \lambda$, then the first eigenvector (associated with the maximum eigenvalue) of the $\boldsymbol{N_j}$ having maximum weight ω_j maximizes $\phi(\boldsymbol{u})$. If all ω's are equal, the first eigenvectors of all $\boldsymbol{N_j}$'s maximize it. It follows that in the particular case where $\boldsymbol{N_j} = \boldsymbol{\Pi}_{S_j}$ and all ω's are equal, any vector $\boldsymbol{u}$ belonging to any $\boldsymbol{S_j}$ maximizes $\phi(\boldsymbol{u})$. So, $\boldsymbol{u}$ will be drawn to the $\boldsymbol{S_j}$ closest to it.

- As previously mentioned, taking $l = 1$ draws the components towards the principal components, while increasing l infinitely leads the component to stick to the closest explanatory variable. Taking $1 < l < \infty$ will make $\boldsymbol{u}$ focus on close local bundles of $\boldsymbol{S_j}$'s. The higher l, the more local the bundle. Hence, l may be considered as a bundle locality parameter.

In this thesis, we present three particular examples of SR worth mentioning.

- **Component Variance**
  Let $\boldsymbol{X}$ being composed of centered numerical variables. We want to find a direction span$[\boldsymbol{u}]$ capturing the highest possible inertia of the observations. We take $\boldsymbol{N} = \{\boldsymbol{X^T W X}\}$, $\boldsymbol{\Omega} = \{1\}$ and $l = 1$. Thus, the SR measure writes

$$\phi(\boldsymbol{u}) = \boldsymbol{u^T X^T W X u} = \|\boldsymbol{X u}\|^2_{\boldsymbol{W}} = \mathbb{V}[\boldsymbol{X u}].$$

  With the constraint $\boldsymbol{u^T M^{-1} u} = 1$, we recognize the maximized criterion of the PCA of $\boldsymbol{X}$ with metric $\boldsymbol{M}$ and weights matrix $\boldsymbol{W}$. However, in practice, explanatory variables are most often a mixture of numerical and nominal variables (see for instance Escofier and Pagès (1984, 1998); Pagès (2021)). We consider

$$\boldsymbol{X} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_P}, \boldsymbol{X_1}, \ldots, \boldsymbol{X_Q}],$$

  where $\boldsymbol{x_1}, \ldots, \boldsymbol{x_P}$ are numerical variables, and $\boldsymbol{X_1}, \ldots, \boldsymbol{X_Q}$ are blocks of centered indicator variables, each block coding a categorical variable ($\boldsymbol{X_q}$ has $L_q - 1$ columns if the corresponding variable has L_q levels, the removed level being taken as reference level). In order to get a relevant PCA of $(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{W})$, we must consider the metric block-diagonal matrix

$$\boldsymbol{M^{-1}} = \text{diag}\left(\boldsymbol{x_1^T W x_1}, \ldots, \boldsymbol{x_P^T W x_P}, \boldsymbol{X_1^T W X_1}, \ldots, \boldsymbol{X_Q^T W X_Q}\right).$$

  This matrix bridges ordinary PCA of numerical variables with Multiple Correspondence Analysis (Greenacre and Blasius, 2006).

- **Block's variance captured by component**
  We assume that $\boldsymbol{X}$ consists of P standardized numerical variables. In this case, we take $\boldsymbol{N} = \{\left(\boldsymbol{X^T W X}\right)^2\}$, $\boldsymbol{\Omega} = \{1\}$ and $l = 1$. We consider

$$\begin{aligned}
\phi(\boldsymbol{u}) &= \boldsymbol{u^T}\left(\boldsymbol{X^T W X}\right)^2 \boldsymbol{u} \\
&= \boldsymbol{u^T}\left(\boldsymbol{X^T W X X^T W X}\right) \boldsymbol{u} \\
&= \sum_{p=1}^{P} \boldsymbol{f^T W x_p x_p^T W f} \\
&= \sum_{p=1}^{P} \langle \boldsymbol{f}, \boldsymbol{x_p} \rangle^2_{\boldsymbol{W}}.
\end{aligned}$$

If we impose that $\|\boldsymbol{f}\|_{\boldsymbol{W}}^2 = 1$, $\phi(\boldsymbol{u})$ is the inertia of the variables along span$[\boldsymbol{X}\boldsymbol{u}]$, it is maximized by the first dual eigenvector of the PCA of $(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{W})$. But the constraint $\|\boldsymbol{f}\|_{\boldsymbol{W}}^2 = 1$ amounts to taking $\boldsymbol{M^{-1}} = \boldsymbol{X^T W X}$, which is a problem when $\boldsymbol{X^T W X}$ is not regular. Therefore, this choice of $\boldsymbol{M}$ will generally be discarded in favor of a regularizing $\boldsymbol{M}$ (see below).

- **Variable Powered Inertia**
  The previous criterion can be extended to something called Variable Powered Inertia (VPI). Taking $J = P$, $\boldsymbol{N_p} = \boldsymbol{X^T W x_p x_p^T W X}$ and $\omega_p = 1/P$ for all $p = 1, \ldots, P$, and $l \geq 1$, the VPI writes

$$\begin{aligned}\phi(\boldsymbol{u}) &= \left(\sum_{p=1}^{P} \omega_p \left(\boldsymbol{u^T X^T W x_p x_p^T W X u}\right)^l\right)^{1/l} \\ &= \left(\frac{1}{P}\sum_{p=1}^{P} \langle \boldsymbol{X}\boldsymbol{u}, \boldsymbol{x_p}\rangle_{\boldsymbol{W}}^{2l}\right)^{1/l}. \end{aligned} \tag{2.9}$$

  One can see how the value of l tunes the locality of bundles considered

  - If $l = 1$ and for all j, $\omega_j = 1$, the VPI gives back the previous block-variance criterion.
  - If $l = 2$ and for all j, $\omega_j = 1$, the VPI yields a varimax criterion initially introduced by Kaiser (1958).

  For a block $\boldsymbol{X}$ consisting of P categorical variables $\boldsymbol{X_p}$, each of which being coded through the set of its centered indicator variables, we will take

$$\phi(\boldsymbol{u}) = \left(\frac{1}{P}\sum_{p=1}^{P} \left\langle \boldsymbol{X}\boldsymbol{u}, \boldsymbol{\Pi}_{\boldsymbol{X_p}}\boldsymbol{X}\boldsymbol{u}\right\rangle_{\boldsymbol{W}}^{l}\right)^{1/l}.$$

  The locality of a bundle of correlated variables is defined by the within-bundle correlation: the higher the correlation, the more local the bundle. The main objective is to focus on the most interpretable directions. On Figure 2.1, we plot the VPI in the particular case of four coplanar variables.

How to choose $\boldsymbol{M}$

To achieve regularization, we shall present how to play on $\boldsymbol{M}$. For any of the above-mentioned SR measures, it is convenient to choose

$$\boldsymbol{M^{-1}} = \tau \boldsymbol{I_P} + (1-\tau)\boldsymbol{X^T W X},$$

where $\tau \in [0, 1]$ is a ridge-like tuning parameter (Hoerl and Kennard, 1970a,b). For instance, if the variables in $\boldsymbol{X}$ are quantitative and standardized, $\boldsymbol{M^{-1}} = \boldsymbol{I_P}$, whereas

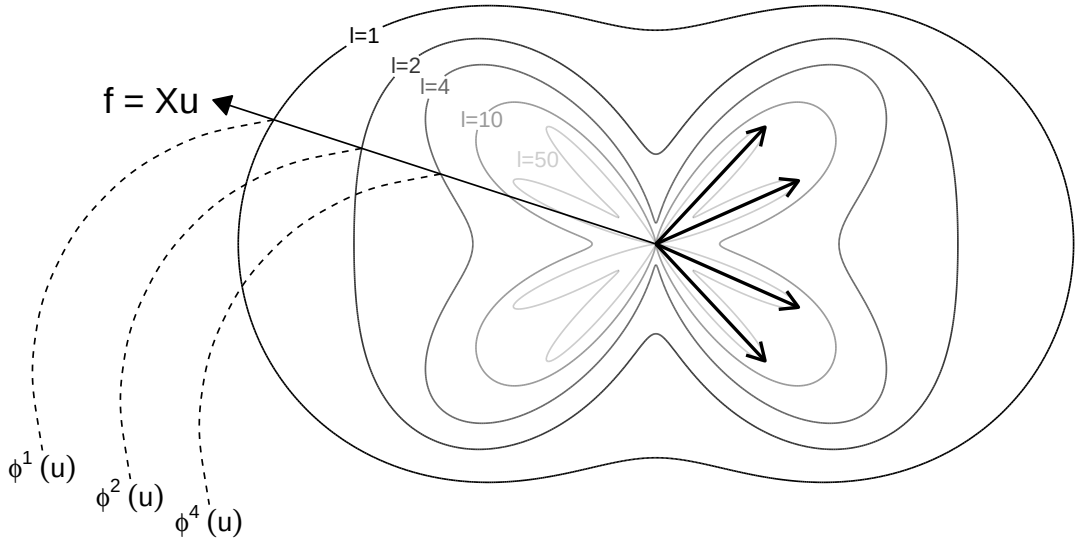


Figure 2.1: Polar representation of the VPI. Vector $\boldsymbol{u}$ is depicted by the complex number $z = e^{i\theta}$ where $\theta \in [0, 2\pi[$. Component $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$ cuts the curve $z_l(\theta) = \left(\phi(e^{i\theta})\right)^l e^{i\theta}$ at a point of radius equal to $\phi(\boldsymbol{u})^l$. Curves z_l are graphed for $l \in \{1, 2, 4, 10, 50\}$.

if they are categorical, $\boldsymbol{M}^{-1}$ will be the metric of the Multiple Correspondance Analysis. This convex combination extends that proposed by Tenenhaus and Tenenhaus (2011) to tune regularization. A main advantage of this constraint is to overcome the limitation of the singularity of the matrix $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$. For more details on the influence of the hyper-parameter τ in the SCGLR context, we refer the reader to Bry et al. (2020a).

2.2.3.4 The SCGLR combined criterion

The SCGLR combined criterion, proposed by Bry et al. (2020b), introduced a hyper-parameter $s \in [0, 1]$ to tune the importance of the SR relative to the GoF. SCGLR attempts a trade-off between ϕ and $\psi_{\boldsymbol{A}}$ by solving

$$\begin{aligned} &\max_{\boldsymbol{u}, \boldsymbol{u}^T\boldsymbol{M}^{-1}\boldsymbol{u}=1} \phi(\boldsymbol{u})^s \psi_{\boldsymbol{A}}(\boldsymbol{u})^{1-s} \\ \Leftrightarrow &\max_{\boldsymbol{u}, \boldsymbol{u}^T\boldsymbol{M}^{-1}\boldsymbol{u}=1} s\ln(\phi(\boldsymbol{u})) + (1-s)\ln(\psi_{\boldsymbol{A}}(\boldsymbol{u})). \end{aligned}$$

When $s = 0$, the criterion identifies with the GoF, while at the other end, taking $s = 1$ makes it identify with the SR. Thus, increasing s intensifies both the focus of components on "strong" dimensions, and the regularization. This role is similar to that of the penalty coefficient in penalty-based methods such as the ridge regression, the least absolute shrinkage and selection operator, or the elastic net (Hoerl and Kennard, 1970a,b; Tibshirani, 1996; Zou and Hastie, 2005).

This compound criterion is quite general. Indeed, the GoF measure adapts any situation where a likelihood function is available for the model taking the components and $\boldsymbol{A}$ as covariates. Generally, this likelihood involves a vector of parameters. The maximization is carried out alternating two steps:

(i) Given $\boldsymbol{u}$, maximize the criterion with respect to the parameter vector. This step is performed using a classical likelihood maximization algorithm relevant to the situation.

(ii) Given the parameter vector, maximize the criterion with respect to $\boldsymbol{u}$ using a dedicated algorithm: PING (for Projected Iterated Normed Gradient) recalled in Section 2.2.3.5.

We hereafter give Algorithm 2 corresponding to the SCGLR method.

Algorithm 2: The SCGLR algorithm

while *not convergence* **do**

Finding the component with the PING algorithm

$$\boldsymbol{u}^{(t+1)} = \underset{\boldsymbol{u},\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{M}^{-\mathbf{1}}\boldsymbol{u}=1}{\operatorname{argmax}} \phi\left(\boldsymbol{u}^{(t)}\right)^{s} \psi_{\boldsymbol{A}}\left(\boldsymbol{u}^{(t)}\right)^{1-s}$$

$$\boldsymbol{f}^{(t+1)} = \boldsymbol{X}\boldsymbol{u}^{(t+1)}$$

$$\boldsymbol{T}^{(t+1)} = \left[\boldsymbol{f}^{(t+1)}, \boldsymbol{A}\right]$$

Computing the model parameters with the IRLS algorithm

for $k = 1, \ldots, K$ **do**

$$(\boldsymbol{\gamma}_{\boldsymbol{k}}^{(t+1)}, \boldsymbol{\delta}_{\boldsymbol{k}}^{(t+1)\boldsymbol{T}})^{T} = \left(\boldsymbol{T}^{(t+1)\,T}\boldsymbol{W}_{\boldsymbol{k}}^{(t)}\boldsymbol{T}^{(t+1)}\right)^{-1}\boldsymbol{T}^{(t+1)\,T}\boldsymbol{W}_{\boldsymbol{k}}^{(t)}\boldsymbol{w}_{\boldsymbol{k}}^{(t)}$$

$$\boldsymbol{\eta}_{\boldsymbol{k}}^{(t+1)} = \boldsymbol{f}^{(t+1)}\boldsymbol{\gamma}_{\boldsymbol{k}}^{(t+1)} + \boldsymbol{A}\boldsymbol{\delta}_{\boldsymbol{k}}^{(t+1)}$$

$$\mu_{nk}^{(t+1)} = h^{-1}\left(\eta_{nk}^{(t+1)}\right), \forall n = 1, \ldots, N$$

$$w_{nk}^{(t+1)} = \eta_{nk}^{(t+1)} + h_k'\left(\mu_{nk}^{(t+1)}\right)\left(y_{nk} - \mu_{nk}^{(t+1)}\right), \forall n = 1, \ldots, N$$

$$\boldsymbol{W}_{k}^{(t+1)} = \operatorname{diag}\left(\left[a_{nk}(\phi_k)v_k\left(\mu_{nk}^{(t+1)}\right)h_k'\left(\mu_{nk}^{(t+1)}\right)^2\right]^{-1}\right)_{n=1,\ldots,N}$$

end

$t \leftarrow t + 1$

end

2.2.3.5 Brief reminder of the PING algorithm

PING is an algorithm designed to maximize, at least locally, any criterion $C(\boldsymbol{v})$ on the unit sphere (Chauvet et al., 2019; Bry et al., 2020a,b). The key idea is to stay close enough to a current starting point $\boldsymbol{v}^{(t)}$ by maximizing the criterion through a Gauss-Newton unidimensional procedure on an arc previously chosen. The new iteration $\boldsymbol{v}^{(t+1)}$ is defined as the maximum of $C(\boldsymbol{v})$ on this arc. The fixed point of the resulting algorithm is a local maximum of the criterion. Further details and developments are given in Supplementary Material.

2.2.4 Extension to a partitioning of explanatory variables

In this section we assume that the explanatory variables are partitioned into R conceptually homogenous thematic variable groups $\boldsymbol{X} = [\boldsymbol{X_1}, \ldots, \boldsymbol{X_R}]$, viewed as explanatory themes. The search for components in multi-block analysis is an ongoing statistical research field, in which several approaches have been proposed, e.g. Multi-Block PLS (MB-PLS, Wangen and Kowalski, 1989), the PCA performed separately on each block (Westerhuis et al., 1998) or the PLS Path Modeling (PLS-PM, Tenenhaus et al., 2005). The term "theme" was first introduced to develop the Structural Equation Exploratory Regression for THEmatic models with Multiple Equations (THEME-SEER, Bry et al., 2009, 2012). We might also cite the work of Bougeard et al. (2018) who propose to rewrite the regularized Generalized Structured Component Analysis (rGSCA, Hwang and Takane, 2004; Hwang, 2009), the regularized Generalized Canonical Correlation Analysis (rGCCA, Tenenhaus and Tenenhaus, 2011) and the THEmatic Equation Model Exploration (THEME, Bry and Verron, 2015) in an unified framework. Finally, Bry et al. (2020b) introduce THEME-SCGLR in order to extend SCGLR so as to deal with a partitioning of the explanatory variables.

2.2.4.1 THEME-SCGLR's components

The conceptual model stating that variables in $\boldsymbol{Y}$ are dependent on R themes $\boldsymbol{X_1}, \ldots, \boldsymbol{X_R}$ plus a set of covariates $\boldsymbol{A}$, and that structurally relevant dimensions should be explicitly identified in the $\boldsymbol{X}_r$'s, will be referred to as "thematic model". For the sake of simplicity, let us momentarily restrict ourselves to the model having a single component per theme. The linear predictor for response $\boldsymbol{y_k}$ writes

$$\boldsymbol{\eta_k} = (\boldsymbol{X_1 u_1})\,\gamma_{k1} + \cdots + (\boldsymbol{X_R u_R})\,\gamma_{kR} + \boldsymbol{A\delta_k}.$$

The main objective of THEME-SCGLR is to track down structurally relevant dimensions in spaces span$[\boldsymbol{X}_r]$, that can ground a good explanatory and predictive model of $\boldsymbol{Y}$. To achieve theme-specific regularization, the SCGLR criterion has to be adapted. Denoting $\boldsymbol{f}_r = \boldsymbol{X}_r\boldsymbol{u}_r$ the (only) component of theme $\boldsymbol{X}_r$, we have $\boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{f_1},\ldots,\boldsymbol{f_R},\boldsymbol{A}]} = \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{f}_r,\boldsymbol{A}_r]}$ where $\boldsymbol{A}_r = [\boldsymbol{f_1}, \ldots, \boldsymbol{f}_{r-1}, \boldsymbol{f}_{r+1}, \ldots, \boldsymbol{f_R}, \boldsymbol{A}]$. For each theme, the GoF measure thus

becomes

$$\psi_{\boldsymbol{A_r}}\left(\boldsymbol{u_r}\right) := \sum_{k=1}^{K} \alpha_k \left\|\boldsymbol{w_k}\right\|_{\boldsymbol{W_k}}^2 \cos_{\boldsymbol{W_k}}^2 \left(\boldsymbol{w_k}, \boldsymbol{\Pi}_{\text{span}[\boldsymbol{X_r u_r}, \boldsymbol{A_r}]}^{\boldsymbol{W_k}} \boldsymbol{w_k}\right). \tag{2.10}$$

The SR measure remains the same $\phi\left(\boldsymbol{u_r}\right)$ as given by Equation (2.9). Finally, the optimization program can be solved by iteratively maximizing in turn the trade-off criterion on every $\boldsymbol{u_r}$

$$\forall r, \quad \max_{\boldsymbol{u_r}, \boldsymbol{u_r^T M^{-1} u_r}=1} s \ln(\phi(\boldsymbol{u_r})) + (1-s) \ln(\psi_{\boldsymbol{A_r}}(\boldsymbol{u_r})). \tag{2.11}$$

2.3 Finite Mixture Models

Finite mixtures of distributions have provided a mathematical approach to the statistical modeling of a wide variety of non homogeneous random phenomena. Because of their extremely flexibility, Finite Mixture Models (FMMs) have continued to receive great attention, from both a practical and a theoretical point of view (see McLachlan and Peel (2004) for a reference book). Indeed, the extent and the application potential of FMMs have widened considerably. Many fields in which FMMs have been applied can be cited, for instance: astronomy (Lee et al., 2012), ecology (Pledger and Phillpot, 2008), psychology (Colder et al., 2002) or sociology (Jones et al., 2001). The FMMs provide a convenient parametric framework in which the objective is to model an unknown probability distribution function (pdf) of a random variable by a finite sum of distributions.

Let $\boldsymbol{y} = (y_1, \ldots, y_N)^T \in \mathbb{R}^N$ be the observed sample vector. The observations are assumed to be modeled through a FMM. The estimation problem of the FMM consists in finding a good approximation of the pdf L, modeled as a finite sum of G densities with a sample $\boldsymbol{y}$ of independent realizations of a random variable

$$L(y_n; \boldsymbol{\Theta}) = \sum_{g=1}^{G} p_g d_g(y_n; \theta_g), \tag{2.12}$$

where d_g is a density, θ_g the parameter of this density and p_g the gth *a priori* mixing probability. The probability conditions are respected with the constraints $p_g \in [0, 1]$ and $\sum_{g=1}^{G} p_g = 1$. $\boldsymbol{\Theta}$ denotes the vector of the model parameters, composed of the parameters θ_g's and p_g's.

2.3.1 The Gaussian mixture

Originally, the FMMs were developed in order to model a linear combination of Gaussian distributions with different means and variances. Indeed, if the data structure is composed of several groups, a single Gaussian distribution cannot capture the entire information. Informally, in the case of multi-modal distribution, a unique Gaussian density does not

allow to fit the curve. As a result, the Gaussian mixture can be seen as an unsupervised classification method (Biernacki et al., 2000; Hunt and Jorgensen, 2003; El Attar, 2012).

In a Gaussian mixture framework with G groups, each having mean μ_g and variance σ_g^2, the pdf writes

$$L(y_n; \boldsymbol{\Theta}) = \sum_{g=1}^{G} \frac{p_g}{\sqrt{2\pi\sigma_g^2}} \exp\left(-\frac{1}{2}\frac{(y_n - \mu_g)^2}{\sigma_g^2}\right),$$

where the set of parameters is $\boldsymbol{\Theta} = \{p_1, \ldots, p_G, \mu_1, \ldots, \mu_G, \sigma_1^2, \ldots, \sigma_G^2\}$. With this in mind, we ought to estimate the parameters of the Gaussian mixture model. To achieve that, we want to maximize the log-likelihood. The observations being independent, the model log-likelihood is given by

$$l(\boldsymbol{\Theta}, \boldsymbol{y}) = \ln\left(L(\boldsymbol{y}; \boldsymbol{\Theta})\right) = \sum_{n=1}^{N} \ln\left(L(y_n; \boldsymbol{\Theta})\right).$$

Originally, the estimation of the mixture parameters was performed by maximum likelihood as proposed by Day (1969) and Wolfe (1970). However, the log-likelihood being difficult to maximize directly, and the group memberships of the observations being unknown, we shall adopt the Expectation Maximization (EM, Dempster et al., 1977) algorithm to estimate the model parameters. In the next section, we give a brief recall on the EM algorithm.

2.3.2 The standard EM algorithm

Initially introduced by Dempster et al. (1977), the EM algorithm aims at finding a local maximum of the likelihood in a context of a random latent variable model. Indeed, in this framework, the (log-)likelihood may be difficult to maximize or may not have an analytic expression. Since the reference paper, many works have described the properties of the EM algorithm, for instance we shall cite Lauritzen (1995); Lin (2010) for missing data, Redner and Walker (1984); Muthén and Shedden (1999) for the mixture model or Rubin and Thayer (1982, 1983) for the factor analysis. For a comprehensive overview on the EM algorithm, we refer the reader to the nice informal tutorial by Chauvet (2019).

Let y be a random variable and $L(y; \theta)$ its probability distribution function, where θ is an unknown parameter. The objective is to maximize the log-likelihood $l(\theta; y) = \ln(L(y; \theta))$ to estimate θ. However, due to the presence of the random latent variable $z \in \Omega_z$, this log-likelihood is too complicated, if not impossible, to maximize. We need to construct a sequence of parameters $\left(\theta^{(t)}\right)_t$ such that the log-likelihood $l\left(\theta^{(t)}; y\right)$ increases

with t. To achieve that, we calculate

$$\begin{aligned} l\left(\theta;y\right) - l\left(\theta^{(t)};y\right) &= \ln\left(L\left(y;\theta\right)\right) - \ln\left(L\left(y;\theta^{(t)}\right)\right) \\ &= \ln\left(\frac{L\left(y;\theta\right)}{L\left(y;\theta^{(t)}\right)}\right) \\ &= \ln\left(\int_{\Omega_Z} \frac{L\left(y,z;\theta\right)}{L\left(y;\theta^{(t)}\right)} dz\right) \\ &= \ln\left(\int_{\Omega_Z} \frac{L\left(y,z;\theta\right)}{L\left(y,z;\theta^{(t)}\right)} L\left(z|y;\theta^{(t)}\right) dz\right) \\ &= \ln\left(\mathbb{E}\left[\left.\frac{L\left(y,z;\theta\right)}{L\left(y,z;\theta^{(t)}\right)}\right| y;\theta^{(t)}\right]\right). \end{aligned}$$

Thanks to the concavity of the logarithm function, we shall use the Jensen's inequality (Cover and Thomas, 2006, pages 25–30)

$$\ln\left(\mathbb{E}\left[\left.\frac{L\left(y,z;\theta\right)}{L\left(y,z;\theta^{(t)}\right)}\right| y;\theta^{(t)}\right]\right) \geq \mathbb{E}\left[\left.\ln\left(\frac{L\left(y,z;\theta\right)}{L\left(y,z;\theta^{(t)}\right)}\right)\right| y;\theta^{(t)}\right].$$

We thus have

$$\begin{aligned} l\left(\theta;y\right) - l\left(\theta^{(t)};y\right) &\geq \int_{\Omega_Z} \ln\left(\frac{L\left(y,z;\theta\right)}{L\left(y,z;\theta^{(t)}\right)}\right) L\left(z|y;\theta^{(t)}\right) dz \\ &:= P\left(\theta,\theta^{(t)}\right) \end{aligned}$$

where the auxiliary function P defined by $P(\theta,\theta^{(t)})$ is seen as a lower bound for the deviation of the log-likelihood from $\theta^{(t)}$ to θ. Henceforth, we need to find $\theta^{(t+1)}$ such that $P(\theta^{(t+1)},\theta^{(t)}) \geq 0$. With this aim in mind, we chose $\theta^{(t+1)}$ as the solution of the maximization of $P(\theta,\theta^{(t)})$ with respect to θ. Then, by definition, we obtain

$$P(\theta^{(t+1)},\theta^{(t)}) \geq P(\theta^{(t)},\theta^{(t)}).$$

Moreover, $P(\theta^{(t)},\theta^{(t)}) = 0$. We thus demonstrate the increase of the log-likelihood with t

$$l\left(\theta^{(t+1)};y\right) - l\left(\theta^{(t)};y\right) \geq P\left(\theta^{(t+1)},\theta^{(t)}\right) \geq 0.$$

The logarithm allows to rewrite the auxiliary function as a difference

$$P\left(\theta,\theta^{(t)}\right) = Q\left(\theta,\theta^{(t)}\right) - Q\left(\theta^{(t)},\theta^{(t)}\right),$$

where

$$Q\left(\theta,\theta^{(t)}\right) = \int_{\Omega_Z} \ln\left(L\left(y,z;\theta\right)\right) L\left(z|y;\theta^{(t)}\right) dz.$$

Since $Q\left(\theta^{(t)}, \theta^{(t)}\right)$ does not depend on θ, maximizing $P\left(\theta, \theta^{(t)}\right)$ with respect to θ is then equivalent to maximizing $Q\left(\theta, \theta^{(t)}\right)$ which can be expressed as the conditional expectation of the complete log-likelihood $l(\theta; y, z)$ conditional on y at the current value $\theta^{(t)}$. Algorithm 3 summarizes the most widespread form of the EM algorithm.

Algorithm 3: The EM algorithm

while *not convergence* **do**

Step E : $Q(\theta, \theta^{(t)}) = \mathbb{E}\left[l(\theta; y, z)|y; \theta^{(t)}\right]$

Step M : $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\ Q\left(\theta, \theta^{(t)}\right)$

$t \leftarrow t + 1$

end

However, we may note that there does not exist any *a priori* on the number of iterations needed. Furthermore, the convergence is ensured to a local maximum of the likelihood. For more discussions on the EM algorithm, see those provided after the paper of Dempster et al. (1977, pages 22–38).

2.3.3 EM algorithm for a Gaussian mixture model

Let z_{ng} be the latent indicator variable equal to 1 if the observation y_n belongs to the gth group, and 0 otherwise. Let $\boldsymbol{z_n} = (z_{ng}\,;\, g = 1, \ldots, G)$ be the indicator vector of group membership of observation y_n, and let $\boldsymbol{Z} = \left[\boldsymbol{z_n^T}\,;\, n = 1, \ldots, N\right]$ be a $N \times G$ matrix. Conditional on $z_{ng} = 1$, the pdf for the observation y_n is d_g, the Gaussian distribution having mean μ_g and variance σ_g^2. The model complete log-likelihood writes

$$
\begin{aligned}
l(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{Z}) &= \ln\left(\prod_{n=1}^{N}\prod_{g=1}^{G}\left[\frac{p_g}{\sqrt{2\pi\sigma_g^2}}\exp\left(-\frac{1}{2}\frac{(y_n-\mu_g)^2}{\sigma_g^2}\right)\right]^{z_{ng}}\right) \\
&= \sum_{n=1}^{N}\sum_{g=1}^{G} z_{ng}\ln\left(\frac{p_g}{\sqrt{2\pi\sigma_g^2}}\exp\left(-\frac{1}{2}\frac{(y_n-\mu_g)^2}{\sigma_g^2}\right)\right).
\end{aligned}
$$

The M step of the EM algorithm consists in maximizing with respect to $\boldsymbol{\Theta}$ the conditional expectation of the complete log-likelihood $\mathbb{E}\left[l(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{Z})|\boldsymbol{y}; \boldsymbol{\Theta'}\right]$. The solution is then injected into $\boldsymbol{\Theta'}$, and the conditional expectation is updated in the E step.

2.3.3.1 The expectation (E) step

With $\boldsymbol{z_n} = g$ meaning that the gth coordinate of the vector $\boldsymbol{z_n}$ equals 1, the conditional expectation of the complete log-likelihood writes

$$\begin{aligned}
\mathbb{E}\left[l(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{Z})|\boldsymbol{y}; \boldsymbol{\Theta'}\right] &= \sum_{n=1}^{N} \mathbb{E}\left[\ln\left(L(y_n, \boldsymbol{z_n}; \theta)\right)|y_n; \theta'\right] \\
&= \sum_{n=1}^{N}\sum_{g=1}^{G} \mathbb{P}(\boldsymbol{z_n} = g|y_n; \theta'_g) \ln\left(L(y_n, \boldsymbol{z_n} = g; \theta_g)\right) \\
&= \sum_{n=1}^{N}\sum_{g=1}^{G} \mathbb{P}(\boldsymbol{z_n} = g|y_n; \theta'_g) \ln\left(\mathbb{P}(\boldsymbol{z_n} = g; \theta_g) L(y_n|\boldsymbol{z_n} = g; \theta_g)\right) \\
&= \sum_{n=1}^{N}\sum_{g=1}^{G} \alpha_{ng} \ln\left(p_g d_g\left(y_n; \theta_g\right)\right) \\
&= \sum_{n=1}^{N}\sum_{g=1}^{G} \alpha_{ng} \left[\ln\left(p_g\right) - \frac{1}{2}\ln\left(2\pi\sigma_g^2\right) - \frac{1}{2}\frac{\left(y_n - \mu_g\right)^2}{\sigma_g^2}\right],
\end{aligned}$$

where the posterior group membership probabilities of each observation y_n are computed as

$$\alpha_{ng} := \mathbb{P}(\boldsymbol{z_n} = g|y_n; \theta_g) = \frac{L(y_n, \boldsymbol{z_n} = g; \theta_g)}{L(y_n; \theta_g)} = \frac{p_g d_g(y_n; \theta_g)}{\sum_{r=1}^{G} p_r d_r(y_n; \theta_r)}.$$

2.3.3.2 The maximization (M) step

The M-step maximizes with respect to $\boldsymbol{\Theta}$ the conditional expectation of the complete log-likelihood, subject to the constraint $\sum_{g=1}^{G} p_g = 1$. We maximize the corresponding Lagrangian

$$\mathcal{L}(\boldsymbol{\Theta}, \lambda) = \mathbb{E}\left[l(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{Z})|\boldsymbol{y}; \boldsymbol{\Theta'}\right] - \lambda\left(\sum_{g=1}^{G} p_g - 1\right).$$

The maximization with respect to p_g yields

$$\begin{aligned}
\nabla_{p_g}\mathcal{L}(\boldsymbol{\Theta}, \lambda) = 0 &\Leftrightarrow \sum_{n=1}^{N} \alpha_{ng} = p_g\lambda \\
&\Leftrightarrow \sum_{n=1}^{N} \underbrace{\sum_{g=1}^{G} \alpha_{ng}}_{=1} = \lambda \underbrace{\sum_{g=1}^{G} p_g}_{=1} \\
&\Leftrightarrow N = \lambda.
\end{aligned}$$

So, the solution p_g is

$$\hat{p}_g = \frac{1}{N}\sum_{n=1}^{N} \alpha_{ng}.$$

The estimates $\hat{\mu}_g$ and $\hat{\sigma}_g^2$ are obtained as the solutions of

$$\begin{aligned}\nabla_{\mu_g}\mathcal{L}(\boldsymbol{\Theta},\lambda)=0 &\Leftrightarrow \nabla_{\mu_g}\left\{\sum_{n=1}^{N}\alpha_{ng}\left(y_n-\mu_g\right)^2\right\}=0\\ &\Leftrightarrow \sum_{n=1}^{N}\alpha_{ng}\left(y_n-\mu_g\right)=0\\ &\Leftrightarrow \sum_{n=1}^{N}\alpha_{ng}y_n=\mu_g\sum_{n=1}^{N}\alpha_{ng}\\ &\Leftrightarrow \mu_g=\frac{\sum_{n=1}^{N}\alpha_{ng}y_n}{\sum_{n=1}^{N}\alpha_{ng}}\end{aligned}$$

and

$$\begin{aligned}\nabla_{\sigma_g^2}\mathcal{L}(\boldsymbol{\Theta},\lambda)=0 &\Leftrightarrow \nabla_{\sigma_g^2}\left\{\sum_{n=1}^{N}\alpha_{ng}\left(\ln\left(\sigma_g^2\right)+\frac{\left(y_n-\mu_g\right)^2}{\sigma_g^2}\right)\right\}=0\\ &\Leftrightarrow \sum_{n=1}^{N}\alpha_{ng}\left(\frac{1}{\sigma_g^2}-\frac{\left(y_n-\mu_g\right)^2}{\left(\sigma_g^2\right)^2}\right)=0\\ &\Leftrightarrow \sum_{n=1}^{N}\alpha_{ng}=\frac{1}{\sigma_g^2}\sum_{n=1}^{N}\alpha_{ng}\left(y_n-\mu_g\right)^2\\ &\Leftrightarrow \sigma_g^2=\frac{\sum_{n=1}^{N}\alpha_{ng}\left(y_n-\mu_g\right)^2}{\sum_{n=1}^{N}\alpha_{ng}}.\end{aligned}$$

2.3.3.3 The FMM estimation algorithm

As a result of the aforementioned developments, we shall use Algorithm 4 to estimate the parameters of the Gaussian mixture model.

2.3.4 Extension to a response mixture model

In a context of multiple and numerous response variables, we have to cluster the outcomes instead of the statistical units as in the original and classical FMM approach. The interest of response clustering has already been shown in several works, e.g. those of Monni and Tadesse (2009); Ovaskainen and Soininen (2011); Pledger and Arnold (2014); Mortier et al. (2015) and Hill et al. (2020). In this thesis, we opt for the modeling approach proposed by Dunstan et al. (2011, 2013). The authors propose the Species Archetype Model (SAM) which supposes that all responses (species) can be clustered into a small number of groups with respect to their responses to environmental gradients. In their model, the responses within a group share the same regression parameters with only an intercept specific to each response. More formally, let $\boldsymbol{Y}\in\mathbb{R}^{N\times K}$ be the multivariate response matrix,

Algorithm 4: The EM algorithm applied to Gaussian mixture

while *not convergence* **do**

Expectation step

for $n = 1, \ldots, N$ **do**

for $g = 1, \ldots, G$ **do**

$$\alpha_{ng}^{(t+1)} = \frac{p_g^{(t)} d_g(y_n; \theta_g^{(t)})}{\sum_{r=1}^{G} p_r^{(t)} d_r(y_n; \theta_r^{(t)})}$$

end

end

Maximization step

for $g = 1, \ldots, G$ **do**

$$p_g^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} \alpha_{ng}^{(t+1)}$$

$$\mu_g^{(t+1)} = \frac{\sum_{n=1}^{N} \alpha_{ng}^{(t+1)} y_n}{\sum_{n=1}^{N} \alpha_{ng}^{(t+1)}}$$

$$\sigma_g^{2\,(t+1)} = \frac{\sum_{n=1}^{N} \alpha_{ng}^{(t+1)} \left(y_n - \mu_g^{(t+1)}\right)^2}{\sum_{n=1}^{N} \alpha_{ng}^{(t+1)}}$$

end

$t \leftarrow t + 1$

end

At the end, we can classify the observations according to their posterior probabilities. An observation y_n is assigned to cluster g if

$$\alpha_{ng}^{(t_{\max})} > \alpha_{nr}^{(t_{\max})}, \forall r \neq g.$$

we call $\boldsymbol{y_k}$ the kth response variable and y_{nk} its nth observation. The model writes

$$L\left(\boldsymbol{y_k};\theta_k\right)=\sum_{g=1}^{G}p_g\prod_{n=1}^{N}d_k(y_{nk},\mu_{nkg}),$$

where d_k is a distribution belonging to the exponential family, with expectation μ_{nkg}. The kth canonical link function h_k is defined by

$$h_k(\mu_{nkg})=\beta_{0k}+\boldsymbol{x_n^T}\boldsymbol{\beta_g},$$

where $\boldsymbol{x_n}$ is the nth row of the explanatory matrix, β_{0k} the specific intercept of response $\boldsymbol{y_k}$ and $\boldsymbol{\beta_g}$ the regression parameter vector of the gth group common to all the statistical units.

2.4 The standard factor model

Standard factor models were introduced by Spearman (1904); Thomson (1916); Thurstone (1931) in a psychology framework. The main goal of these models is to find a reduced number of non observed variables (called factors) to synthesize the information enclosed in multivariate data. They are a way to sum up and model the dependency between observed variables. Ever since then, these methods have been largely developed and diversified, see for instance Bartholomew (1995); Saidane (2006); Meyer (2009) or Tami (2016) for good reviews and more examples of factor models developments. The factors are uncorrelated random latent variables summarizing a set of observed variables correlated to some extent. The observed variables are described as linear combination of factors plus a mean parameter (a.k.a. the intercept) and an error term. The factors not being directly observed, they must be predicted together with the model parameters estimation. The frequentist approaches developed in order to estimate the model regression parameters are either based on the maximum likelihood of the sample variance-covariance matrix (Jöreskog, 1967, 1969) or on the EM algorithm (Rubin and Thayer, 1982; Jamshidian, 1997). For identification purposes, all of the methods need to impose constraints on the parameters. We refer the reader to Bartholomew et al. (2011, Chapter 3) for an overview of the subject.

2.4.1 Writing the factor model

Let $\boldsymbol{Y}\in\mathbb{R}^{N\times K}$ be the matrix of the observed variables. We call y_{nk} the value of the kth variable for the nth observation. Hereafter, the vector indexed by n is composed of the nth row of the associated matrix, while the vector indexed by k is its kth column. Denoting $\boldsymbol{g_n^T}=(g_{n1},\ldots,g_{nJ})$ the vector comprising a realization of J factors for the nth statistical unit, the model for y_{nk} writes

$$y_{nk}=\mu_{nk}+\boldsymbol{g_n^T}\boldsymbol{b_k}+\varepsilon_{nk},$$

where μ_{nk} is the intercept, $\boldsymbol{b_k}$ the regression parameters (or "loadings") on the factors and ε_{nk} the error term. This writes matrix-wise

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{GB} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{N\times K}$ is the matrix of the intercepts, $\boldsymbol{G} \in \mathbb{R}^{N\times J}$ the matrix of factors, $\boldsymbol{B} \in \mathbb{R}^{J\times K}$ the matrix of loadings and $\boldsymbol{\varepsilon} \in \mathbb{R}^{N\times K}$ the matrix of errors. For all $n = 1, \ldots, N$, the model writes

$$\boldsymbol{y_n} = \boldsymbol{\mu_n} + \boldsymbol{B^T g_n} + \boldsymbol{\varepsilon_n},$$

where $\boldsymbol{y_n}$, $\boldsymbol{\mu_n}$, $\boldsymbol{g_n}$ and $\boldsymbol{\varepsilon_n}$ are the vectors composed of the nth rows of the matrices $\boldsymbol{Y}$, $\boldsymbol{\mu}$, $\boldsymbol{G}$ and $\boldsymbol{\varepsilon}$ respectively. The vectors of factors are assumed drawn from a multivariate normal distribution and independent across statistical units, that is $\boldsymbol{g_n} \sim \mathcal{N}_J(0, \boldsymbol{I_J})$. The latter are moreover supposed independent from the error measures drawn from $\boldsymbol{\varepsilon_n} \sim \mathcal{N}_K(0, \boldsymbol{\Psi})$, with $\boldsymbol{\Psi} = \text{diag}\,(\sigma_k^2)_{k=1,\ldots,K}$. As a result, we obtain

$$\boldsymbol{y_n} \sim \mathcal{N}_K(\boldsymbol{\mu_n}, \boldsymbol{B^T B} + \boldsymbol{\Psi}).$$

The model is thus constructed such that all the covariance between variables is accounted by the J factors. Besides

$$\boldsymbol{y_n}|\boldsymbol{g_n} \sim \mathcal{N}_K(\boldsymbol{\mu_n} + \boldsymbol{B^T g_n}, \boldsymbol{\Psi}).$$

Denoting $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{B}, \boldsymbol{\Psi}\}$ the set of parameters, the complete log-likelihood of the model writes

$$\begin{aligned}
l(\boldsymbol{\Theta}; \boldsymbol{Y}, \boldsymbol{G}) &= \ln\left(L(\boldsymbol{Y}, \boldsymbol{G}; \boldsymbol{\Theta})\right) \\
&= \sum_{n=1}^{N} \ln\left(L(\boldsymbol{y_n}|\boldsymbol{g_n}; \boldsymbol{\Theta})\right) + \ln\left(L(\boldsymbol{g_n}; \boldsymbol{\Theta})\right) \\
&= \sum_{n=1}^{N} \Big\{ -\ln\left((2\pi)^{K/2}\text{det}(\boldsymbol{\Psi})^{1/2}\right) - \ln\left((2\pi)^{J/2}\right) \\
&\qquad -\frac{1}{2}\left(\boldsymbol{y_n} - \boldsymbol{\mu_n} - \boldsymbol{B^T g_n}\right)^T \boldsymbol{\Psi^{-1}} \left(\boldsymbol{y_n} - \boldsymbol{\mu_n} - \boldsymbol{B^T g_n}\right) - \frac{1}{2}\boldsymbol{g_n^T g_n} \Big\} \\
&= -\frac{1}{2}\sum_{n=1}^{N} \Bigg\{ (K+J)\ln(2\pi) + \sum_{k=1}^{K} \ln(\sigma_k^2) + \boldsymbol{g_n^T g_n} \\
&\qquad + \sum_{k=1}^{K} \frac{1}{\sigma_k^2}\left(y_{nk} - \mu_{nk} - \boldsymbol{g_n^T b_k}\right)^2 \Bigg\}.
\end{aligned}$$

Thanks to the previous development, we can characterize the distribution of

$(\boldsymbol{y}_n^T, \boldsymbol{g}_n^T)^T$. We have

$$\begin{aligned}
\operatorname{cov}\left(\boldsymbol{g}_n, \boldsymbol{y}_n^T\right) &= \mathbb{E}\left[\boldsymbol{g}_n \boldsymbol{y}_n^T\right] - \mathbb{E}\left[\boldsymbol{g}_n\right] \mathbb{E}\left[\boldsymbol{y}_n^T\right] \\
&= \mathbb{E}\left[\boldsymbol{g}_n \left(\boldsymbol{\mu}_n + \boldsymbol{B}^T \boldsymbol{g}_n + \boldsymbol{\varepsilon}_n\right)^T\right] \\
&= \mathbb{E}\left[\boldsymbol{g}_n \boldsymbol{g}_n^T\right] \boldsymbol{B} \\
&= \mathbb{V}\left[\boldsymbol{g}_n\right] \boldsymbol{B} \\
&= \boldsymbol{B}. \qquad (2.13)
\end{aligned}$$

As a result, we obtain

$$\begin{pmatrix} \boldsymbol{y}_n \\ \boldsymbol{g}_n \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_n \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{B}^T \boldsymbol{B} + \boldsymbol{\Psi} & \boldsymbol{B}^T \\ \boldsymbol{B} & \boldsymbol{I}_J \end{pmatrix}\right). \qquad (2.14)$$

2.4.2 The identification constraints

The structure of the factor model is characterized by a large number of parameters to estimate. This characteristic involves identification problems (e.g. an infinity of solutions), so we need to introduce some constraints on the model.

2.4.2.1 The rotation constraint

For identification purposes, we need to constrain the matrix $\boldsymbol{B}$. As shown by Geweke and Zhou (1996), if $\boldsymbol{\Omega}$ is an orthogonal matrix, we can rewrite the model as

$$\begin{aligned}
\boldsymbol{y}_n &= \boldsymbol{\mu}_n + \boldsymbol{B}^T \boldsymbol{\Omega}^T \boldsymbol{\Omega} \boldsymbol{g}_n + \boldsymbol{\varepsilon}_n \\
&= \boldsymbol{\mu}_n + \boldsymbol{B}_0^T \boldsymbol{g}_{0n} + \boldsymbol{\varepsilon}_n,
\end{aligned}$$

where the new factors $\boldsymbol{g}_{0n} = \boldsymbol{\Omega} \boldsymbol{g}_n$ are a rotation of the original factors $\boldsymbol{g}_n$. The same moment conditions valid for the old factors are also valid for the rotated ones, that is

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{g}_{0n}\right] &= \boldsymbol{\Omega} \mathbb{E}[\boldsymbol{g}_n] = 0 \\
\mathbb{V}\left[\boldsymbol{g}_{0n}\right] &= \boldsymbol{\Omega} \mathbb{V}[\boldsymbol{g}_n] \boldsymbol{\Omega}^T = \boldsymbol{\Omega} \boldsymbol{\Omega}^T = \boldsymbol{I}_J.
\end{aligned}$$

Moreover, parameters are also rotated. The new parameters are linked to the old ones through $\boldsymbol{B}_0^T = \boldsymbol{B}^T \boldsymbol{\Omega}^T$. Because these new parameters and factors lead to the same distribution for the responses, they cannot be identified from the observed variables unless further restrictions are imposed.

Since $\boldsymbol{B}$ has rank J, we assume that the first J columns of $\boldsymbol{B}$ are independent. Let $\boldsymbol{B_1}$ be the $J \times J$ sub-matrix composed of the first J columns of $\boldsymbol{B}$ and $\boldsymbol{B_2}$ be the matrix composed by the last $K - J$ columns such that $\boldsymbol{B} = [\boldsymbol{B_1}, \boldsymbol{B_2}]$. By theorem A9.8 of Muirhead (1982, page 592), there exists a unique orthogonal matrix $\boldsymbol{\Omega_1}$ such that $\boldsymbol{\Omega_1 B_1}$ is

an upper triangular matrix with positive diagonal elements. Thus, to get a unique solution for $\boldsymbol{B}$, for $J < K$, we impose the form

$$\boldsymbol{B} = \begin{pmatrix} b_{11} & \dots & b_{1J} & b_{1,J+1} & \dots & b_{1K} \\ & \ddots & \vdots & \vdots & \ddots & \vdots \\ & & b_{JJ} & b_{J,J+1} & \dots & b_{JK} \end{pmatrix}, \tag{2.15}$$

where for all $k, j = 1, \dots, J$, $k < j$, $b_{kj} = 0$ and $b_{jj} > 0$. This choice of $\boldsymbol{B}$ is largely used in the literature, although $\boldsymbol{B}$ may take various forms according to an arbitrary orthogonal rotation. More examples of matrices with *a priori* zeros are given by Jöreskog (1969).

2.4.2.2 The factor number constraint

Another identification problem is caused by the condition on the variance-covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Psi}$. We need to constrain the number of factors. Indeed, the number of distinct elements of $\boldsymbol{\Sigma}$ is equal to $K(K+1)/2$, however the number of free parameters in the model's variance-covariance matrix is JK elements for $\boldsymbol{B}$ plus K elements for $\boldsymbol{\Psi}$. The rotation constraint imposes $J(J-1)/2$ *a priori* zeroes on $\boldsymbol{B}$. Finally, the right hand side has $JK + K - J(J-1)/2$ distinct elements. To determine those parameters, the difference between the number of distinct elements of $\boldsymbol{\Sigma}$ and $\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Psi}$ must be positive

- If $K(K+1)/2 < JK + K - J(J-1)/2$, there is more parameters than equations, so the number of solutions is infinite.
- If $K(K+1)/2 = JK + K - J(J-1)/2$, there exists a solution, but the number of parameters being equal to the number of equations, we have no reduction of the number of parameters to estimate.
- If $K(K+1)/2 > JK + K - J(J-1)/2$, there exists a unique solution and the number of parameters is reduced.

In other words, the number of factors should satisfy

$$J \leq \left\lfloor \left(2K + 1 - \sqrt{8K+1}\right)/2 \right\rfloor.$$

Table 2.1 gives few examples of the maximal number of factors with respect to the number of variables.

Table 2.1: The maximum number of factors with respect to the number of responses

K	1	2	3	4	5	6	7	8	9	10
J max	0	0	1	1	2	3	3	4	5	6

2.4.3 The EM algorithm for factor analysis

In order to estimate the parameters, we want to maximize the log-likelihood. Due to the latent variables $\boldsymbol{G}$, this log-likelihood has a complex expression which makes it difficult to maximize directly. We thus use the EM algorithm recalled in Section 2.3.2. We calculate and then maximize the expectation of the complete log-likelihood conditional on the observed data: $\mathbb{E}[l(\boldsymbol{\Theta}; \boldsymbol{Y}, \boldsymbol{G})|\boldsymbol{Y}, \boldsymbol{\Theta'}]$.

2.4.3.1 The Expectation step

To perform the E step of the algorithm, we need to explicitly calculate the conditional expectation of the complete log-likelihood

$$\begin{aligned}\mathbb{E}[l(\boldsymbol{\Theta}; \boldsymbol{Y}, \boldsymbol{G})|\boldsymbol{Y}, \boldsymbol{\Theta'}] &= \sum_{n=1}^{N} \mathbb{E}[l(\boldsymbol{\Theta}; \boldsymbol{y_n}, \boldsymbol{g_n})|\boldsymbol{y_n}, \boldsymbol{\Theta'}] \\ &= \sum_{n=1}^{N} \int L\left(\boldsymbol{g_n}|\boldsymbol{y_n}; \boldsymbol{\Theta'}\right) \ln\left(L(\boldsymbol{y_n}, \boldsymbol{g_n}; \boldsymbol{\Theta})\right) d\boldsymbol{g_n} \\ &= \sum_{n=1}^{N} \int \ln\left\{L(\boldsymbol{y_n}|\boldsymbol{g_n}; \boldsymbol{\Theta}) L\left(\boldsymbol{g_n}; \boldsymbol{\Theta}\right)\right\} L\left(\boldsymbol{g_n}|\boldsymbol{y_n}; \boldsymbol{\Theta'}\right) d\boldsymbol{g_n}.\end{aligned}$$

We first need to find the distribution of $\boldsymbol{g_n}|\boldsymbol{y_n}$. Since the random vector $(\boldsymbol{y_n^T}, \boldsymbol{g_n^T})^{\boldsymbol{T}}$ is Gaussian as shown by Equation (2.14), we have, due to the conditioning rule of the multivariate Gaussian distribution, $\boldsymbol{g_n}|\boldsymbol{y_n} \sim \mathcal{N}_J\left(\boldsymbol{\alpha}\left(\boldsymbol{y_n} - \boldsymbol{\mu_n}\right), \boldsymbol{I_J} - \boldsymbol{\alpha B^T}\right)$ with $\boldsymbol{\alpha} = \boldsymbol{B}\left(\boldsymbol{B^T B} + \boldsymbol{\Psi}\right)^{-1}$. The conditional moments of this distribution are then given by

$$\begin{aligned}\tilde{\boldsymbol{g}}_{\boldsymbol{n}} &:= \mathbb{E}\left[\boldsymbol{g_n}|\boldsymbol{y_n}; \boldsymbol{\Theta}\right] \\ &= \boldsymbol{\alpha}(\boldsymbol{y_n} - \boldsymbol{\mu_n})\end{aligned}$$

and

$$\begin{aligned}\tilde{\boldsymbol{R}}_{\boldsymbol{n}} &:= \mathbb{E}\left[\boldsymbol{g_n g_n^T}|\boldsymbol{y_n}; \boldsymbol{\Theta}\right] \\ &= \mathbb{V}\left[\boldsymbol{g_n}|\boldsymbol{y_n}; \boldsymbol{\Theta}\right] + \mathbb{E}\left[\boldsymbol{g_n}|\boldsymbol{y_n}; \boldsymbol{\Theta}\right] \mathbb{E}\left[\boldsymbol{g_n}|\boldsymbol{y_n}; \boldsymbol{\Theta}\right]^T \\ &= \boldsymbol{I_J} - \boldsymbol{\alpha B^T} + \tilde{\boldsymbol{g}}_{\boldsymbol{n}} \tilde{\boldsymbol{g}}_{\boldsymbol{n}}^{\boldsymbol{T}}.\end{aligned}$$

The conditional expectation of the complete log-likelihood thus becomes

$$
\begin{aligned}
&\mathbb{E}\left[l(\boldsymbol{\Theta};\boldsymbol{Y},\boldsymbol{G})|\boldsymbol{Y},\boldsymbol{\Theta'}\right] \\
&= -\frac{1}{2}\sum_{n=1}^{N}\Bigg\{(K+J)\ln(2\pi)+\sum_{k=1}^{K}\ln(\sigma_k^2)+ \\
&\quad \mathbb{E}\left[\boldsymbol{g_n^T g_n}+\sum_{k=1}^{K}\frac{1}{\sigma_k^2}\left(y_{nk}-\mu_{nk}-\boldsymbol{g_n^T b_k}\right)^2\middle|\boldsymbol{y_n};\boldsymbol{\Theta'}\right]\Bigg\} \\
&= -\frac{1}{2}\sum_{n=1}^{N}\Bigg\{(K+J)\ln(2\pi)+\sum_{k=1}^{K}\ln(\sigma_k^2)+\mathbb{E}\left[\boldsymbol{g_n^T g_n}|\boldsymbol{y_n};\boldsymbol{\Theta'}\right]+ \\
&\quad \mathbb{E}\Bigg[\sum_{k=1}^{K}\frac{1}{\sigma_k^2}\Big(\left(y_{nk}-\mu_{nk}\right)^2+\boldsymbol{b_k^T}\left(\boldsymbol{g_n g_n^T}\right)\boldsymbol{b_k}- \\
&\quad 2\left(y_{nk}-\mu_{nk}\right)\boldsymbol{g_n^T b_k}\Big)|\boldsymbol{y_n};\boldsymbol{\Theta'}\Bigg]\Bigg\} \\
&= -\frac{1}{2}\sum_{n=1}^{N}\Bigg\{(K+J)\ln(2\pi)+\sum_{k=1}^{K}\ln(\sigma_k^2)+\mathbb{E}\left[\boldsymbol{g_n^T g_n}|\boldsymbol{y_n};\boldsymbol{\Theta'}\right]+ \\
&\quad \sum_{k=1}^{K}\frac{1}{\sigma_k^2}\left(\left(y_{nk}-\mu_{nk}\right)^2+\boldsymbol{b_k^T}\tilde{\boldsymbol{R}}_{\boldsymbol{n}}\boldsymbol{b_k}-2\left(y_{nk}-\mu_{nk}\right)\tilde{\boldsymbol{g}}_{\boldsymbol{n}}^{\boldsymbol{T}}\boldsymbol{b_k}\right)\Bigg\} \\
&= -\frac{1}{2}\Bigg\{N(K+J)\ln(2\pi)+N\sum_{k=1}^{K}\ln(\sigma_k^2)+\sum_{n=1}^{N}\mathbb{E}\left[\boldsymbol{g_n^T g_n}|\boldsymbol{y_n};\boldsymbol{\Theta'}\right]+ \\
&\quad \sum_{k=1}^{K}\frac{1}{\sigma_k^2}\left(\left\|\boldsymbol{y_k}-\boldsymbol{\mu_k}\right\|^2+\boldsymbol{b_k^T}\tilde{\boldsymbol{R}}\boldsymbol{b_k}-2\left(\tilde{\boldsymbol{G}}\boldsymbol{b_k}\right)^T\left(\boldsymbol{y_k}-\boldsymbol{\mu_k}\right)\right)\Bigg\},
\end{aligned}
$$

where the rows of the matrix $\tilde{\boldsymbol{G}}$ are the $\tilde{\boldsymbol{g}}_{\boldsymbol{n}}^{\boldsymbol{T}}$'s and $\tilde{\boldsymbol{R}}=\sum_{n=1}^{N}\tilde{\boldsymbol{R}}_{\boldsymbol{n}}$ is the sum of the order two conditional moments.

2.4.3.2 The Maximization step

The M-step maximizes, with respect to $\boldsymbol{\Theta}$, the conditional expectation of the complete log-likelihood, subject to the constraint on matrix $\boldsymbol{B}$ presented in Section 2.4.2.1. However, the parameters $\boldsymbol{\mu_k}$ and σ_k^2 are not concerned by the constraint. Thus, the first order conditions of the maximization yield

$$
\begin{aligned}
&\nabla_{\boldsymbol{\mu_k}}\mathbb{E}\left[l(\boldsymbol{\Theta};\boldsymbol{Y},\boldsymbol{G})|\boldsymbol{Y},\boldsymbol{\Theta'}\right]=0 \\
&\Leftrightarrow \nabla_{\boldsymbol{\mu_k}}\left\{\left\|\boldsymbol{y_k}-\boldsymbol{\mu_k}\right\|^2-2\left(\tilde{\boldsymbol{G}}\boldsymbol{b_k}\right)^T\left(\boldsymbol{y_k}-\boldsymbol{\mu_k}\right)\right\}=0 \\
&\Leftrightarrow \left(\boldsymbol{y_k}-\boldsymbol{\mu_k}\right)-\tilde{\boldsymbol{G}}\boldsymbol{b_k}=0 \\
&\Leftrightarrow \boldsymbol{\mu_k}=\boldsymbol{y_k}-\tilde{\boldsymbol{G}}\boldsymbol{b_k}.
\end{aligned}
$$

Besides,

$$
\begin{aligned}
&\nabla_{\sigma_k^2}\mathbb{E}[l(\boldsymbol{\Theta};\boldsymbol{Y},\boldsymbol{G})|\boldsymbol{Y},\boldsymbol{\Theta'}] = 0 \\
&\Leftrightarrow \nabla_{\sigma_k^2}\left\{N\ln(\sigma_k^2) + \frac{1}{\sigma_k^2}\left(\|\boldsymbol{y_k}-\boldsymbol{\mu_k}\|^2 + \boldsymbol{b_k^T}\tilde{\boldsymbol{R}}\boldsymbol{b_k} - 2\left(\tilde{\boldsymbol{G}}\boldsymbol{b_k}\right)^T(\boldsymbol{y_k}-\boldsymbol{\mu_k})\right)\right\} = 0 \\
&\Leftrightarrow N - \frac{1}{\sigma_k^2}\left(\|\boldsymbol{y_k}-\boldsymbol{\mu_k}\|^2 + \boldsymbol{b_k^T}\tilde{\boldsymbol{R}}\boldsymbol{b_k} - 2\left(\tilde{\boldsymbol{G}}\boldsymbol{b_k}\right)^T(\boldsymbol{y_k}-\boldsymbol{\mu_k})\right) = 0 \\
&\Leftrightarrow \sigma_k^2 = \frac{1}{N}\left(\|\boldsymbol{y_k}-\boldsymbol{\mu_k}\|^2 + \boldsymbol{b_k^T}\tilde{\boldsymbol{R}}\boldsymbol{b_k} - 2\left(\tilde{\boldsymbol{G}}\boldsymbol{b_k}\right)^T(\boldsymbol{y_k}-\boldsymbol{\mu_k})\right)
\end{aligned}
$$

Now, we need to maximize vector $\boldsymbol{b_k}$ under the constraint given by Equation (2.15). For each $k = 1,\ldots,J$, let $\boldsymbol{b_k^T} = (\boldsymbol{b_{1:k,k}^T}, \boldsymbol{0}^T)$ be the regression parameters, where $\boldsymbol{b_{1:k,k}^T} = (b_{1k},\ldots,b_{kk})$ is a vector of length k to be estimated and $\boldsymbol{0}$ is a null vector of length $(J-k)$ *a priori* fixed. In this case, we define $\tilde{\boldsymbol{R}}_{\boldsymbol{1:k,1:k}}$ the submatrix of size $k\times k$ of $\tilde{\boldsymbol{R}}$ and $\tilde{\boldsymbol{G}}_{\boldsymbol{1:k}}$ the matrix composed by the first k columns of $\tilde{\boldsymbol{G}}$. The maximization yields

$$
\begin{aligned}
&\nabla_{\boldsymbol{b_{1:k,k}}}\mathbb{E}[l(\boldsymbol{\Theta};\boldsymbol{Y},\boldsymbol{G})|\boldsymbol{Y},\boldsymbol{\Theta'}] = 0 \\
&\Leftrightarrow \nabla_{\boldsymbol{b_{1:k,k}}}\left\{\boldsymbol{b_{1:k,k}^T}\left(\tilde{\boldsymbol{R}}_{\boldsymbol{1:k,1:k}}\right)\boldsymbol{b_{1:k,k}} - 2\left(\tilde{\boldsymbol{G}}_{\boldsymbol{1:k}}\boldsymbol{b_{1:k,k}}\right)^T(\boldsymbol{y_k}-\boldsymbol{\mu_k})\right\} = 0 \\
&\Leftrightarrow \left(\tilde{\boldsymbol{G}}_{\boldsymbol{1:k}}\right)^T(\boldsymbol{y_k}-\boldsymbol{\mu_k}) - \left(\tilde{\boldsymbol{R}}_{\boldsymbol{1:k,1:k}}\right)\boldsymbol{b_{1:k,k}} = 0 \\
&\Leftrightarrow \boldsymbol{b_{1:k,k}} = \left(\tilde{\boldsymbol{R}}_{\boldsymbol{1:k,1:k}}\right)^{-1}\left(\tilde{\boldsymbol{G}}_{\boldsymbol{1:k}}\right)^T(\boldsymbol{y_k}-\boldsymbol{\mu_k}).
\end{aligned}
$$

Likewise, for $k = J+1,\ldots,K$, $\boldsymbol{b_k}$ is given by

$$
\boldsymbol{b_k} = \tilde{\boldsymbol{R}}^{-1}\tilde{\boldsymbol{G}}^T(\boldsymbol{y_k}-\boldsymbol{\mu_k}).
$$

2.4.3.3 The factor model estimation algorithm

As a result of the aforementioned developments, we shall use Algorithm 5 to estimate factor model parameters.

2.4.4 Extension to GLMs

In this section, we explore approaches allowing to find some factors in the case of multivariate abundance data. In the literature, two ways are developed to find the factors. In the first one, the factor are non-random and then must be estimated as simple parameters. The second way assumes the multivariate normality of the factors and looks for a method to maximize the log-likelihood. A multivariate abundance dataset can be defined by a matrix $\boldsymbol{Y}\in\mathbb{R}^{N\times K}$ where each column is composed by realizations of a GLM.

Algorithm 5: The EM algorithm applied to factor models

while *not convergence* **do**

Expectation step

for $n = 1, \dots, N$ **do**

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{B}^{(t)} \left(\boldsymbol{B}^{(t)\,T} \boldsymbol{B}^{(t)} + \boldsymbol{\Psi}^{(t)} \right)^{-1}$$

$$\tilde{\boldsymbol{g}}_n^{(t+1)} = \boldsymbol{\alpha}^{(t+1)} \left(\boldsymbol{y}_n - \boldsymbol{\mu}_n^{(t)} \right)$$

$$\tilde{\boldsymbol{R}}_n^{(t+1)} = \boldsymbol{I}_L - \boldsymbol{\alpha}^{(t+1)} \boldsymbol{B}^{(t)\,T} + \tilde{\boldsymbol{g}}_n^{(t+1)} \tilde{\boldsymbol{g}}_n^{(t+1)\,T}$$

end

Maximization step

for $k = 1, \dots, K$ **do**

$$\boldsymbol{\mu}_k^{(t+1)} = \boldsymbol{y}_k - \tilde{\boldsymbol{G}}^{(t+1)} \boldsymbol{b}_k^{(t)}$$

$$\sigma_k^{2\,(t+1)} = \frac{1}{N} \left(\left\| \boldsymbol{y}_k - \boldsymbol{\mu}_k^{(t+1)} \right\|^2 + \left(\boldsymbol{b}_k^{(t)} \right)^T \tilde{\boldsymbol{R}}^{(t+1)} \boldsymbol{b}_k^{(t)} - 2 \left(\tilde{\boldsymbol{G}}^{(t+1)} \boldsymbol{b}_k^{(t)} \right)^T \left(\boldsymbol{y}_k - \boldsymbol{\mu}_k^{(t+1)} \right) \right)$$

if $k \leq L$ **then**

$$\boldsymbol{b}_{\boldsymbol{1:k,k}}^{(t+1)} = \left(\tilde{\boldsymbol{R}}_{\boldsymbol{1:k,1:k}}^{(t+1)} \right)^{-1} \left(\tilde{\boldsymbol{G}}_{\boldsymbol{1:k}}^{(t+1)} \right)^T \left(\boldsymbol{y}_k - \boldsymbol{\mu}_k^{(t+1)} \right)$$

else

$$\boldsymbol{b}_k^{(t+1)} = \left(\tilde{\boldsymbol{R}}^{(t+1)} \right)^{-1} \left(\tilde{\boldsymbol{G}}^{(t+1)} \right)^T \left(\boldsymbol{y}_k - \boldsymbol{\mu}_k^{(t+1)} \right)$$

end

end

$t \leftarrow t + 1$

end

2.4.4.1 Non random factors

In the case of non random factors, the main objective consists in identifying low-dimensional features in high-dimensional multivariate abundance data (Lee et al., 2013; Sohn and Li, 2018; Xu et al., 2021). To encourage dimension reduction, we assume that matrix $\boldsymbol{\eta} = \boldsymbol{GB} \in \mathbb{R}^{N \times K}$ has a low-rank structure with rank $J < \min(N, K)$, where $\boldsymbol{G}$ is the factor matrix and $\boldsymbol{B}$ the loading matrix. Thus, the linear predictor associated with response y_{nk} is defined as

$$\eta_{nk} = g_{n1} b_{1k} + \cdots + g_{nJ} b_{Jk},$$

where g_{nj} and b_{jk} are elements of $\boldsymbol{G}$ and $\boldsymbol{B}$ respectively. The matrices $\boldsymbol{G}$ and $\boldsymbol{B}$ being non random, the key of the estimation is to alternate two generalized linear regressions,

the first one consists in estimating $\boldsymbol{B}$ given $\boldsymbol{G}$ and the second one in obtaining $\boldsymbol{G}$ given $\boldsymbol{B}$.

2.4.4.2 Random factors

In a context of latent variables, an approach to the statistical modeling of multivariate abundances is the Generalized Linear Latent Variable Model (GLLVM, Skrondal and Rabe-Hesketh, 2004). In a particular case of random factors, the GLLVM extends the basic GLM by expressing the mean parameter as a linear combination of the covariates and the factors. The model writes

$$\eta_{nk} = \boldsymbol{x_n^T}\boldsymbol{\beta_k} + \boldsymbol{g_n^T}\boldsymbol{b_k},$$

where $\boldsymbol{x_n}$ is the vector of covariates, $\boldsymbol{g_n}$ the vector of factors and $\boldsymbol{\beta_k}$ and $\boldsymbol{b_k}$ their regression parameters. As in standard factor models, $\boldsymbol{g_n}$ is supposed drawn from a multivariate Gaussian distribution with zero mean and identity variance-covariance matrix. The marginal log-likelihood is obtained by integrating over the latent variables $\boldsymbol{g_n}$

$$\begin{aligned} l(\boldsymbol{\Theta}; \boldsymbol{Y}) &= \sum_{n=1}^{N} \ln\left(L(\boldsymbol{y_n}; \boldsymbol{\Theta})\right) \\ &= \sum_{n=1}^{N} \ln\left(\int \prod_{k=1}^{K} L(y_{nk}|\boldsymbol{g_n}; \boldsymbol{\Theta}) L(\boldsymbol{g_n})\, d\boldsymbol{g_n}\right), \end{aligned}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\beta_k}, \boldsymbol{b_k}; k = 1, \ldots, K\}$ is the set of parameters. Unfortunately, the previous log-likelihood derived from GLLVM cannot be expressed analytically. Several works propose to maximize this log-likelihood, but some of them suffer from a consuming computation time. We may cite for instance the works using the adaptive quadrature (Rabe-Hesketh et al., 2002), the EM algorithm in conjunction with Monte Carlo integration (Hui et al., 2015) or the works using Bayesian Markov Chain Monte Carlo (MCMC) (Hui, 2016; Tikhonov et al., 2020). A few methods cut down the computation time by taking closed form approximations of the log-likelihood. For instance, the approaches employing a variational approximation (Hui et al., 2017), a Laplace approximation (Niku et al., 2017, 2019a) or an extended variational approximation (Korhonen et al., 2023) deserve mentioning.

Alternatively to the previous approaches, Saidane et al. (2013) propose to estimate the combination of a GLM with a factor model by using the EM algorithm for the factor model as a step in the IRLS algorithm. Indeed, the main idea is to consider two alternate steps: (i) Conditional on the factors, the linearized model writes

$$w_{nk} = \boldsymbol{x_n^T}\boldsymbol{\beta_k} + \boldsymbol{g_n^T}\boldsymbol{b_k} + \zeta_{nk},$$

where the expectation of the errors is given by $\mathbb{E}[\zeta_{nk}] = 0$ and the variance by the weight associated with GLR. The working variables $\boldsymbol{w_n}$ and the variance matrix $\boldsymbol{\Psi_n}$ can be estimated through the IRLS. (ii) Given $\boldsymbol{w_n}$ and $\boldsymbol{\Psi_n}$, the log-likelihood $l(\boldsymbol{\Theta}, \boldsymbol{\mathcal{W}})$, where $\boldsymbol{\mathcal{W}}$

denotes the set of working variables, is maximized through the EM algorithm presented in Section 2.4.3.

CHAPTER 3

RESPONSE MIXTURE MODELS BASED ON SUPERVISED COMPONENTS

Contents

This chapter is a modified version of a published paper freely available at https://hal.archives-ouvertes.fr /hal-03547177/.

3.1 Response Mixture SCGLR

In this section, we present the framework and the modeling objectives for which we propose to combine SCGLR with a Finite Mixture Model (FMM). Section 3.1.1 presents our mixture model for the responses. Section 3.1.2 introduces the EM algorithm we develop, of which Sub-sections 3.1.2.1 and 3.1.2.2 detail the expectation and maximization steps respectively. The explicit EM algorithm is given in Sub-section 3.1.2.3. In Section 3.1.3 the method to calculate successive components is recalled: Sub-section 3.1.3.2 deals with the first component, and Sub-section 3.1.3.3 explains how to calculate further components. The overall algorithm is shown in Section 3.1.4. Finally, Section 3.1.5 gives the heuristic we use to tune the hyper-parameters real values.

3.1.1 The response mixture model

Let $\boldsymbol{Y} = [\boldsymbol{y_1}, \ldots, \boldsymbol{y_K}] \in \mathbb{R}^{N \times K}$ be the response matrix. The responses are assumed to be modeled through a finite mixture of regression models, comprising G groups. The probability distribution function (pdf) of response $\boldsymbol{y_k}$ is thus

$$L(\boldsymbol{y_k}; \boldsymbol{\theta_k}) = \sum_{g=1}^{G} p_g \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkg}),$$

where the nth observation in the kth response of the gth group has a pdf d_k belonging to the exponential family, with expectation μ_{nkg}. $\boldsymbol{\theta_k}$ denotes the vector of parameters, including the regression parameters $\boldsymbol{\gamma_{kg}}$ and $\boldsymbol{\delta_{kg}}$, as defined in Section 2.2.3.1, and p_g the gth mixing probability with $p_g \in [0, 1]$ and $\sum_{g=1}^{G} p_g = 1$. Denoting h_k the kth canonical link function, we assume, for a single component model

$$h_k(\mu_{nkg}) = (\boldsymbol{x_n^T u_g})\boldsymbol{\gamma_{kg}} + \boldsymbol{a_n^T \delta_{kg}},$$

where $\boldsymbol{u_g}$ is the loading vector of the (first) component of group g, and $\boldsymbol{x_n}$ and $\boldsymbol{a_n}$ are the nth rows of matrices $\boldsymbol{X}$ and $\boldsymbol{A}$ respectively. Thus, the responses in group g are predicted by component $\boldsymbol{f_g} = \boldsymbol{X u_g}$, together with $\boldsymbol{A}$. For each $k = 1, \ldots, K$, d_k and h_k are chosen so as to suit the type of response $\boldsymbol{y_k}$ (e.g. binary, count, categorical, continuous etc.).

Conditional on the explanatory variables, the response variables are assumed independent. The group memberships of the responses being unknown, the model log-likelihood

$$l(\boldsymbol{\Theta}; \boldsymbol{Y}) = \sum_{k=1}^{K} \ln \left(L(\boldsymbol{y_k}; \boldsymbol{\theta_k}) \right),$$

where the set of parameters is $\boldsymbol{\Theta} = \{\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_K}\}$, being difficult to maximize directly, we adopt the EM algorithm to estimate the model parameters.

Let z_{kg} be the latent indicator variable equal to 1 if the response $\boldsymbol{y_k}$ belongs to the gth group, and 0 otherwise. Let $\boldsymbol{z_k} = (z_{kg}\,;\, g = 1, \ldots, G)$ be the vector of group membership

indicators of response $\boldsymbol{y_k}$, and let $\boldsymbol{Z} = [\boldsymbol{z_k}\,;\, k = 1, \ldots, K]$ be their $G \times K$ matrix. Conditional on $z_{kg} = 1$, the pdf of response $\boldsymbol{y_k}$ for unit n is $d_k(y_{nk}; \mu_{nkg})$. The model complete log-likelihood writes

$$\begin{aligned} l(\boldsymbol{\Theta}; \boldsymbol{Y}, \boldsymbol{Z}) &= \ln \left(\prod_{k=1}^{K} \prod_{g=1}^{G} \left[p_g \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkg}) \right]^{z_{kg}} \right) \\ &= \sum_{k=1}^{K} \sum_{g=1}^{G} z_{kg} \ln \left(p_g \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkg}) \right). \end{aligned}$$

3.1.2 Adapting the EM algorithm to a response mixture model

Step (i) in Section 2.2.3.4 boils down to maximizing the likelihood of the component-based model. Owing to the latent variable $\boldsymbol{Z}$, this step will be performed using the EM algorithm.

3.1.2.1 The expectation (E) step

With $z_k = g$ meaning that the gth coordinate of the vector $\boldsymbol{z_k}$ equals 1, the conditional expectation of the complete log-likelihood writes

$$\begin{aligned} \mathbb{E}\left[l(\boldsymbol{\Theta}; \boldsymbol{Y}, \boldsymbol{Z}) | \boldsymbol{Y}; \boldsymbol{\Theta'} \right] &= \sum_{k=1}^{K} \mathbb{E}\left[\ln \left(L(\boldsymbol{y_k}, \boldsymbol{z_k}; \boldsymbol{\theta_k}) \right) | \boldsymbol{y_k}; \boldsymbol{\theta'_k} \right] \\ &= \sum_{k=1}^{K} \sum_{g=1}^{G} \mathbb{P}(\boldsymbol{z_k} = g | \boldsymbol{y_k}; \boldsymbol{\theta'_k}) \ln \left(L(\boldsymbol{y_k}, \boldsymbol{z_k} = g; \boldsymbol{\theta_k}) \right) \\ &= \sum_{k=1}^{K} \sum_{g=1}^{G} \mathbb{P}(\boldsymbol{z_k} = g | \boldsymbol{y_k}; \boldsymbol{\theta'_k}) \ln \left(\mathbb{P}(\boldsymbol{z_k} = g; \boldsymbol{\theta_k}) L(\boldsymbol{y_k} | \boldsymbol{z_k} = g; \boldsymbol{\theta_k}) \right) \\ &= \sum_{k=1}^{K} \sum_{g=1}^{G} \alpha_{kg} \ln \left(p_g \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkg}) \right). \end{aligned}$$

The posterior group membership probabilities of each response $\boldsymbol{y_k}$ are computed as

$$\alpha_{kg} := \mathbb{P}(\boldsymbol{z_k} = g | \boldsymbol{y_k}; \boldsymbol{\theta_k}) = \frac{L(\boldsymbol{y_k}, \boldsymbol{z_k} = g; \boldsymbol{\theta_k})}{L(\boldsymbol{y_k}; \boldsymbol{\theta_k})} = \frac{p_g \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkg})}{\sum_{r=1}^{G} p_r \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkr})}.$$

3.1.2.2 The maximization (M) step

The M-step maximizes with respect to $\boldsymbol{\Theta}$ the conditional expectation of the complete log-likelihood, subject to the constraint $\sum_{g=1}^{G} p_g = 1$. We maximize the corresponding Lagrangian

$$\mathcal{L}(\boldsymbol{\Theta}, \lambda) = \mathbb{E}\left[l(\boldsymbol{\Theta}; \boldsymbol{Y}, \boldsymbol{Z}) | \boldsymbol{Y}; \boldsymbol{\Theta'} \right] - \lambda \left(\sum_{g=1}^{G} p_g - 1 \right).$$

The maximization with respect to p_g yields

$$\begin{aligned}\nabla_{p_g}\mathcal{L}(\boldsymbol{\Theta},\lambda)=0 &\Leftrightarrow \sum_{k=1}^{K}\alpha_{kg}=p_g\lambda\\ &\Leftrightarrow \sum_{k=1}^{K}\underbrace{\sum_{g=1}^{G}\alpha_{kg}}_{=1}=\lambda\underbrace{\sum_{g=1}^{G}p_g}_{=1}\\ &\Leftrightarrow K=\lambda.\end{aligned}$$

So, the solution p_g is

$$\hat{p}_g=\frac{1}{K}\sum_{k=1}^{K}\alpha_{kg}.$$

The estimates of the regression parameters $\boldsymbol{\gamma_{kg}}$ and $\boldsymbol{\delta_{kg}}$ are obtained as the solutions of

$$\nabla_{\boldsymbol{\gamma_{kg}}}\mathbb{E}\left[l(\boldsymbol{\Theta};\boldsymbol{Y},\boldsymbol{Z})|\boldsymbol{Y};\boldsymbol{\Theta'}\right]=0\Leftrightarrow\nabla_{\boldsymbol{\gamma_{kg}}}\sum_{n=1}^{N}\ln(d_k(y_{nk};\mu_{nkg}))=0$$

and

$$\nabla_{\boldsymbol{\delta_{kg}}}\mathbb{E}\left[l(\boldsymbol{\Theta};\boldsymbol{Y},\boldsymbol{Z})|\boldsymbol{Y};\boldsymbol{\Theta'}\right]=0\Leftrightarrow\nabla_{\boldsymbol{\delta_{kg}}}\sum_{n=1}^{N}\ln(d_k(y_{nk};\mu_{nkg}))=0.$$

These equations characterize the maximum likelihood estimation of the GLM of $\boldsymbol{y_k}$ in each group g. This estimate can be obtained as the fixed point of the FSA.

Assuming the response variable $\boldsymbol{y_k}$ belongs to the gth group, the working variable associated with y_{nk} is calculated as

$$w_{nkg}=h_k\left(\mu_{nkg}\right)+\left(y_{nk}-\mu_{nkg}\right)h_k'\left(\mu_{nkg}\right)=\eta_{nkg}+\zeta_{nkg},$$

where $\zeta_{nkg}=\left(y_{nk}-\mu_{nkg}\right)h_k'\left(\mu_{nkg}\right)$. In view of the conditional independence assumption, the variance matrix for w_{nkg} is

$$\mathbb{V}\left[w_{kg}\right]=W_{kg}^{-1}=\text{diag}\left(a_{nk}(\phi_k)v_k\left(\mu_{nkg}\right)h_k'\left(\mu_{nkg}\right)^2\right)_{n=1,\ldots,N},$$

where a_{nk} and v_k are known functions and ϕ_k is the dispersion parameter related to $\boldsymbol{y_k}$. Thus, to optimize the regression parameters, we perform a generalized least squares step in the linearized model defined by

$$\boldsymbol{w_{kg}}=\left(\boldsymbol{X}\boldsymbol{u_g}\right)\boldsymbol{\gamma_{kg}}+\boldsymbol{A}\boldsymbol{\delta_{kg}}+\boldsymbol{\zeta_{kg}},$$

with $\mathbb{E}[\boldsymbol{\zeta_{kg}}]=0$ and $\mathbb{V}[\boldsymbol{\zeta_{kg}}]=\boldsymbol{W_{kg}^{-1}}$.

3.1.2.3 The response mixture estimation algorithm

As a result of the aforementioned developments, we shall use Algorithm 6 to estimate the parameters of the response mixture model.

Algorithm 6: The EM algorithm adapted to the response mixture

$A_g := [f_g, A]$

while *not convergence* **do**

Expectation step

for $k = 1, \ldots, K$ **do**

for $g = 1, \ldots, G$ **do**

$$\alpha_{kg}^{(t+1)} = \frac{p_g^{(t)} \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkg}^{(t)})}{\sum_{r=1}^{G} p_r^{(t)} \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkr}^{(t)})}$$

end

end

Maximization step

for $g = 1, \ldots, G$ **do**

$p_g^{(t+1)} = \frac{1}{K} \sum_{k=1}^{K} \alpha_{kg}^{(t+1)}$

for $k = 1, \ldots, K$ **do**

$$\left(\boldsymbol{\gamma}_{\boldsymbol{kg}}^{(t+1)}, \boldsymbol{\delta}_{\boldsymbol{kg}}^{(t+1)\,T}\right)^T = \left(\boldsymbol{A}_{\boldsymbol{g}}^{\boldsymbol{T}} \boldsymbol{W}_{\boldsymbol{kg}}^{(t)} \boldsymbol{A}_{\boldsymbol{g}}\right)^{-1} \boldsymbol{A}_{\boldsymbol{g}}^{\boldsymbol{T}} \boldsymbol{W}_{\boldsymbol{kg}}^{(t)} \boldsymbol{w}_{\boldsymbol{kg}}^{(t)}$$

$$\boldsymbol{\eta}_{\boldsymbol{kg}}^{(t+1)} = \boldsymbol{f}_{\boldsymbol{g}} \boldsymbol{\gamma}_{\boldsymbol{kg}}^{(t+1)} + \boldsymbol{A} \boldsymbol{\delta}_{\boldsymbol{kg}}^{(t+1)}$$

$$\mu_{nkg}^{(t+1)} = h_k^{-1}\left(\eta_{nkg}^{(t+1)}\right), \forall n = 1, \ldots, N$$

$$w_{nkg}^{(t+1)} = \eta_{nkg}^{(t+1)} + h_k'\left(\mu_{nkg}^{(t+1)}\right)\left(y_{nk} - \mu_{nkg}^{(t+1)}\right), \forall n = 1, \ldots, N$$

$$\boldsymbol{W}_{kg}^{(t+1)} = \text{diag}\left(\left[a_{nk}(\phi_k) v_k\left(\mu_{nkg}^{(t+1)}\right) h_k'\left(\mu_{nkg}^{(t+1)}\right)^2\right]^{-1}\right)_{n=1,\ldots,N}$$

end

end

$t \leftarrow t + 1$

end

3.1.3 Calculating rank 1 components of response groups

When clustering the responses according to their group-specific SCGLR components, we must ensure that the explanatory subspaces spanned by the components associated to response clusters be reasonably separated (otherwise the EM algorithm may fail to converge). To achieve that, when calculating a component explanatory of a response cluster, we must prevent that it be too close to the explanatory subspaces of other clusters.

3.1.3.1 An additional sub-criterion to better separate explanatory sub-spaces

Let $\boldsymbol{F}_{-g} = \{\boldsymbol{f_1}, \ldots, \boldsymbol{f_{g-1}}, \boldsymbol{f_{g+1}}, \ldots, \boldsymbol{f_G}\}$ be the set of components from which we remove the component of group g. The space spanned by the component $\boldsymbol{f_g}$ may be uniquely represented by the orthogonal projector on it: $\boldsymbol{\varpi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{f_g}]}$. With this in mind, we propose to measure the separation of span$[\boldsymbol{f_g}]$ from span$[\boldsymbol{f_r}]$'s, for all $r \neq g$, through the following function of $\boldsymbol{u_g}$

$$\varphi_{\boldsymbol{F}_{-g}}(\boldsymbol{u_g}) = 1 - \frac{1}{G-1} \sum_{r \neq g} \left\langle \boldsymbol{\varpi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{f_g}]}, \boldsymbol{\varpi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{f_r}]} \right\rangle_{\text{Frob}}. \tag{3.1}$$

Indeed, if, for all $r \neq g$, span$[\boldsymbol{f_g}]$ is orthogonal to span$[\boldsymbol{f_r}]$ then the Frobenius product will be zero, so that the criterion will be equal to 1. At the other end, if, for all $r \neq g$, span$[\boldsymbol{f_g}]$ = span$[\boldsymbol{f_r}]$, the Frobenius product will be equal to 1, and the criterion to 0.

The new program optimizing the combined criterion we propose for group g is thus

$$\max_{\boldsymbol{u_g}, \boldsymbol{u_g^T M^{-1} u_g}=1} s \ln(\phi(\boldsymbol{u_g})) + t \ln(\varphi_{\boldsymbol{F}_{-g}}(\boldsymbol{u_g})) + (1-s-t) \ln(\psi_{\boldsymbol{A}}(\boldsymbol{u_g})), \tag{3.2}$$

where $s, t, (s+t) \in [0, 1]$.

3.1.3.2 Rank 1 components

Let us now address step (ii) of the combined criterion maximization as detailed in Section 2.2.3.4. The GoF measure applied to group g is given by

$$\psi_{\boldsymbol{A}}(\boldsymbol{u_g}) = \sum_{k=1}^{K} \alpha_{kg} \left\| \boldsymbol{w_{kg}} \right\|^2_{\boldsymbol{W_{kg}}} \cos^2_{\boldsymbol{W_{kg}}} \left(\boldsymbol{w_{kg}}, \boldsymbol{\Pi}^{\boldsymbol{W_{kg}}}_{\text{span}[\boldsymbol{f_g}, \boldsymbol{A}]} \boldsymbol{w_{kg}} \right),$$

where the weights reflecting the degrees of membership to group g of responses are the posterior probabilities $\{\alpha_{1g}, \ldots, \alpha_{Kg}\}$. The functions ϕ and $\varphi_{\boldsymbol{F}_{-g}}$ are respectively given by Equation (2.9) and Equation (3.1). The explicit expression of the criterion is given in Supplementary Material.

3.1.3.3 Higher rank components

We shall henceforth calculate the higher rank components. Let $\boldsymbol{f}_g^h = \boldsymbol{X}\boldsymbol{u}_g^h$ be the hth component of group g, and let $\boldsymbol{F}_g^h = [\boldsymbol{f}_g^1, \ldots, \boldsymbol{f}_g^h]$, where $h \leq H_g$, be the matrix of the first h components of group g. We adopt the local nesting (LocNes) principle presented by Bry et al. (2009) and extended by Bry et al. (2012). According to the LocNes principle, the new component $\boldsymbol{f}_g^{h+1}$ must best complement both the existing ones and $\boldsymbol{A}$, that is $\boldsymbol{A}_g^h := [\boldsymbol{F}_g^h, \boldsymbol{A}]$. So $\boldsymbol{f}_g^{h+1}$ has to be calculated using $\boldsymbol{A}_g^h$ as the new set of additional covariates. Moreover, to avoid linear redundancy of components, we impose that $\boldsymbol{f}_g^{h+1}$ be orthogonal to $\boldsymbol{F}_g^h$, *i.e.* $\boldsymbol{F}_g^{h\,T}\boldsymbol{W}\boldsymbol{f}_g^{h+1} = 0$.

We calculate every new component as the solution of the optimization program given by Equation (3.2), with the additional constraint: $\boldsymbol{\Delta}_g^{h\,T}\boldsymbol{u}_g^{h+1} = 0$, where $\boldsymbol{\Delta}_g^h = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{F}_g^h$, and loop on g until overall convergence of the component system. Taking $\boldsymbol{A}_g^h = [\boldsymbol{F}_g^h, \boldsymbol{A}]$ and $\boldsymbol{F}_{-g} = \{\boldsymbol{F}_1^{H_1}, \ldots, \boldsymbol{F}_{g-1}^{H_{g-1}}, \boldsymbol{F}_{g+1}^{H_{g+1}}, \ldots, \boldsymbol{F}_G^{H_G}\}$, the sub-criteria become

$$\psi_{\boldsymbol{A}_g^h}\left(\boldsymbol{u}_g^{h+1}\right) = \sum_{k=1}^{K} \alpha_{kg} \|\boldsymbol{w}_{kg}\|^2_{\boldsymbol{W}_{kg}} \cos^2_{\boldsymbol{W}_{kg}}\left(\boldsymbol{w}_{kg}, \boldsymbol{\Pi}^{\boldsymbol{W}_{kg}}_{\text{span}[\boldsymbol{f}_g^{h+1}, \boldsymbol{A}_g^h]}\boldsymbol{w}_{kg}\right)$$

and

$$\varphi_{\boldsymbol{F}_{-g}}\left(\boldsymbol{u}_g^{h+1}\right) = 1 - \frac{1}{G-1}\sum_{r\neq g}\left\langle \boldsymbol{\varpi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F}_g^{h+1}]}, \boldsymbol{\varpi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F}_r^{H_r}]}\right\rangle_{\text{Frob}}.$$

For all $g = 1, \ldots, G$, the rank-1 component of group g is calculated using the same program with $\boldsymbol{A}_g^0 = \boldsymbol{A}$ and $\boldsymbol{\Delta}_g^0 = 0$.

3.1.3.4 Optimizing the cluster-specific components

In order to identify the groups, one may have to put a heavy weight on the separation sub-criterion. As a result, the supervised components output by the former maximization may be artificially too strongly separated between groups. So, this maximization is used merely to identify groups having specific explanatory dimensions. Posterior to that, we must optimize the group-specific components for prediction in a second phase, performing classical SCGLR separately on each group.

3.1.4 The overall algorithm

The method comprising these two phases (clustering, and component optimization), is named response mixture SCGLR (rmSCGLR). Algorithm 7 consists in alternating the following steps: (i) Given the current set of components, estimate the response mixture parameters through the EM algorithm; (ii) Given the current group memberships of responses, calculate all the components of all the groups. To give our algorithm a good starting point, namely well separated response clusters and strong initial components, we use the **ClustOfVar** package (Chavent et al., 2012) to determine G initial response

groups, and then, the **pls** package (Mevik and Wehrens, 2007) in each group, to find initial supervised components. In the component optimization phase, SCGLR is performed on each response group separately, each having specific components. This phase includes determining the best number of components for prediction by means of cross-validation.

Algorithm 7: The clustering phase algorithm

while *not convergence* **do**

Update mixture parameters with the EM algorithm described in Algorithm 6

$$\boldsymbol{\Theta}^{(n+1)} = \arg\max_{\boldsymbol{\Theta}} l(\boldsymbol{\Theta}^{(n)}; \boldsymbol{Y}, \boldsymbol{Z})$$

Update loading vectors with the PING algorithm described in Supplementary Material

for $g = 1, \ldots, G$ **do**

for $h = 1, \ldots, H_g$ **do**

$$\boldsymbol{u}_g^{h\,(n+1)} = \underset{\substack{\boldsymbol{u}_g^{h\,T}\boldsymbol{M}^{-1}\boldsymbol{u}_g^h = 1 \\ \boldsymbol{\Delta}_g^{h-1\,T}\boldsymbol{u}_g^h = 0}}{\arg\max} \phi\left(\boldsymbol{u}_g^h\right)^s \varphi_{\boldsymbol{F}_{-g}}\left(\boldsymbol{u}_g^h\right)^t \psi_{\boldsymbol{A}_g^{h-1}}\left(\boldsymbol{u}_g^h\right)^{1-s-t}$$

end

end

$n \leftarrow n + 1$

end

At the end, we can classify the responses according to their posterior probabilities. A response $\boldsymbol{y_k}$ is assigned to cluster g if

$$\alpha_{kg}^{(n_{\max})} > \alpha_{kr}^{(n_{\max})}, \forall r \neq g.$$

3.1.5 A hyper-parameter calibration heuristic

The hyper-parameters are calibrated minimizing the Bayesian Information Criterion (BIC, Schwarz, 1978), defined by

$$\text{BIC} = -2l\left(\boldsymbol{\Theta}; \boldsymbol{Y}\right) + \ln\left(N\right) \times \left(\text{number of parameters}\right).$$

Keribin (2000) shows the reliability of the BIC in a context of mixture model. The hyper-parameters are many $(s, l, t, G, H_1, \ldots, H_G)$, so that using the BIC to compare all their combinations on a cross-product grid is out of the question in practice. We choose instead to study the effects of varying the hyper-parameters following a heuristic. Even if these

parameters have different purposes, which can to some extent be dealt with sequentially, they are not completely independent. For instance: the higher s is, the higher H_g is likely to be. In practice, we propose the following heuristic: first, we perform an optimization on the hyper-parameters s and l with standard SCGLR (e.g. without mixture) on a grid $(s, l) \in \{0.1, 0.3, 0.5\} \times \{1, 2, 3, 4, 5\}$, calculating only one component. In a second step, we chose the number of groups by varying G from 1 to 5, keeping s and l fixed to their previously optimized values, still calculating a single component. The decision of distinguishing the groups only through their first component is justified by the simulation study presented in Section 3.2.2. Next, we implement forward selection to determine a suitable number of components in each group. We add one component in each group alternatively, and then choose the combination minimizing the BIC. We repeat that until the BIC rises. Finally, we vary the hyper-parameter t in $\{0.1, 0.2 \ldots, 0.9\}$, subject to the constraint $s + t \leq 1$, in order to better separate the components which might cause confusion between groups.

3.2 Simulation study

Two simulation studies have been implemented to assess the performance of rmSCGLR. The first one, presented in Section 3.2.1, focuses on the identification of response groups in a case of high correlations between latent variables spanning the explanatory spaces. In this simulation, we first present the component combination found by the previous heuristic for $s \in \{0.1, 0.3, 0.5\}$. Then we study the determination of the best value of hyper-parameter t. In Section 3.2.2, we present a simulation, in which we study the recovering of the true numbers of components, in a context of low correlation between the latent variables. In both simulations, we set $l = 4$ in order to facilitate the interpretation of components. For more information on the effects of hyper-parameters s and l, we refer the reader to Chauvet et al. (2019) and Bry et al. (2020a,b). The R package **rmSCGLR** and the simulation codes are available at https://github.com/julien-gibaud/rmSCGLR.

In the simulation study, we use the Rand Index (RI, Rand, 1971) and the Adjusted Rand Index (ARI, Hubert and Arabie, 1985) to assess the correctness of the classification decisions. In addition, to measure the quality of the latent variables recovery, we calculate the square correlation between the latent variable $\boldsymbol{\xi}$ and the components

$$\rho^2(\boldsymbol{\xi}, .) = \max_{g,h} \rho\left(\boldsymbol{\xi}, \boldsymbol{f}_g^h\right)^2,$$

where $\boldsymbol{f}_g^h$ denotes the hth component of group g. The RI, ARI, square correlation and BIC are all given through mean values over a hundred samples.

3.2.1 Generation of the simulated data

The variables are simulated on $N = 100$ observations. Two latent variables $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ are simulated with a correlation $\rho = 0.9$, while two others, $\boldsymbol{\xi_3}$ and $\boldsymbol{\xi_4}$, are simulated independent of any other. The $\boldsymbol{X}$ matrix consists in five blocks: $\boldsymbol{X} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{X_5}]$, where $\boldsymbol{X_1} \in \mathbb{R}^{N\times 20}$, $\boldsymbol{X_2} \in \mathbb{R}^{N\times 20}$, $\boldsymbol{X_3} \in \mathbb{R}^{N\times 10}$ and $\boldsymbol{X_4} \in \mathbb{R}^{N\times 10}$ are bundles of variables distributed about $\boldsymbol{\xi_1}$, $\boldsymbol{\xi_2}$, $\boldsymbol{\xi_3}$ and $\boldsymbol{\xi_4}$ respectively. More formally, for all $i = 1, \ldots, 4$, a variable $\boldsymbol{x_p}$ is simulated as $\boldsymbol{x_p} = \boldsymbol{\xi}_i + \boldsymbol{\varepsilon_p}$, where $\boldsymbol{\varepsilon_p} \sim \mathcal{N}_N(0, 0.1\boldsymbol{I_N})$. The $\boldsymbol{X_5}$ block contains 40 unstructured noise variables constructed as $\boldsymbol{x_p} \sim \mathcal{N}_N(0, \boldsymbol{I_N})$. The response matrix $\boldsymbol{Y}$ is partitioned into two groups of responses only distinguished by their explanatory latent variables. The first group consists of Poisson and Gaussian responses whose linear predictors are combinations of $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_3}$, while the second group gathers Gaussian and binary responses with linear predictors combining $\boldsymbol{\xi_2}$ and $\boldsymbol{\xi_4}$. The matrix $\boldsymbol{Y}$ is generated as

$$
\begin{aligned}
\forall k = 1, \ldots, 20, \quad & \boldsymbol{y_k} \sim \mathcal{N}_N\left(\boldsymbol{\mu} = \boldsymbol{\gamma_{1k}}\boldsymbol{\xi_1} + \boldsymbol{\gamma_{2k}}\boldsymbol{\xi_3}, \boldsymbol{\Sigma} = \boldsymbol{I_N}\right), \\
\forall k = 21, \ldots, 70, \quad & \boldsymbol{y_k} \sim \mathcal{P}\left(\boldsymbol{\lambda} = \exp\left[0.25\boldsymbol{\gamma_{1k}}\boldsymbol{\xi_1} + 0.25\boldsymbol{\gamma_{2k}}\boldsymbol{\xi_3}\right]\right), \\
\forall k = 71, \ldots, 80, \quad & \boldsymbol{y_k} \sim \mathcal{N}_N\left(\boldsymbol{\mu} = \boldsymbol{\gamma_{1k}}\boldsymbol{\xi_2} + \boldsymbol{\gamma_{2k}}\boldsymbol{\xi_4}, \boldsymbol{\Sigma} = \boldsymbol{I_N}\right), \\
\forall k = 81, \ldots, 100, \quad & \boldsymbol{y_k} \sim \mathcal{B}\left(\boldsymbol{p} = \text{logit}^{-1}\left[\boldsymbol{\gamma_{1k}}\boldsymbol{\xi_2} + \boldsymbol{\gamma_{2k}}\boldsymbol{\xi_4}\right]\right),
\end{aligned}
$$

where for all k, $\boldsymbol{\gamma_{1k}}$ and $\boldsymbol{\gamma_{2k}}$ are uniformly generated, with $\boldsymbol{\gamma_{1k}} \in [-4, 4]$ and $\boldsymbol{\gamma_{2k}} \in [-2, 2]$.

The purpose of this simulation scheme is to mix different types of response distributions, modeled through explanatory dimensions specific to response groups which must be recovered. Explanatory variables are many, and exhibit both bundles of highly redundant variables and isolated variables. Such a data structure is often encountered in practice when no pre-selection of explanatory variables has been carried out, and causes difficulties in modeling and estimation, which our method intends to solve.

3.2.1.1 Results and interpretation

Table 3.1 sums up the heuristic performed to find the best component combination for $s \in \{0.1, 0.3, 0.5\}$. We observe that the three values of s lead to detect two groups of responses. In this simulation, taking a higher value of s is not recommended. Indeed, as s increases, the components get closer to the principal components (Bry et al., 2020a). Thus, for $s > 0.5$, the first component of each group being drawn towards the same first principal component, they tend to be similar. This similarity hinders the distinction between groups. Performing a forward selection step and opting for the minimal value of the BIC, we see that only $s = 0.1$ and $s = 0.3$ lead to the right combination of components. However, $s = 0.5$ leads to identify the true overall number of directions central to the explanatory bundles. Thus, in the sequel, the analysis is done with combinations $(H_1, H_2) = (2, 2)$ for $s = 0.1$ and $s = 0.3$, while we set $(H_1, H_2) = (3, 1)$ for $s = 0.5$.

Table 3.1: Mean values of the BIC over a hundred samples, for a high correlation value ($\rho = 0.9$) between the latent variables $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$, for $s \in \{0.1, 0.3, 0.5\}$ and different combinations of H_1 and H_2.

$s = 0.1$			$s = 0.3$			$s = 0.5$		
H_1	H_2	BIC	H_1	H_2	BIC	H_1	H_2	BIC
1	1	30802.37	1	1	31281.99	1	1	31862.52
2	1	29577.02	2	1	29896.88	2	1	30416.06
1	2	29538.69	1	2	29821.90	1	2	30431.90
2	**2**	**29091.21**	**2**	**2**	**29549.27**	**3**	**1**	**29593.27**
1	3	29513.46	1	3	29561.49	2	2	29811.69
3	2	30030.12	3	2	30296.23	4	1	30054.50
2	3	30108.98	2	3	30292.89	3	2	30450.21

On the last step of the heuristic, summed up in Table 3.2, we can see the impact of the hyper-parameter t. For $s = 0.1$, the RI and the ARI increase as t goes from 0 to 0.4, and then decrease. Our criterion allows to distinguish two sub-spaces close to one another: for $t = 0.4$, the RI and the ARI values are respectively equal to 0.883 and 0.764 despite the high correlation between the first latent variables of the two groups. These observations are consistent with the BIC which decreases from $t = 0$ to $t = 0.4$. When t is too high, the RI and the ARI decrease, while the BIC increases, as observed for $t \geq 0.5$. In such cases, the weight on the separation sub-criterion φ is too heavy, and prevents the first components of the two groups to be close enough, which precludes the correct identification of the latent variables, hence of the groups. As a result, the square correlations between the rank-1 components and the corresponding latent variables are lower than 0.9 for $t \geq 0.4$. Moreover, when $t \geq 0.5$, the correlations $\rho^2(\boldsymbol{\xi_3}, .)$ and $\rho^2(\boldsymbol{\xi_4}, .)$ are higher than $\rho^2(\boldsymbol{\xi_1}, .)$ and $\rho^2(\boldsymbol{\xi_2}, .)$. The reason for this is that for such a high value of t as 0.5, $\boldsymbol{\xi_3}$ and $\boldsymbol{\xi_4}$ are found before $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$, because they provide more separated explanatory spaces. For $s = 0.3$, we observe, likewise, that the best values of RI and ARI, corresponding to the minimal value of the BIC, are reached for $t = 0.2$ but they are lower than that in the $s = 0.1$ case. As noticed by Chauvet et al. (2019), the thinner the bundles, the greater the value of s has to be, to recover the latent variables correctly. Here, indeed, the error variance being low ($\sigma^2 = 0.1$), the square correlations are, on the whole, greater for $s = 0.3$ than for $s = 0.1$. As in the case $s = 0.1$, $\rho^2(\boldsymbol{\xi_1}, .)$ and $\rho^2(\boldsymbol{\xi_2}, .)$ decrease with t. However, contrary to the case $s = 0.1$, the increase of the square correlations $\rho^2(\boldsymbol{\xi_3}, .)$ and $\rho^2(\boldsymbol{\xi_4}, .)$ could not be observed, since t could not exceed 0.6. In the $s = 0.5$ case, we observe the dramatic effect of taking too many components in a group. For all values of t, the RI and the ARI are respectively close to 0.5 and 0. This indicates that for $s = 0.5$ and $(H_1, H_2) = (3, 1)$, the obtained classification is not better than a random one.

For the sake of visualization, Figure 3.1 shows the correlation scatterplots of plane $(1, 2)$ for each group. We can see that the components $\boldsymbol{f}$ are well aligned with the corre-

Table 3.2: Mean values of RI, ARI, square correlation and BIC over a hundred samples, for a high correlation value ($\rho = 0.9$) between the latent variables $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$, for $s \in \{0.1, 0.3, 0.5\}$, the optimized combination of components and t ranging from 0 to 0.8.

				group 1		group 2		
s	t	RI	ARI	$\rho^2(\boldsymbol{\xi_1}, .)$	$\rho^2(\boldsymbol{\xi_3}, .)$	$\rho^2(\boldsymbol{\xi_2}, .)$	$\rho^2(\boldsymbol{\xi_4}, .)$	BIC
	0	0.860	0.718	0.971	0.950	0.963	0.927	29095.04
	0.1	0.861	0.721	0.970	0.951	0.955	0.907	29085.84
	0.2	0.865	0.729	0.966	0.939	0.938	0.888	28963.32
	0.3	0.870	0.738	0.931	0.889	0.913	0.878	28955.93
0.1	**0.4**	**0.883**	**0.764**	**0.899**	**0.889**	**0.893**	**0.874**	**28950.69**
	0.5	0.873	0.745	0.857	0.878	0.858	0.847	29531.91
	0.6	0.853	0.705	0.835	0.859	0.856	0.861	29705.92
	0.7	0.844	0.684	0.841	0.907	0.853	0.881	30302.17
	0.8	0.693	0.378	0.788	0.934	0.865	0.900	31805.47
	0	0.799	0.595	0.958	0.967	0.957	0.927	29497.32
	0.1	0.814	0.626	0.956	0.967	0.956	0.939	29493.86
	0.2	**0.815**	**0.629**	**0.957**	**0.956**	**0.965**	**0.970**	**29489.26**
0.3	0.3	0.812	0.623	0.957	0.958	0.955	0.959	29518.38
	0.4	0.796	0.591	0.951	0.958	0.950	0.951	29519.18
	0.5	0.794	0.589	0.919	0.915	0.917	0.911	29528.79
	0.6	0.792	0.582	0.813	0.795	0.814	0.815	29532.73
	0	0.560	0.039	0.948	0.918	0.948	0.911	29581.08
	0.1	0.572	0.054	0.945	0.896	0.951	0.897	29518.82
0.5	0.2	0.562	0.044	0.943	0.872	0.949	0.883	29541.87
	0.3	0.556	0.033	0.945	0.902	0.947	0.902	29531.97
	0.4	0.551	0.025	0.945	0.938	0.945	0.910	29606.57

sponding simulated latent variables $\boldsymbol{\xi}$, except for $\boldsymbol{f_2^2}$, which slightly deviates from $\boldsymbol{\xi_4}$. Due to the high correlation between $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$, the bundles $\boldsymbol{X_1}$ (in red) and $\boldsymbol{X_2}$ (in blue) are both well aligned with the first component of each group.

Finally, keeping the response groups obtained with the hyper-parameter values minimizing the BIC ($s = 0.1$ and $t = 0.4$), we go through the component optimization phase by performing SCGLR on each group separately. The square correlations of these final components with the latent variables are the following: $\rho^2(\boldsymbol{\xi_1}, .) = 0.971$, $\rho^2(\boldsymbol{\xi_2}, .) = 0.976$, $\rho^2(\boldsymbol{\xi_3}, .) = 0.957$ and $\rho^2(\boldsymbol{\xi_4}, .) = 0.948$. As expected, the recovery of the latent variables is much better.

As reference values for comparison, we calculated the RI and ARI of the partitions output by, on the one hand, the R package **ClustOfVar**, employed to initialize our algorithm, and, on the other hand, the R package **ecomix** implementing the approach proposed

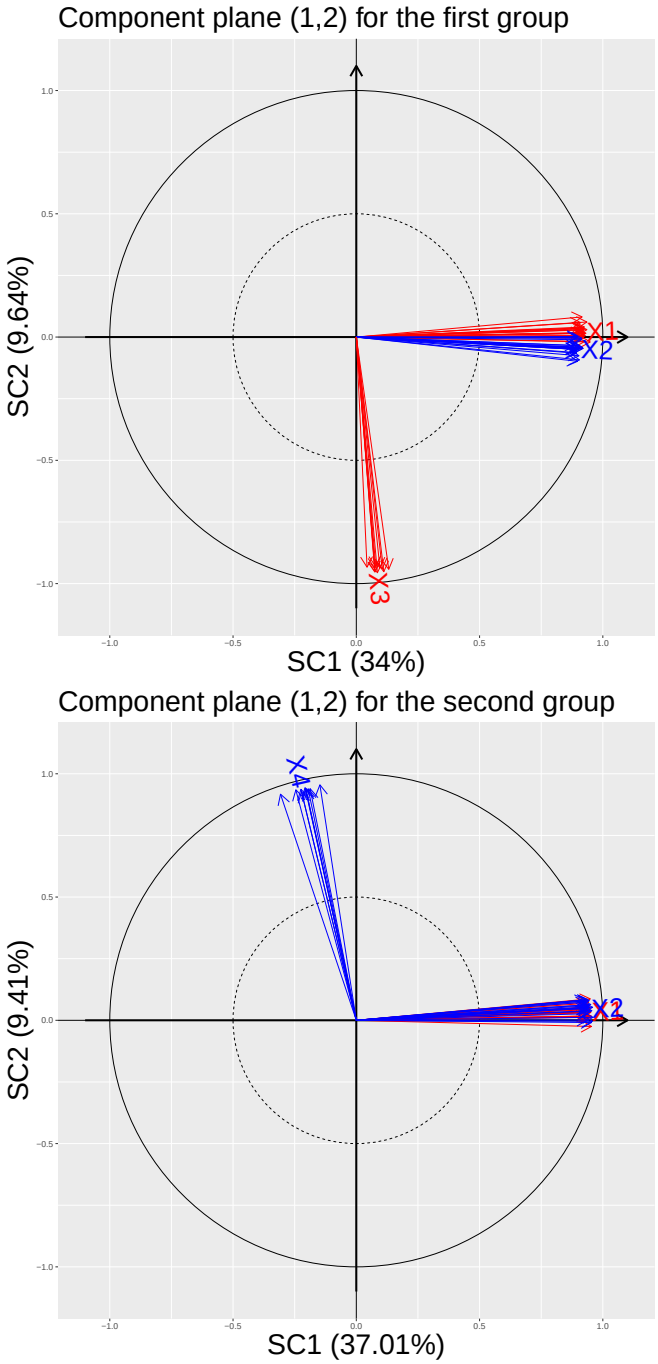


Figure 3.1: Correlation scatterplot of plane (1,2) for the two groups obtained by the rm-SCGLR algorithm with $s = 0.1$ and $t = 0.4$. The red arrows represent the bundles $\boldsymbol{X_1}$ and $\boldsymbol{X_3}$ which explain the first group. The blue ones represent the bundles $\boldsymbol{X_2}$ and $\boldsymbol{X_4}$ which explain the second group. The percentage of inertia captured by each component is given in parentheses.

by Dunstan et al. (2011, 2013). The computation time in seconds of the three packages is also eventually mentioned. However, **ecomix** not allowing to consider different distribution families for the responses, we restricted the comparison to the case of Gaussian responses. Thus, with the previous generated data, we have twenty responses in the first group and ten in the second one. Table 3.3 presents the results. As expected, in a context of component-based model, the **ecomix** classification does not outperform the random classification. The classification output by **ClustOfVar** is slightly better, but only provides a good starting point for **rmSCGLR**, which leads to high values of RI and ARI. We may note that **rmSCGLR** offers a greater classification performance than in the case of mixed distribution families. Even though **rmSCGLR** gives the best classification decisions, it is the slowest package, followed by **ecomix** and **ClustOfVar**.

Table 3.3: Mean values and standard deviations (in parentheses) of RI, ARI and computation time, in seconds, over a hundred samples for the R packages **rmSCGLR**, **ClustOfVar** and **ecomix**.

rmSCGLR		**ClustOfVar**		**ecomix**	
RI	0.964 (0.101)	RI	0.538 (0.070)	RI	0.507 (0.037)
ARI	0.929 (0.195)	ARI	0.104 (0.121)	ARI	0.045 (0.061)
Time	5.110 (2.359)	Time	0.192 (0.028)	Time	1.107 (0.197)

3.2.2 Varying the numbers of components

This simulation is devoted to recovering the true numbers of components, in a context of low correlation between the latent variables spanning the explanatory spaces. We assume unrealistically that the number of groups is known. s is fixed to 0.1, in order to study the behavior of the results when we vary the number of components per group and the weight t of the separation sub-criterion φ.

Three latent variables $\boldsymbol{\xi_1}$, $\boldsymbol{\xi_3}$ and $\boldsymbol{\xi_5}$ are simulated with a pairwise correlation $\rho = 0.5$. Two more latent variables $\boldsymbol{\xi_2}$ and $\boldsymbol{\xi_4}$ are independently simulated. The $\boldsymbol{X}$ matrix consists in six blocks $\boldsymbol{X} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{X_5}, \boldsymbol{X_6}]$, where $\boldsymbol{X_1} \in \mathbb{R}^{N\times 50}$, $\boldsymbol{X_2} \in \mathbb{R}^{N\times 40}$, $\boldsymbol{X_3} \in \mathbb{R}^{N\times 30}$, $\boldsymbol{X_4} \in \mathbb{R}^{N\times 20}$ and $\boldsymbol{X_5} \in \mathbb{R}^{N\times 10}$ are bundles aligned with $\boldsymbol{\xi_1}$, $\boldsymbol{\xi_2}$, $\boldsymbol{\xi_3}$, $\boldsymbol{\xi_4}$ and $\boldsymbol{\xi_5}$, respectively. The $\boldsymbol{X_6}$ block contains a set of 50 unstructured noise variables. The response matrix $\boldsymbol{Y}$ is partitioned into three groups of responses. The first group is composed of Gaussian responses, the expectations of which are linear combinations of $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_4}$. The second group gathers Poisson responses whose linear predictors are combinations of $\boldsymbol{\xi_2}$ and $\boldsymbol{\xi_5}$. The third group is made of binary responses depending only on $\boldsymbol{\xi_3}$. The matrix

$\boldsymbol{Y}$ is generated as

$$\begin{aligned}
\forall k = 1, \ldots, 20, \quad & \boldsymbol{y_k} \sim \mathcal{N}_N \left(\boldsymbol{\mu} = \boldsymbol{\gamma_{1k}\xi_1} + \boldsymbol{\gamma_{2k}\xi_4}, \boldsymbol{\Sigma} = \boldsymbol{I_N}\right), \\
\forall k = 21, \ldots, 70, \quad & \boldsymbol{y_k} \sim \mathcal{P} \left(\boldsymbol{\lambda} = \exp\left[0.25\boldsymbol{\gamma_{1k}\xi_2} + 0.25\boldsymbol{\gamma_{2k}\xi_5}\right]\right), \\
\forall k = 71, \ldots, 100, \quad & \boldsymbol{y_k} \sim \mathcal{B} \left(\boldsymbol{p} = \text{logit}^{-1}\left[\boldsymbol{\gamma_{1k}\xi_3}\right]\right),
\end{aligned}$$

where for all k, $\boldsymbol{\gamma_{1k}}$ and $\boldsymbol{\gamma_{2k}}$ are uniformly simulated such that $|\boldsymbol{\gamma_{1k}}| \in [2, 4]$ and $|\boldsymbol{\gamma_{2k}}| \in [1, 2]$.

3.2.2.1 Results and interpretation

The results of rmSCGLR on this simulation are given in Table 3.4. H_g denotes the number of components calculated in group g, and several triplets $H = (H_1, H_2, H_3)$ are tried. For none of these do we observe a clear difference of the RI and ARI across values of t. This was expected, since in this simulation, the explanatory subspaces are only weakly redundant. So, the separation sub-criterion φ proves almost useless here, and has practically no impact on the results. For $(H_1, H_2, H_3) = (1, 1, 1)$, the lowest values of RI and ARI are respectively 0.980 and 0.958, without the help of rank $h > 1$ components. The first component of each group perfectly recovers the latent explanatory variable which has the largest effect in the linear predictor of its responses. No component is aligned with the latent variable $\boldsymbol{\xi_4}$. The latent variable $\boldsymbol{\xi_5}$ having a correlation of 0.5 with $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_3}$, we find that $\sqrt{\rho^2(\boldsymbol{\xi_5}, .)} \simeq 0.5$ for all values of t. Taking $(H_1, H_2, H_3) = (2, 2, 1)$ does not improve the RI and ARI. We notice that the latent variable $\boldsymbol{\xi_5}$ is not as well recovered as the other latent variables, owing to the small size of the $\boldsymbol{X_5}$ bundle. However, the BIC is considerably reduced, which illustrates the importance of taking the right number of components to correctly predict the responses. The last case, where $(H_1, H_2, H_3) = (1, 3, 1)$ highlights the importance of getting a truly explanatory and strong first component in each group, and of not calculating too many components in a group. Like in the former cases, the third group is perfectly recovered using the true number of explanatory components $H_3 = 1$. But some confusion arises between the first two groups. Indeed, the extra component $\boldsymbol{f_2^3}$ of the second group is drawn towards the heaviest bundle $\boldsymbol{X_1}$. Then, the responses predictable from $\boldsymbol{X_1}$ tend to be scattered between the first and the second groups instead of being assigned to the first one, which causes a decrease of RI and ARI. Furthermore, owing to the correlation between $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_5}$, the components of the second group cannot be properly aligned with these latent variables. When $t = 0.8$ the weight on the separation criterion φ is heavy enough to recover $\boldsymbol{\xi_1}$, $\boldsymbol{\xi_2}$ and $\boldsymbol{\xi_5}$ in the second group, and $\boldsymbol{\xi_4}$ in the first group. To sum up this simulation, we observe that the role played by the first component in recovering the groups is crucial. Indeed, in the first case, the groups are determined by the first component only. In the second case, their prediction is completed by further rank components. However, in the third case, we see that calculating too many components may lead to impede group recovery.

Table 3.4: Mean values of RI and square correlations between latent variables and supervised components, over a hundred samples, for a weak pairwise correlation value ($\rho = 0.5$) between the latent variables $\boldsymbol{\xi_1}$, $\boldsymbol{\xi_3}$ and $\boldsymbol{\xi_5}$, and for various numbers H_g of components per group.

				group 1		group 2		group 3	
H	t	RI	ARI	$\rho^2(\boldsymbol{\xi_1}, .)$	$\rho^2(\boldsymbol{\xi_4}, .)$	$\rho^2(\boldsymbol{\xi_2}, .)$	$\rho^2(\boldsymbol{\xi_5}, .)$	$\rho^2(\boldsymbol{\xi_3}, .)$	BIC
	0	0.992	0.983	0.971	0.030	0.980	0.309	0.976	33525
	0.1	0.986	0.970	0.962	0.037	0.978	0.303	0.969	33580
	0.2	0.985	0.967	0.965	0.033	0.976	0.317	0.972	33435
1	0.3	0.987	0.972	0.968	0.037	0.978	0.314	0.973	33577
1	0.4	0.991	0.980	0.971	0.032	0.980	0.297	0.975	33435
1	0.5	0.980	0.958	0.960	0.036	0.974	0.298	0.961	33612
	0.6	0.992	0.983	0.960	0.043	0.979	0.295	0.974	33631
	0.7	0.994	0.987	0.954	0.046	0.983	0.295	0.975	33837
	0.8	0.992	0.983	0.944	0.044	0.979	0.298	0.964	34304
	0	0.984	0.966	0.968	0.921	0.975	0.816	0.966	29945
	0.1	0.983	0.964	0.971	0.938	0.977	0.809	0.971	29878
	0.2	0.989	0.977	0.974	0.951	0.979	0.835	0.975	29838
2	0.3	0.994	0.988	0.974	0.952	0.981	0.865	0.978	29783
2	0.4	0.993	0.984	0.968	0.946	0.981	0.876	0.975	29936
1	0.5	0.991	0.981	0.957	0.934	0.981	0.856	0.972	30150
	0.6	0.984	0.966	0.944	0.928	0.976	0.844	0.960	30348
	0.7	0.997	0.993	0.932	0.946	0.983	0.864	0.976	30733
	0.8	0.983	0.965	0.916	0.925	0.973	0.827	0.971	31131
	0	0.878	0.750	0.871	0.264	0.945	0.514	0.978	30483
	0.1	0.874	0.742	0.856	0.214	0.956	0.506	0.965	30245
	0.2	0.859	0.712	0.858	0.230	0.970	0.555	0.932	30020
1	0.3	0.871	0.776	0.853	0.242	0.969	0.545	0.946	31090
3	0.4	0.868	0.724	0.839	0.370	0.961	0.580	0.980	30052
1	0.5	0.876	0.748	0.804	0.308	0.977	0.585	0.977	30322
	0.6	0.891	0.774	0.806	0.320	0.976	0.656	0.977	30815
	0.7	0.882	0.759	0.732	0.353	0.975	0.657	0.974	30572
	0.8	0.790	0.592	0.877	0.772	0.956	0.614	0.963	33790

Figure 3.2 shows the correlation scatterplots in the component planes $(1, 2)$ for the first two groups. As for the first simulation, the components are almost perfectly aligned with the explanatory bundles. Because of the weak correlation between $\boldsymbol{\xi_1}$, $\boldsymbol{\xi_3}$ and $\boldsymbol{\xi_5}$, the three bundles $\boldsymbol{X_1}$, $\boldsymbol{X_3}$ and $\boldsymbol{X_5}$ are visible on the same component for each of the two groups.

3.3 Analysis of the floristic ecology data

3.3.1 Data description

We apply rmSCGLR to the *CoForTaxa* dataset available on demand at http://dx.doi.org/10.18167/DVN1/UCNCA7. The sample we consider gives the abundances of $K = 193$ floristic taxa in the Congo basin rainforest over a $N = 1571$ 10×10-km^2 grid cells across central Africa. To predict abundances, we have $P = 24$ climatic variables and $Q = 3$ non-climatic additional variables. $\boldsymbol{X}$ consists of all the climatic variables, i.e.: eleven temperature variables coded "C1",...,"C11", eight precipitation variables coded "C12",...,"C19", three climatic water deficit variables coded "sumCWD", "maxCWD" and "MCWD" respectively, one climatic water balance coded "meanCWB" and one evapotranspiration variable coded "meanET0". Figure 3.3 shows the correlation plot given by the PCA of the climatic variables. Since it appears that the explanatory variables exhibit a clear bundle structure, a methodology such as SCGLR is necessary to regularize the model estimation and reduce the dimension of the explanatory space. Besides, the non-climatic variables, i.e. the soil type (Harmonized World Soil Database, "HWSD") and the human-induced forest-disturbance intensity index ("Anthr2"), as well as its logarithm ("logAnthr2") to account for nonlinear effects, are few and weakly correlated with the variables in $\boldsymbol{X}$ as well as between themselves, and interesting per se. We shall then consider them as additional explanatory variables, and gather them in matrix $\boldsymbol{A}$. The response variables are assumed to be Poisson random variables, independent conditional on $\boldsymbol{X}$ and $\boldsymbol{A}$. Moreover, the variable corresponding to the number of plots within each grid cell is taken as the offset of the Poisson regression. For more information about the *CoForTaxa* dataset, we refer the reader to Réjou-Méchain et al. (2021).

3.3.2 Hyper-parameter calibration

We present the results obtained when following the parameter-varying scheme presented in Section 3.1.5. As noticed by Réjou-Méchain et al. (2021), the tuning parameters $s = 0.1$ and $l = 1$ allow to optimize SCGLR on *CoForTaxa* dataset. Here, thanks to the heuristic, $G = 3$ groups are retained to carry on with the analysis, using the previously found values of the tuning parameters. Starting with one component per group, we increment the number of components by one in each group alternately. Only adding one in the third group improves the criterion. When Réjou-Méchain et al. (2021) applied the basic SCGLR (without response mixture) to these data, three relevant components were found. The combination $H = (1, 1, 2)$ thus does not seem irrelevant. To get a refined model with this

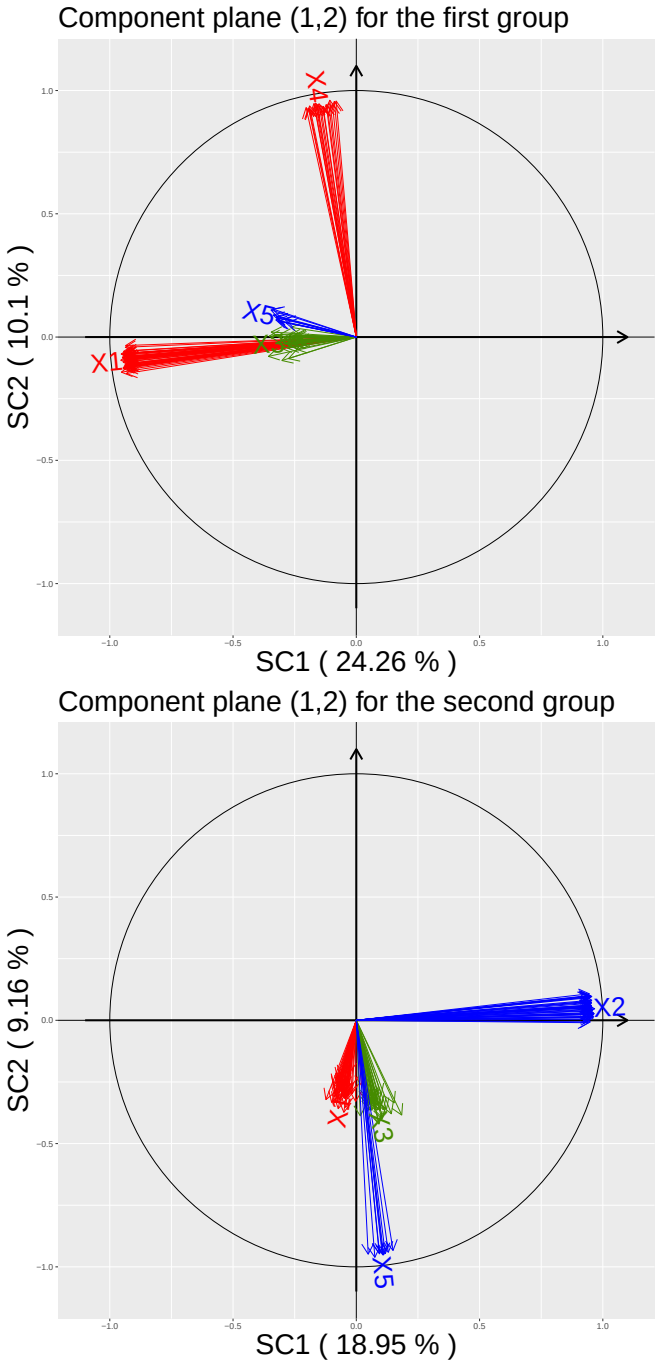


Figure 3.2: Correlation scatterplots of plane (1,2) for the first two groups obtained by rmSCGLR with $(H_1, H_2, H_3) = (2, 2, 1)$. The red arrows represent the bundles $\boldsymbol{X_1}$ and $\boldsymbol{X_4}$, explanatory of the first group. The blue ones represent the bundles $\boldsymbol{X_2}$ and $\boldsymbol{X_5}$, explanatory of the second group. The green bundle $\boldsymbol{X_3}$ is explanatory of the third group. The percentage of inertia captured by each component is given in parentheses.

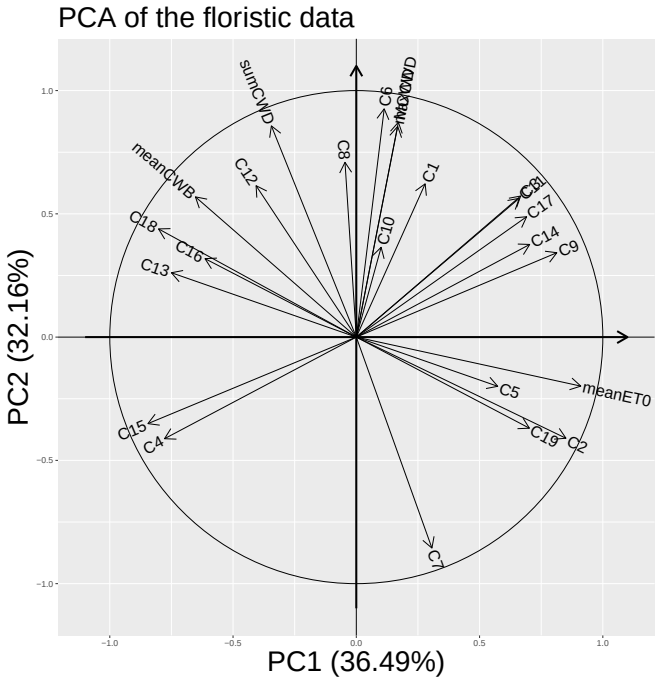


Figure 3.3: Component plane (1,2) of the explanatory climatic variables obtained through PCA. The percentage of inertia captured by each principal component is given in parentheses.

combination of components, the tuning parameter t needs to be raised to 0.5 to allow to better distinguish the groups, and minimize the BIC.

3.3.3 Results and interpretation

The clustering phase of rmSCGLR led to three groups of taxa. Two of them were associated with a single explanatory component, and the last one with two components. The groups respectively comprise 44, 67 and 82 taxa. The contents of the groups are given in Table 3.5.

Table 3.5: Here is the list of the taxa used in this study (the family classification follows Angiosperm Phylogeny Group III).

Group	Family	Genus	Species
1	Huaceae	Afrostyrax	lepidophyllus
1	Fabaceae	Afzelia	spp.
1	Fabaceae	Albizia	ferruginea
1	Fabaceae	Albizia	spp.
1	Gentianaceae	Anthocleista	spp.

Continued on next page

Continued from previous page

Group	Family	Genus	Species
1	Fabaceae	Anthonotha	spp.
1	Phyllanthaceae	Antidesma	spp.
1	Fabaceae	Aphanocalyx	spp.
1	Fabaceae	Aubrevillea	kerstingii
1	Zygophyllaceae	Balanites	wilsoniana
1	Passifloraceae	Barteria	spp.
1	Lauraceae	Beilschmiedia	spp.
1	Malvaceae	Bombax	spp.
1	Malvaceae	Ceiba	pentandra
1	Cannabaceae	Celtis	spp.
1	Sapotaceae	Chrysophyllum	spp.
1	Annonaceae	Cleistopholis	spp.
1	Malvaceae	Cola	spp.
1	Boraginaceae	Cordia	spp.
1	Fabaceae	Detarium	macrocarpum
1	Fabaceae	Dialium	spp.
1	Euphorbiaceae	Discoglypremna	caloneura
1	Malvaceae	Duboscia	spp.
1	Arecaceae	Elaeis	guineensis
1	Malvaceae	Eribroma	oblongum
1	Hypericaceae	Harungana	madagascariensis
1	Annonaceae	Hexalobus	spp.
1	Ulmaceae	Holoptelea	grandis
1	Meliaceae	Khaya	spp.
1	Irvingiaceae	Klainedoxa	spp.
1	Meliaceae	Lovoa	trichilioides
1	Malvaceae	Mansonia	altissima
1	Urticaceae	Myrianthus	arboreus
1	Apocynaceae	Picralima	nitida
1	Sapotaceae	Pouteria	spp.
1	Malvaceae	Pterygota	spp.
1	Euphorbiaceae	Ricinodendron	heudelotii
1	Malvaceae	Sterculia	spp.
1	Olacaceae	Strombosiopsis	spp.
1	Myrtaceae	Syzygium	spp.
1	Combretaceae	Terminalia	superba
1	Fabaceae	Tetrapleura	tetraptera
1	Malvaceae	Triplochiton	scleroxylon
1	Lamiaceae	Vitex	spp.

Continued on next page

Continued from previous page

Group	Family	Genus	Species
2	Clusiaceae	Allanblackia	spp.
2	Apocynaceae	Alstonia	spp.
2	Fabaceae	Angylocalyx	spp.
2	Anisophylleaceae	Anisophyllea	spp.
2	Moraceae	Antiaris	toxicaria
2	Fabaceae	Aubrevillea	platycarpa
2	Burseraceae	Aucoumea	klaineana
2	Sapotaceae	Autranella	congolensis
2	Sapotaceae	Baillonella	toxisperma
2	Fabaceae	Bikinia	spp.
2	Sapindaceae	Blighia	spp.
2	Sapotaceae	Breviea	sericea
2	Burseraceae	Canarium	schweinfurthii
2	Myristicaceae	Coelocaryon	spp.
2	Rubiaceae	Corynanthe	pachyceras
2	Fabaceae	Cylicodiscus	gabunensis
2	Burseraceae	Dacryodes	spp.
2	Fabaceae	Daniellia	spp.
2	Achariaceae	Dasylepis	seretii
2	Malvaceae	Desplatsia	spp.
2	Ebenaceae	Diospyros	crassiflora
2	Fabaceae	Distemonanthus	benthamianus
2	Meliaceae	Entandrophragma	angolense
2	Meliaceae	Entandrophragma	candollei
2	Meliaceae	Entandrophragma	cylindricum
2	Meliaceae	Entandrophragma	utile
2	Vochysiaceae	Erismadelphus	exsul
2	Bignoniaceae	Fernandoa	adolfi
2	Moraceae	Ficus	spp.
2	Fabaceae	Gilbertiodendron	spp.
2	Euphorbiaceae	Gymnanthes	inopinata
2	Irvingiaceae	Irvingia	grandifolia
2	Lepidobotryaceae	Lepidobotrys	staudtii
2	Euphorbiaceae	Macaranga	spp.
2	Rhamnaceae	Maesopsis	eminii
2	Sapotaceae	Manilkara	spp.
2	Phyllanthaceae	Margaritaria	discoidea
2	Moraceae	Milicia	excelsa
2	Moraceae	Morus	mesozygia
2	Urticaceae	Musanga	cecropioides

Continued on next page

Continued from previous page

Group	Family	Genus	Species
2	Rubiaceae	Nauclea	spp.
2	Malvaceae	Nesogordonia	spp.
2	Fabaceae	Newtonia	spp.
2	Picrodendraceae	Oldfieldia	africana
2	Salicaceae	Oncoba	spp.
2	Chrysobalanaceae	Parinari	spp.
2	Fabaceae	Pentaclethra	eetveldeana
2	Euphorbiaceae	Plagiostyles	africana
2	Combretaceae	Pteleopsis	hylodendron
2	Fabaceae	Pterocarpus	spp.
2	Violaceae	Rinorea	spp.
2	Burseraceae	Santiria	spp.
2	Oleaceae	Schrebera	arborea
2	Myristicaceae	Scyphocephalium	mannii
2	Anacardiaceae	Sorindeia	spp.
2	Clusiaceae	Symphonia	globulifera
2	Sapotaceae	Synsepalum	spp.
2	Ochnaceae	Testulea	gabonensis
2	Fabaceae	Tetraberlinia	bifoliolata
2	Euphorbiaceae	Tetrorchidium	didymostemon
2	Sapotaceae	Tieghemella	africana
2	Moraceae	Treculia	spp.
2	Meliaceae	Trichilia	spp.
2	Anacardiaceae	Trichoscypha	spp.
2	Sapotaceae	Tridesmostemon	omphalocarpoides
2	Moraceae	Trilepisium	madagascariense
2	Dipterocarpaceae	Trillesanthus	excelsus
3	Fabaceae	Amphimas	spp.
3	Annonaceae	Annickia	spp.
3	Annonaceae	Anonidium	mannii
3	Rhizophoraceae	Anopyxis	klaineana
3	Euphorbiaceae	Anthostema	aubryanum
3	Anacardiaceae	Antrocaryon	spp.
3	Fabaceae	Berlinia	spp.
3	Fabaceae	Bobgunnia	fistuloides
3	Fabaceae	Brachystegia	spp.
3	Rubiaceae	Brenania	brieyi
3	Phyllanthaceae	Bridelia	spp.
3	Fabaceae	Calpocalyx	spp.

Continued on next page

Continued from previous page

Group	Family	Genus	Species
3	Meliaceae	Carapa	spp.
3	Sapotaceae	Chrysophyllum	lacourtianum
3	Fabaceae	Copaifera	spp.
3	Olacaceae	Coula	edulis
3	Euphorbiaceae	Croton	spp.
3	Fabaceae	Cryptosepalum	spp.
3	Olacaceae	Diogoa	zenkeri
3	Ebenaceae	Diospyros	spp.
3	Asparagaceae	Dracaena	spp.
3	Putranjivaceae	Drypetes	spp.
3	Annonaceae	Duguetia	spp.
3	Fabaceae	Erythrophleum	spp.
3	Erythroxylaceae	Erythroxylum	mannii
3	Fabaceae	Eurypetalum	spp.
3	Fabaceae	Fillaeopsis	discophora
3	Apocynaceae	Funtumia	spp.
3	Clusiaceae	Garcinia	spp.
3	Fabaceae	Gilbertiodendron	dewevrei
3	Malvaceae	Grewia	spp.
3	Salicaceae	Homalium	spp.
3	Fabaceae	Hylodendron	gabunense
3	Fabaceae	Hymenostegia	spp.
3	Irvingiaceae	Irvingia	spp.
3	Fabaceae	Julbernardia	spp.
3	Phyllanthaceae	Keayodendron	bridelioides
3	Meliaceae	Leplaea	spp.
3	Sapotaceae	Letestua	durissima
3	Ochnaceae	Lophira	alata
3	Calophyllaceae	Mammea	africana
3	Chrysobalanaceae	Maranthes	spp.
3	Bignoniaceae	Markhamia	spp.
3	Fabaceae	Millettia	spp.
3	Rubiaceae	Morinda	lucida
3	Fabaceae	Neochevalierodendron	stephanii
3	Ochnaceae	Ochna	spp.
3	Ixonanthaceae	Ochthocosmus	spp.
3	Sapotaceae	Omphalocarpum	spp.
3	Olacaceae	Ongokea	gore
3	Fabaceae	Pachyelasma	tessmannii
3	Pandaceae	Panda	oleosa

Continued on next page

Continued from previous page

Group	Family	Genus	Species
3	Rubiaceae	Pausinystalia	spp.
3	Fabaceae	Pentaclethra	macrophylla
3	Fabaceae	Pericopsis	elata
3	Lecythidaceae	Petersianthus	macrocarpus
3	Fabaceae	Piptadeniastrum	africanum
3	Annonaceae	Polyalthia	suaveolens
3	Fabaceae	Prioria	spp.
3	Anacardiaceae	Pseudospondias	spp.
3	Myristicaceae	Pycnanthus	angolensis
3	Simaroubaceae	Quassia	spp.
3	Apocynaceae	Rauvolfia	spp.
3	Rubiaceae	Rothmannia	spp.
3	Euphorbiaceae	Sapium	spp.
3	Fabaceae	Scorodophloeus	zenkeri
3	Achariaceae	Scottellia	spp.
3	Lecythidaceae	Scytopetalum	klaineanum
3	Bignoniaceae	Spathodea	campanulata
3	Fabaceae	Stachyothyrsus	staudtii
3	Myristicaceae	Staudtia	kamerunensis
3	Fabaceae	Stemonocoleus	micranthus
3	Combretaceae	Strephonema	spp.
3	Olacaceae	Strombosia	spp.
3	Apocynaceae	Tabernaemontana	spp.
3	Fabaceae	Tessmannia	spp.
3	Fabaceae	Tetraberlinia	polyphylla
3	Phyllanthaceae	Uapaca	spp.
3	Annonaceae	Xylopia	aethiopica
3	Annonaceae	Xylopia	hypolampra
3	Annonaceae	Xylopia	quintasii
3	Rutaceae	Zanthoxylum	spp.

Let us first try to interpret the groups and components output by the clustering phase of rmSCGLR. We sum up the first two groups in Table 3.6, stating the explanatory variables most correlated with the components. Table 3.6 does not deal with the third group, as this one appears in the sequel to be something of a "junk" group with no homogeneous interpretation.

The component of the first group is highly correlated with the variable "C7" (difference between the maximum of temperature of the warmest month and the minimum of tempera-

Table 3.6: Lists of explanatory variables most correlated with the component in each of the first two groups. Only correlations over 0.8 in absolute value are given.

Groups	Explanatory variables	Correlation
1	C7, sumCWD, MCWD, maxCWD,	0.956, 0.955, 0.885, 0.880
2	C2, meanET0, C18, C19	0.930, 0.929, 0.925, 0.862

ture of the coldest month), and with the three climatic water deficit variables: "sumCWD", "MCWD" and "maxCWD". Thus, the abundances of taxa composing the first group would be linked to a gradient of temperature, and sensitive to a water deficit. The component of the second group is highly correlated with "C2" (the mean diurnal range), "meanET0" (the mean monthly evapotranspiration) and with "C18" and "C19" (the precipitations of the warmest quarter and the coldest quarter, respectively). This component is very similar to the first component found if we apply SCGLR on all the responses ($\rho = -0.965$). According to Réjou-Méchain et al. (2021), this component is highly related to a regional floristic gradient contrasting areas with a cool and light-deficient dry season (coastal Gabon) and areas with high evapotranspiration rates (northern limit of the central African forests). The components of the third group fail to be aligned with any bundle of variables. The corresponding scatterplot is given in Figure 3.4. As mentioned by Réjou-Méchain et al. (2021), a majority of taxon abundances may relate with climate only by chance. Thus, by contrast to the first and second group, where the abundances are linked to water deficit or precipitation, the taxa composing the third group are not connected with any specific gradient but with various combinations of climatic variables.

In the optimization phase, SCGLR is performed separately on each group. In the first group, SCGLR finds a single component, highly correlated ($\rho = 0.960$) with $\boldsymbol{f}_1^1$ of the clustering phase. Three components are calculated by SCGLR to best predict the second group. However, on Figure 3.5a, we can see that all the linear predictors of the taxa's abundances composing the second group are highly correlated with the bundle found by $\boldsymbol{f}_2^1$ of the clustering phase. The second and third components only provide a secondary improvement in predicting the abundances. The correlation between the first SCGLR-component of the second group and rmSCGLR's $\boldsymbol{f}_2^1$ is equal to -0.991. As expected for the third group of taxa, Figure 3.5b shows no particular correlation pattern between the linear predictors and any bundle, which highlights the absence of specific climatic gradient in this group's explanatory space. The planes spanned by the higher rank components are respectively given by Figure 3.6 and Figure 3.7.

Let us evaluate the benefits obtained in the prediction by taking into account the clustering found by rmSCGLR. In Réjou-Méchain et al. (2021), the quality of prediction was given by the mean of ten-fold cross-validation Mean Squared Prediction Errors (MSPE), and we shall use the same index for comparison. We shall thus compare: (i) the prediction error we get with SCGLR on all taxa, named MSPE_{all}, (ii) the prediction error obtained

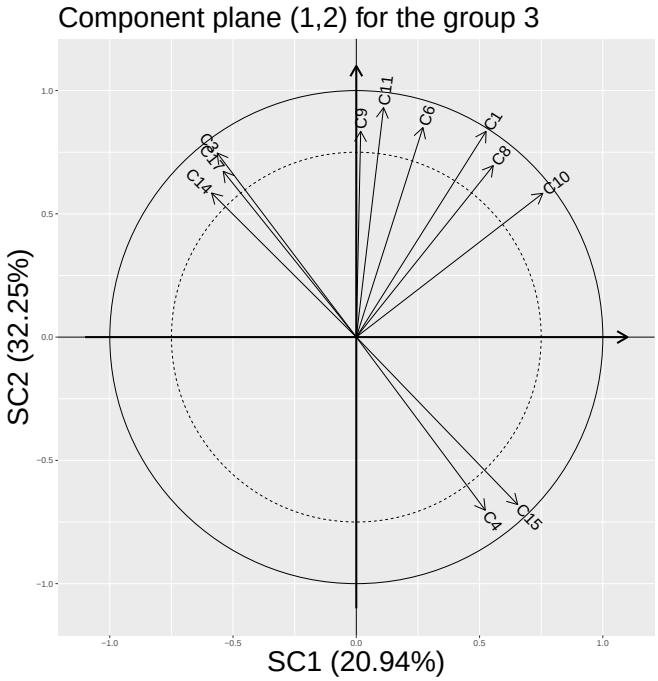
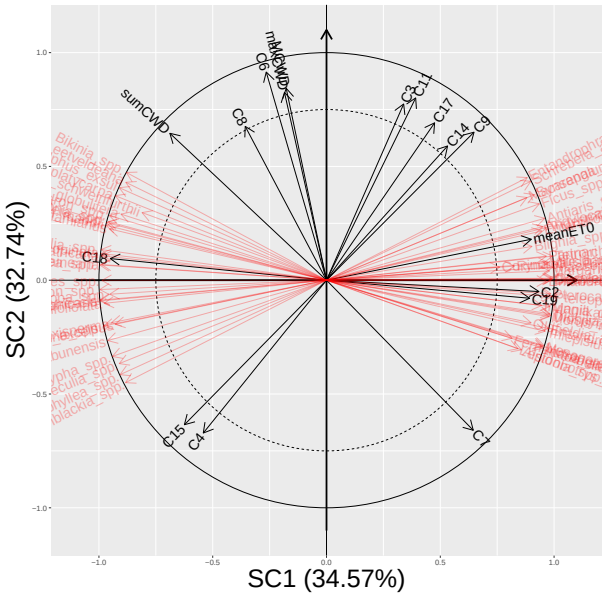
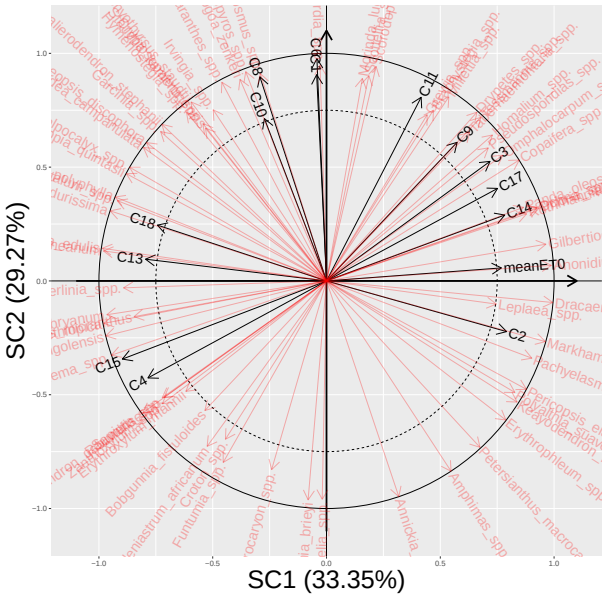


Figure 3.4: Component plane (1,2) for the group 3 output by rmSCGLR on the *CoForTaxa* dataset, with optimal hyper-parameter $(s, l, t) = (0.1, 1, 0.5)$. The plot displays only variables having cosine greater than 0.75. The percentage of inertia captured by each component is given in parentheses.

with SCGLR on the three groups separately, named MSPE_1, MSPE_2 and MSPE_3 respectively, with their weighted mean named $\text{MSPE}_{\text{mean}}$, and (iii) the mean of the prediction error on random partitions into three groups of taxa, obtained over a hundred samples, named $\text{MSPE}_{\text{random}}$. The prediction error of SCGLR on all taxa was calculated by Réjou-Méchain et al. (2021), and found to be $\text{MSPE}_{\text{all}} = 3.23$ (1.13). SCGLR, performed separately on the first and second groups, gave the following prediction errors: $\text{MSPE}_1 = 3.07$ (0.87) and $\text{MSPE}_2 = 2.94$ (1.07) respectively, which indicates an improved quality of prediction. However, the prediction error of the third group rises to $\text{MSPE}_3 = 3.41$ (0.99), which indicates that group 3 is composed by taxa the abundances of which are poorly predictable from the sheer observed climatic variables. Finally, the mean prediction error of SCGLR accounting for the partition is: $\text{MSPE}_{\text{mean}} = 3.17$ (0.99). The mean prediction error accounting for a random three-group partition is: $\text{MSPE}_{\text{random}} = 3.20$ (1.09). This shows that rmSCGLR was able to, if only slightly, better capture the explanatory structure of the floristic data. It should be noted that prediction of taxa abundances from merely such climatic variables is usually poor (Beale et al., 2008).



(a) Component plane $(1, 2)$ for the group 2



(b) Component plane $(1, 2)$ for the group 3

Figure 3.5: Correlation scatterplots of plane (1,2) with linear predictors for the second and third separated groups obtained by the SCGLR algorithm. The black arrows represent the covariates. The red ones are the linear predictors of the responses. The plot displays only variables having a cosine over 0.75. The percentage of inertia captured by each component is given in parentheses.

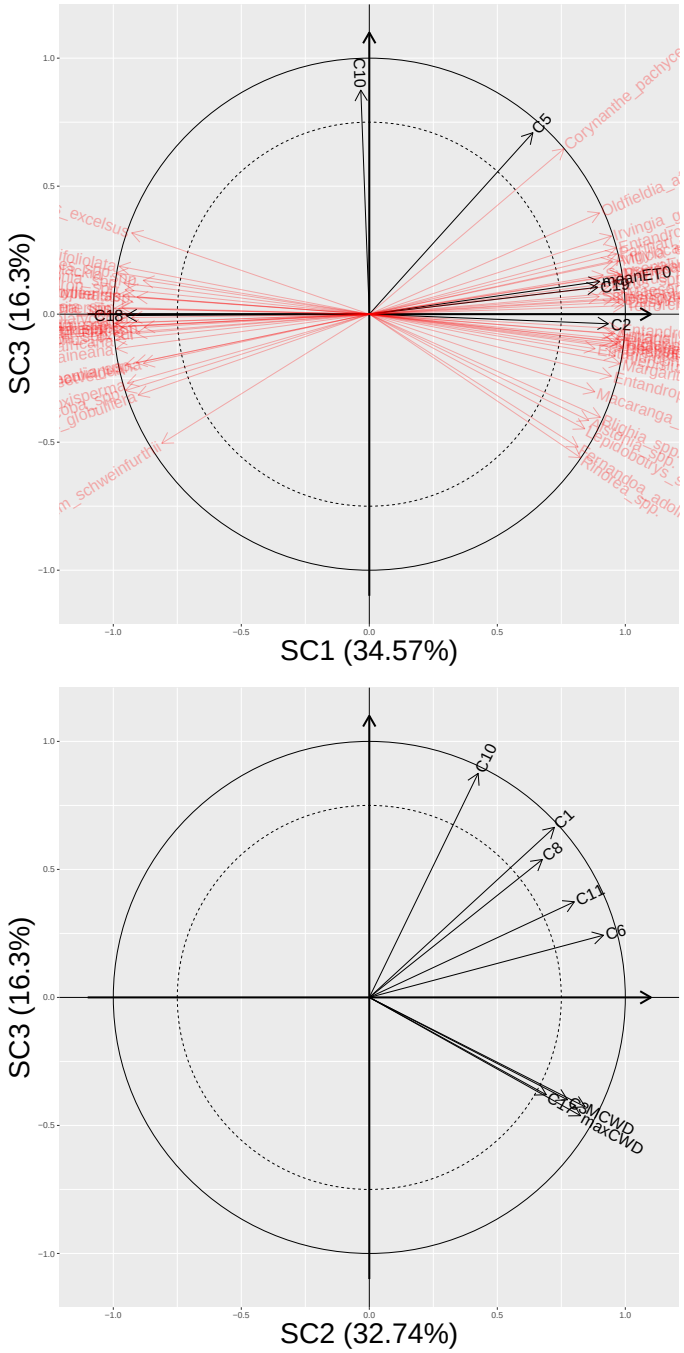


Figure 3.6: Correlation scatterplots of planes (1,3) and (2,3) with linear predictors obtained by applying SCGLR to the second group separately. The blacks arrows represent the covariates. The red ones represent the linear predictors. The plot displays only variables having a cosine greater than 0.75 with the plane. The percentage of inertia captured by each component is given in parentheses.

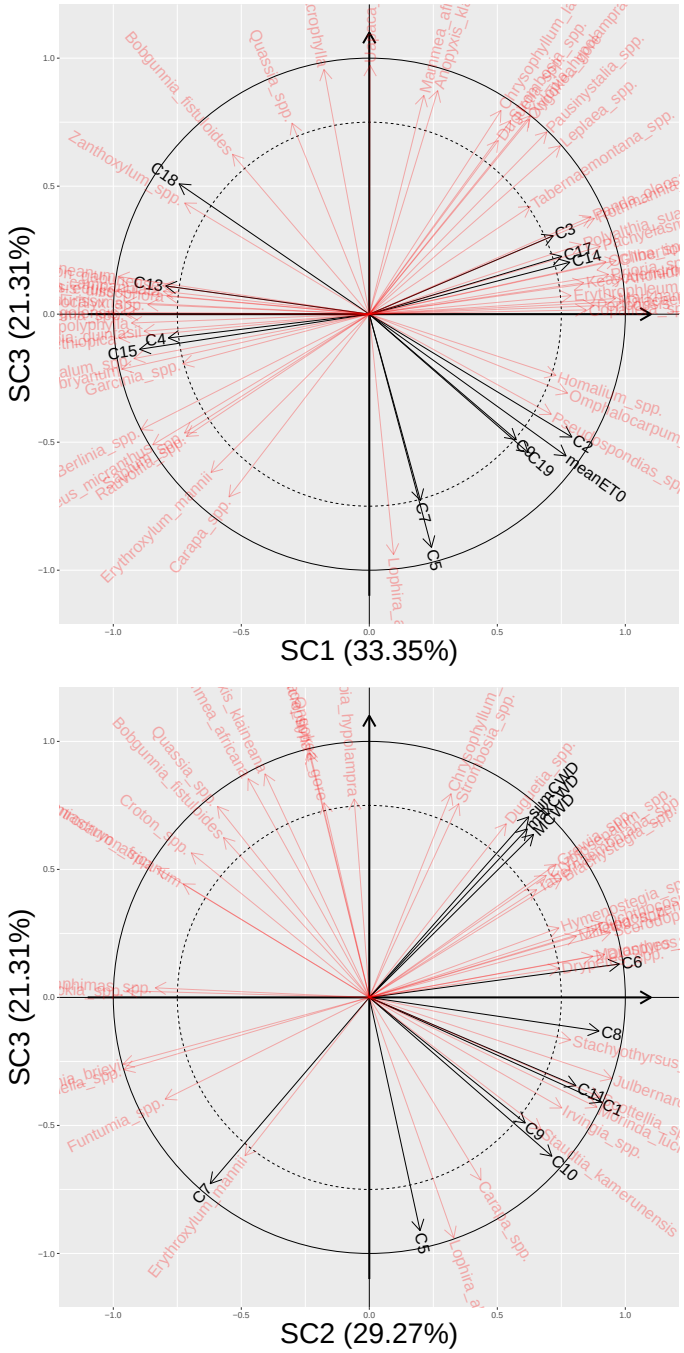


Figure 3.7: Correlation scatterplots of planes (1,3) and (2,3) with linear predictors obtained by applying SCGLR to the third group separately. The blacks arrows represent the covariates. The red ones represent the linear predictors. The plot displays only variables having a cosine greater than 0.75 with the plane. The percentage of inertia captured by each component is given in parentheses.

3.4 Conclusion and discussion

In the context we address, we have multiple responses to be modeled through many covariates. All responses may not depend on the same explanatory dimensions, captured by components. Therefore, we both need to model the responses and to cluster them with respect to their common explanatory components. Unfortunately, no available method jointly performs response clustering and search for explanatory components. Among the methods searching for common explanatory components, the original SCGLR was designed to regularize GLM estimation and reduce the explanatory space through components, so as to decompose the linear predictor in an interpretable way. It allowed to find strong and interpretable supervised components common to response variables, by achieving a trade-off between Goodness-of-Fit and a Structural Relevance measure. Methods as proposed by Dunstan et al. (2011, 2013) or Mortier et al. (2015) cluster responses by imposing that the regression coefficients of the covariates be the same within each cluster, which does not allow to model responses in a flexible enough manner. Moreover, their modeling is not based on strong dimensions as components. The response mixture SCGLR extends SCGLR in two major ways: (i) Through a mixture model on the response variables, it identifies groups of responses that can be predicted from group-specific components. Doing so, this method improves both the prediction quality of the response groups, and the interpretation of what explains the responses. In our ecological framework, we detected communities of taxa sensitive to specific gradients of climate variables. (ii) It extends the criterion to be maximized by introducing a separation sub-criterion, which allows to specify sub-spaces which components had better keep away from. In the context of response mixture, this sub-criterion helped distinguish the groups by better separating their explanatory sub-spaces.

In our simulation study, rmSCGLR proved to behave as expected regarding groups. In a context of very close explanatory sub-spaces, it recovered the original groups, and provided components aligned with the latent variables. On the floristic ecology dataset, we found three communities of taxa. The first one is linked to a gradient of temperature, while the second one is connected to a regional floristic gradient contrasting two main areas. The last group gathers the taxa related to no specific gradient, but to many combinations of the observed climatic variables. More predictive climatic components could likely be generated after removing these taxa.

Our method still has some limitations. Just as the original SCGLR, it does not allow to deal with a thematic partition of the explanatory variables. To overcome this limitation, we could extend THEME-SCGLR (Bry et al., 2020b) to a response mixture. For instance, the temperature and precipitation variables would be seen as pertaining to two distinct themes and each community of taxa would be predicted by common components in each theme. Another way of extending our model would be to create sparse components, in the spirit of Durif et al. (2018), with intent to select relevant climatic variables. Another limitation

is that the heuristic presented in Section 3.1.5 does not guarantee to find the best values of the hyper-parameters. Several parameter-varying schemes could be implemented and the results compared. Hutter et al. (2015) propose a review of works allowing to best optimize the hyper-parameters.

CHAPTER 4

GENERALIZED LINEAR LATENT VARIABLE MODEL BASED ON SUPERVISED COMPONENTS

Contents

This chapter is inspired by a work in progress which should be submitted soon in a statistical journal.

4.1 Relaxing the independence hypothesis

This part is dedicated to the relaxation of the independence hypothesis in a component-based model framework. Section 4.1.1 recalls the method to calculate successive components for an extension of SCGLR to multiple explanatory variable subsets: THEME-SCGLR. The factor model and the THEME-SCGLR are combined in Section 4.1.2. Section 4.1.3 presents the estimation process we use: Sub-section 4.1.3.1 linearizes the model, while Sub-section 4.1.3.2 assumes Gaussian distributions on the pseudo-variables. The EM algorithm we develop is presented in Section 4.1.4. More particularly, Sub-sections 4.1.4.1 and 4.1.4.2 detail the expectation and maximization steps respectively, while the explicit EM algorithm is given in Sub-section 4.1.4.3. The overall algorithm is shown in Section 4.1.5. Finally, Section 4.1.6 gives the posterior clustering steps used to detect the groups.

4.1.1 Reminder of THEME-SCGLR

We recall that this chapter is developed in the THEME-SCGLR context as defined in Section 2.2.4. Let $\boldsymbol{f}_r^h = \boldsymbol{X}_r \boldsymbol{u}_r^h$ be the rank-h component of theme $\boldsymbol{X}_r$, and let $\boldsymbol{F}_r^h = [\boldsymbol{f}_r^1, \ldots, \boldsymbol{f}_r^h]$, where $h \leq H_r$, be the matrix of the first h components for this theme. According to the local nesting principle, the new component $\boldsymbol{f}_r^{h+1}$ must best complement both the existing ones and $\boldsymbol{A}$, that is $\boldsymbol{A}_r^h := [\boldsymbol{F}_1^{H_1}, \ldots, \boldsymbol{F}_{r-1}^{H_{r-1}}, \boldsymbol{F}_r^h, \boldsymbol{F}_{r+1}^{H_{r+1}}, \ldots, \boldsymbol{F}_R^{H_R}, \boldsymbol{A}]$. So $\boldsymbol{f}_r^{h+1}$ has to be calculated using $\boldsymbol{A}_r^h$ as the new set of additional covariates. Moreover, to avoid linear redundancy of components, we impose that $\boldsymbol{f}_r^{h+1}$ be orthogonal to $\boldsymbol{F}_r^h$, *i.e.* $\boldsymbol{F}_r^{h\,T} \boldsymbol{W} \boldsymbol{f}_r^{h+1} = \boldsymbol{0}$. The sub-criteria ϕ and $\psi_{\boldsymbol{A}_r}$ are respectively given by Equation (2.9) and Equation (2.10), while the weight α_k reflecting the *a priori* relative importance of working variable w_k is set to 1 for all k. We calculate every new component as the solution of the optimization program given by Equation (2.11), with the additional constraint: $\boldsymbol{\Delta}_r^{h\,T} \boldsymbol{u}_r^{h+1} = \boldsymbol{0}$, where $\boldsymbol{\Delta}_r^h = \boldsymbol{X}_r^T \boldsymbol{W} \boldsymbol{F}_r^h$, and loop on r until overall convergence of the component system. For all $r = 1, \ldots, R$, the rank-1 component of theme $\boldsymbol{X}_r$ is calculated using the same program with $\boldsymbol{F}_r^0 = \emptyset$ and $\boldsymbol{\Delta}_r^0 = \boldsymbol{0}$. For the sake of simplicity, in the following, we shall consider the matrix $\boldsymbol{F} = [\boldsymbol{F}_1^{H_1}, \ldots, \boldsymbol{F}_R^{H_R}]$ as the new set of explanatory variables and $\boldsymbol{\gamma_k} = (\gamma_{k1}, \ldots, \gamma_{kR})^T$ its vector of regression parameters associated with the response $\boldsymbol{y_k}$.

4.1.2 THEME-SCGLR in a factor model context

Let $\boldsymbol{Y} = [\boldsymbol{y_1}, \ldots, \boldsymbol{y_K}] \in \mathbb{R}^{N \times K}$ be the response matrix. For unit n, each response is assumed to be linearly modeled using the components and additional covariates, plus J random latent factors $\boldsymbol{g_n} = (g_{n1}, \ldots, g_{nJ})^T$

$$\eta_{nk} = \boldsymbol{f_n^T \gamma_k} + \boldsymbol{a_n^T \delta_k} + \boldsymbol{g_n^T b_k},$$

where $\boldsymbol{f_n}$ and $\boldsymbol{a_n}$ are the vectors composed of the nth rows of matrices $\boldsymbol{F}$ and $\boldsymbol{A}$ respectively, and $\boldsymbol{b_k}$ is the vector of regression parameters associated with $\boldsymbol{g_n}$. The factors are

assumed drawn from a multivariate normal distribution $\boldsymbol{g_n} \sim \mathcal{N}_J(0, \boldsymbol{I_J})$ and independent across statistical units. This model is designed so that the J factors capture as much as possible of the covariance between the responses not accounted for by the components and additional covariates, hence their conditional covariance. Denoting $\boldsymbol{G} \in \mathbb{R}^{N \times J}$ the matrix containing all the realizations of factors, the linear predictor associated with the response $\boldsymbol{y_k}$ expressed column-wise becomes

$$\boldsymbol{\eta_k} = \boldsymbol{F}\boldsymbol{\gamma_k} + \boldsymbol{A}\boldsymbol{\delta_k} + \boldsymbol{G}\boldsymbol{b_k}.$$

Let $\boldsymbol{B} = [\boldsymbol{b_1}, \ldots, \boldsymbol{b_K}] \in \mathbb{R}^{J \times K}$ be the loading matrix. Jöreskog (1969) noticed that the loading matrix $\boldsymbol{B}$ is defined up to an arbitrary orthogonal rotation. To guarantee the identification of the model, we choose to constrain the $J \times J$ sub-matrix of $\boldsymbol{B}$ to be an upper triangular matrix with positive diagonal elements (Geweke and Zhou, 1996). An advantage of the factor model is to yield the matrix $\boldsymbol{\Sigma} = \boldsymbol{B^T B} \in \mathbb{R}^{K \times K}$, storing the conditional covariance of the responses, in a parsimonious manner. Indeed, the number of factors retained may remain small with respect to the size of the covariance matrix. For the sake of clarity, Figure 4.1 presents the path diagram of THEME-SCGLR with factors.

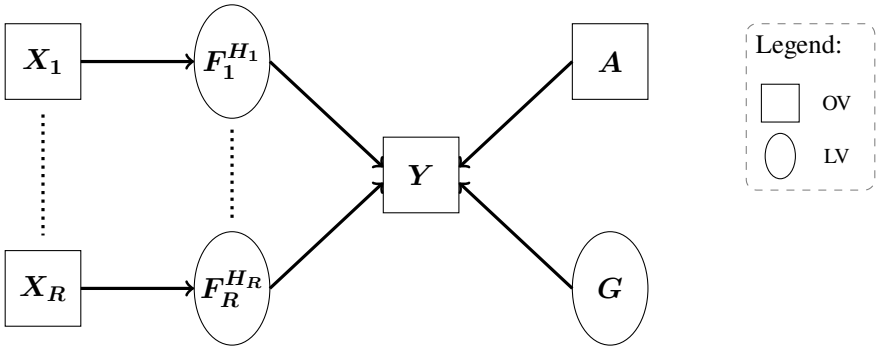


Figure 4.1: Path diagram of THEME-SCGLR with latent factors. The observed variables (OV) are presented in squares while the latent variables (LV) are shown in ovals. The arrows represent the influence links.

4.1.3 Estimating the parameters of a GLM with factors

Let $\boldsymbol{\Theta} = \{\boldsymbol{\gamma_k}, \boldsymbol{\delta_k}, \boldsymbol{b_k}; k = 1, \ldots, K\}$ be the set of parameters. The marginal log-likelihood of the model is obtained by integrating over latent variables $\boldsymbol{g_n}$

$$\begin{aligned} l(\boldsymbol{\Theta}; \boldsymbol{Y}) &= \sum_{n=1}^{N} \ln\left(L(\boldsymbol{y_n}; \boldsymbol{\Theta})\right) \\ &= \sum_{n=1}^{N} \ln\left(\int \prod_{k=1}^{K} L(y_{nk}|\boldsymbol{g_n}; \boldsymbol{\Theta}) L(\boldsymbol{g_n})\, d\boldsymbol{g_n}\right). \end{aligned}$$

In a context of non-Gaussian responses, the maximization of this log-likelihood is not allowed. In the spirit of Saidane et al. (2013), the estimation of the parameters is performed in two steps: first, we linearize the model; then, we maximize the pseudo-likelihood of the linearized model under a Gaussian assumption.

4.1.3.1 The linearization step

Temporarily considering the factors given, i.e. conditional on $\boldsymbol{G}$, the above log-likelihood is that of a classic multivariate GLM. Let h_k denote the canonical link function associated with the response $\boldsymbol{y_k}$, h'_k its first derivative and μ_{nk} the mean parameter for statistical unit n. The working variable w_{nk} associated with y_{nk} is then calculated as the first order development of h_k at point μ_{nk}

$$\begin{aligned} w_{nk} &= h_k\left(\mu_{nk}\right) + \left(y_{nk} - \mu_{nk}\right) h'_k\left(\mu_{nk}\right) \\ &= \eta_{nk} + \zeta_{nk}, \end{aligned}$$

where $\zeta_{nk} = \left(y_{nk} - \mu_{nk}\right) h'_k\left(\mu_{nk}\right)$. This development leads to the conditional linearized model expressed column-wise

$$\boldsymbol{w_k} = \boldsymbol{F\gamma_k} + \boldsymbol{A\delta_k} + \boldsymbol{Gb_k} + \boldsymbol{\zeta_k},$$

where $\mathbb{E}[\boldsymbol{w_k}|\boldsymbol{G}] = \boldsymbol{F\gamma_k} + \boldsymbol{A\delta_k} + \boldsymbol{Gb_k}$ and $\mathbb{V}[\boldsymbol{w_k}|\boldsymbol{G}] = \mathbb{V}[\boldsymbol{\zeta_k}] = \boldsymbol{W_k^{-1}}$.

4.1.3.2 The estimation step

In this step, we assume that the distribution of the working variables given $\boldsymbol{F}$, $\boldsymbol{A}$ and $\boldsymbol{G}$ is Gaussian, and view the factors as latent variables. The model pseudo-log-likelihood $l(\boldsymbol{\Theta}; \boldsymbol{\mathcal{W}})$, where $\boldsymbol{\mathcal{W}}$ denotes the matrix of working variables, being difficult to maximize directly, we use the EM algorithm to estimate the model parameters. We calculate and then maximize the expectation of the complete log-likelihood $l(\boldsymbol{\Theta}; \boldsymbol{\mathcal{W}}, \boldsymbol{G})$ of the working variables. Further details of the EM algorithm are given in Section 4.1.4.

4.1.4 The EM algorithm for a GLM with factors

We are now dealing with the linearized model, where the factors are latent. So, we shall use the EM algorithm to estimate the parameters. The previous developments lead to the conditional linearized model

$$\boldsymbol{w_k} = \boldsymbol{F\gamma_k} + \boldsymbol{A\delta_k} + \boldsymbol{Gb_k} + \boldsymbol{\zeta_k},$$

where $\mathbb{E}[\boldsymbol{w_k}|\boldsymbol{G}] = \boldsymbol{F\gamma_k} + \boldsymbol{A\delta_k} + \boldsymbol{Gb_k}$ and

$$\mathbb{V}[\boldsymbol{w_k}|\boldsymbol{G}] = \mathbb{V}[\boldsymbol{\zeta_k}] = \boldsymbol{W_k^{-1}} = \operatorname{diag}\left(v_{nk}^{-1}\right)_{n=1,\ldots,N},$$

with $\upsilon_{nk}^{-1} := a_{nk}(\phi_k)v_k\left(\mu_{nk}\right)h_k'\left(\mu_{nk}\right)^2$, where a_{nk} and v_k are known functions and ϕ_k is the dispersion parameter related to $\boldsymbol{y_k}$. The linearized model expressed row-wise thus writes

$$\boldsymbol{w_n} = \boldsymbol{\Gamma^T f_n} + \boldsymbol{\Delta^T a_n} + \boldsymbol{B^T g_n} + \boldsymbol{\zeta_n},$$

where $\boldsymbol{\Gamma} = [\boldsymbol{\gamma_1}, \ldots, \boldsymbol{\gamma_K}]$, $\boldsymbol{\Delta} = [\boldsymbol{\delta_1}, \ldots, \boldsymbol{\delta_K}]$, $\boldsymbol{B} = [\boldsymbol{b_1}, \ldots, \boldsymbol{b_K}]$, and where $\boldsymbol{w_n}$, $\boldsymbol{f_n}$, $\boldsymbol{a_n}$ and $\boldsymbol{g_n}$ are the vectors composed of the nth rows of matrices $\boldsymbol{\mathcal{W}}$, $\boldsymbol{F}$, $\boldsymbol{A}$ and $\boldsymbol{G}$ respectively. The expectation and the variance are given by $\mathbb{E}[\boldsymbol{w_n}] = \boldsymbol{\Gamma^T f_n} + \boldsymbol{\Delta^T a_n}$ and $\mathbb{V}[\boldsymbol{w_n}] = \boldsymbol{B^T B} + \boldsymbol{\Upsilon_n^{-1}}$, where

$$\boldsymbol{\Upsilon_n^{-1}} = \text{diag}\left(\upsilon_{nk}^{-1}\right)_{k=1,\ldots,K}.$$

Denoting $\boldsymbol{\Theta} = \{\boldsymbol{\Gamma}, \boldsymbol{\Delta}, \boldsymbol{B}\}$ the set of parameters, the complete log-likelihood writes

$$\begin{aligned}
&l(\boldsymbol{\Theta}; \boldsymbol{\mathcal{W}}, \boldsymbol{G}) \\
&= \ln\left(L(\boldsymbol{\mathcal{W}}, \boldsymbol{G}; \boldsymbol{\Theta})\right) \\
&= \sum_{n=1}^{N} \ln\left(L(\boldsymbol{w_n}|\boldsymbol{g_n}; \boldsymbol{\Theta})\right) + \ln\left(L(\boldsymbol{g_n}; \boldsymbol{\Theta})\right) \\
&= \sum_{n=1}^{N} \Bigg[-\ln\left((2\pi)^{K/2}\det\left(\boldsymbol{\Upsilon_n^{-1}}\right)^{1/2}\right) \\
&\qquad - \frac{1}{2}\left(\boldsymbol{w_n} - \boldsymbol{\Gamma^T f_n} - \boldsymbol{\Delta^T a_n} - \boldsymbol{B^T g_n}\right)^T \boldsymbol{\Upsilon_n}\left(\boldsymbol{w_n} - \boldsymbol{\Gamma^T f_n} - \boldsymbol{\Delta^T a_n} - \boldsymbol{B^T g_n}\right) \\
&\qquad - \ln\left((2\pi)^{J/2}\right) - \frac{1}{2}\boldsymbol{g_n^T g_n}\Bigg] \\
&= -\frac{1}{2}\sum_{n=1}^{N}\Bigg[\sum_{k=1}^{K}\ln\left(\upsilon_{nk}^{-1}\right) + \boldsymbol{g_n^T g_n} + (K+J)\ln(2\pi) \\
&\qquad + \sum_{k=1}^{K}\upsilon_{nk}\left(w_{nk} - \boldsymbol{f_n^T\gamma_k} - \boldsymbol{a_n^T\delta_k} - \boldsymbol{g_n^T b_k}\right)^2\Bigg].
\end{aligned}$$

4.1.4.1 The expectation (E) step

We first calculate the expectation of the complete log-likelihood conditional on the data $\boldsymbol{\mathcal{W}}$

$$\mathbb{E}[l\left(\boldsymbol{\Theta}; \boldsymbol{\mathcal{W}}, \boldsymbol{G}\right)|\boldsymbol{\mathcal{W}}; \boldsymbol{\Theta'}] = \sum_{n=1}^{N}\int \ln\left(L\left(\boldsymbol{w_n}|\boldsymbol{g_n}; \boldsymbol{\Theta}\right)L\left(\boldsymbol{g_n}; \Theta\right)\right)L\left(\boldsymbol{g_n}|\boldsymbol{w_n}; \boldsymbol{\Theta'}\right)d\boldsymbol{g_n}.$$

Thus, we need to first find the law of $\boldsymbol{g_n}|\boldsymbol{w_n}$. Since the random vector $(\boldsymbol{w_n^T}, \boldsymbol{g_n^T})^T$ is Gaussian, we have

$$\begin{pmatrix}\boldsymbol{w_n}\\ \boldsymbol{g_n}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\Gamma^T f_n} + \boldsymbol{\Delta^T a_n}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{B^T B} + \boldsymbol{\Upsilon_n^{-1}} & \boldsymbol{B^T}\\ \boldsymbol{B} & \boldsymbol{I_J}\end{pmatrix}\right).$$

Thanks to the conditioning rule of the multivariate Gaussian , we get

$$\boldsymbol{g_n}|\boldsymbol{w_n} \sim \mathcal{N}\left(\boldsymbol{\alpha_n}\left(\boldsymbol{w_n} - \boldsymbol{\Gamma^T f_n} - \boldsymbol{\Delta^T a_n}\right), \boldsymbol{I_J} - \boldsymbol{\alpha_n B^T}\right),$$

where $\boldsymbol{\alpha_n} = \boldsymbol{B}(\boldsymbol{B^T B} + \boldsymbol{\Upsilon_n^{-1}})^{-1}$. The moments of the random variable $\boldsymbol{g_n}|\boldsymbol{w_n}$ are given by

$$\begin{aligned}\tilde{\boldsymbol{g}}_{\boldsymbol{n}} &:= \mathbb{E}\left[\boldsymbol{g_n}|\boldsymbol{w_n}; \boldsymbol{\Theta}\right] \\ &= \boldsymbol{\alpha_n}\left(\boldsymbol{w_n} - \boldsymbol{\Gamma^T f_n} - \boldsymbol{\Delta^T a_n}\right)\end{aligned}$$

and

$$\begin{aligned}\tilde{\boldsymbol{R}}_{\boldsymbol{n}} &:= \mathbb{E}\left[\boldsymbol{g_n g_n^T}|\boldsymbol{w_n}; \boldsymbol{\Theta}\right] \\ &= \mathbb{V}\left[\boldsymbol{g_n}|\boldsymbol{w_n}; \boldsymbol{\Theta}\right] + \mathbb{E}\left[\boldsymbol{g_n}|\boldsymbol{w_n}; \boldsymbol{\Theta}\right]\mathbb{E}\left[\boldsymbol{g_n}|\boldsymbol{w_n}; \boldsymbol{\Theta}\right]^T \\ &= \boldsymbol{I_J} - \boldsymbol{\alpha_n B^T} + \tilde{\boldsymbol{g}}_{\boldsymbol{n}}\tilde{\boldsymbol{g}}_{\boldsymbol{n}}^{\boldsymbol{T}}.\end{aligned}$$

Finally, we have the explicit form of the expectation of the complete log-likelihood

$$\begin{aligned}&\mathbb{E}[l\left(\boldsymbol{\Theta}; \boldsymbol{\mathcal{W}}, \boldsymbol{G}\right)|\boldsymbol{\mathcal{W}}, \boldsymbol{\Theta'}] \\ &= -\frac{1}{2}\sum_{n=1}^{N}\Bigg\{(K+J)\ln(2\pi) + \sum_{k=1}^{K}\ln\left(\upsilon_{nk}^{-1}\right) + \\ &\quad \mathbb{E}\left[\boldsymbol{g_n^T g_n} + \sum_{k=1}^{K}\upsilon_{nk}\left(w_{nk} - \boldsymbol{f_n^T\gamma_k} - \boldsymbol{a_n^T\delta_k} - \boldsymbol{g_n^T b_k}\right)^2\Bigg|\boldsymbol{w_n}; \boldsymbol{\Theta'}\right]\Bigg\} \\ &= -\frac{1}{2}\sum_{n=1}^{N}\Bigg\{(K+J)\ln(2\pi) + \sum_{k=1}^{K}\ln\left(\upsilon_{nk}^{-1}\right) + \mathbb{E}\left[\boldsymbol{g_n^T g_n}|\boldsymbol{w_n}; \boldsymbol{\Theta'}\right] + \\ &\quad \mathbb{E}\Bigg[\sum_{k=1}^{K}\upsilon_{nk}\Big(\left(w_{nk} - \boldsymbol{f_n^T\gamma_k} - \boldsymbol{a_n^T\delta_k}\right)^2 + \boldsymbol{b_k^T}\left(\boldsymbol{g_n g_n^T}\right)\boldsymbol{b_k} - \\ &\quad 2\left(w_{nk} - \boldsymbol{f_n^T\gamma_k} - \boldsymbol{a_n^T\delta_k}\right)\boldsymbol{g_n^T b_k}\Big)\Big|\boldsymbol{w_n}; \boldsymbol{\Theta'}\Bigg]\Bigg\} \\ &= -\frac{1}{2}\sum_{n=1}^{N}\Bigg\{(K+J)\ln(2\pi) + \sum_{k=1}^{K}\ln\left(\upsilon_{nk}^{-1}\right) + \mathbb{E}\left[\boldsymbol{g_n^T g_n}|\boldsymbol{w_n}; \boldsymbol{\Theta'}\right] + \\ &\quad \sum_{k=1}^{K}\upsilon_{nk}\Big[\left(w_{nk} - \boldsymbol{f_n^T\gamma_k} - \boldsymbol{a_n^T\delta_k}\right)^2 + \boldsymbol{b_k^T}\tilde{\boldsymbol{R}}_{\boldsymbol{n}}\boldsymbol{b_k} - \\ &\quad 2\left(w_{nk} - \boldsymbol{f_n^T\gamma_k} - \boldsymbol{a_n^T\delta_k}\right)\tilde{\boldsymbol{g}}_{\boldsymbol{n}}^{\boldsymbol{T}}\boldsymbol{b_k}\Big]\Bigg\} \\ &= -\frac{1}{2}\Bigg\{N(K+J)\ln(2\pi) + \sum_{n=1}^{N}\sum_{k=1}^{K}\ln\left(\upsilon_{nk}^{-1}\right) + \sum_{n=1}^{N}\mathbb{E}\left[\boldsymbol{g_n^T g_n}|\boldsymbol{w_n}; \boldsymbol{\Theta'}\right] + \\ &\quad \sum_{k=1}^{K}\Bigg[\|\boldsymbol{w_k} - \boldsymbol{F\gamma_k} - \boldsymbol{A\delta_k}\|_{\boldsymbol{W_k}}^2 + \boldsymbol{b_k^T}\left(\sum_{n=1}^{N}\upsilon_{nk}\tilde{\boldsymbol{R}}_{\boldsymbol{n}}\right)\boldsymbol{b_k} - \\ &\quad 2\left(\tilde{\boldsymbol{G}}\boldsymbol{b_k}\right)^T\boldsymbol{W_k}\left(\boldsymbol{w_k} - \boldsymbol{F\gamma_k} - \boldsymbol{A\delta_k}\right)\Bigg]\Bigg\},\end{aligned}$$

where the rows of the matrix $\tilde{\boldsymbol{G}}$ are composed by $\tilde{\boldsymbol{g}}_n^T$'s.

4.1.4.2 The maximization (M) step

The maximization step maximizes the conditional expectation of the complete log-likelihood with respect to $\boldsymbol{\Theta}$, subject to the upper triangular constraint on matrix $\boldsymbol{B}$. However, for all k, the parameters $\boldsymbol{\gamma_k}$ and $\boldsymbol{\delta_k}$ are not concerned by the constraint. Denoting $\boldsymbol{\beta_k^T} = (\boldsymbol{\gamma_k^T}, \boldsymbol{\delta_k^T})$ and $\tilde{\boldsymbol{X}} = [\boldsymbol{F}, \boldsymbol{A}]$, the first order conditions of the maximization yield

$$
\begin{aligned}
&\nabla_{\boldsymbol{\beta_k}} \mathbb{E}[l(\boldsymbol{\Theta}; \boldsymbol{\mathcal{W}}, \boldsymbol{G}) | \boldsymbol{\mathcal{W}}, \boldsymbol{\Theta'}] = 0 \\
&\Leftrightarrow \nabla_{\boldsymbol{\beta_k}} \left\{ \left\| \boldsymbol{w_k} - \tilde{\boldsymbol{X}} \boldsymbol{\beta_k} \right\|_{\boldsymbol{W_k}}^2 - 2 \left(\tilde{\boldsymbol{G}} \boldsymbol{b_k} \right)^T \boldsymbol{W_k} \left(\boldsymbol{w_k} - \tilde{\boldsymbol{X}} \boldsymbol{\beta_k} \right) \right\} = 0 \\
&\Leftrightarrow \tilde{\boldsymbol{X}}^{\boldsymbol{T}} \boldsymbol{W_k} \left(\boldsymbol{w_k} - \tilde{\boldsymbol{X}} \boldsymbol{\beta_k} \right) - \tilde{\boldsymbol{X}}^{\boldsymbol{T}} \boldsymbol{W_k} \tilde{\boldsymbol{G}} \boldsymbol{b_k} = 0 \\
&\Leftrightarrow \tilde{\boldsymbol{X}}^{\boldsymbol{T}} \boldsymbol{W_k} \tilde{\boldsymbol{X}} \boldsymbol{\beta_k} = \tilde{\boldsymbol{X}}^{\boldsymbol{T}} \boldsymbol{W_k} \left(\boldsymbol{w_k} - \tilde{\boldsymbol{G}} \boldsymbol{b_k} \right) \\
&\Leftrightarrow \boldsymbol{\beta_k} = \left(\tilde{\boldsymbol{X}}^{\boldsymbol{T}} \boldsymbol{W_k} \tilde{\boldsymbol{X}} \right)^{-1} \tilde{\boldsymbol{X}}^{\boldsymbol{T}} \boldsymbol{W_k} \left(\boldsymbol{w_k} - \tilde{\boldsymbol{G}} \boldsymbol{b_k} \right).
\end{aligned}
$$

If a response is drawn from a Gaussian law $\boldsymbol{y_k} \sim \mathcal{N}_N \left(\tilde{\boldsymbol{X}} \boldsymbol{\beta_k}, \sigma_k^2 \boldsymbol{I_N} \right)$, the residual variance σ_k^2 must be estimated. Besides,

$$
\begin{aligned}
&\nabla_{\sigma_k^2} \mathbb{E}[l(\boldsymbol{\Theta}; \boldsymbol{\mathcal{W}}, \boldsymbol{G}) | \boldsymbol{\mathcal{W}}, \boldsymbol{\Theta'}] = 0 \\
&\Leftrightarrow \nabla_{\sigma_k^2} \left\{ N \ln \left(\sigma_k^2 \right) + \frac{1}{\sigma_k^2} \left[\left\| \boldsymbol{w_k} - \tilde{\boldsymbol{X}} \boldsymbol{\beta_k} \right\|^2 + \boldsymbol{b_k^T} \left(\sum_{n=1}^{N} \tilde{\boldsymbol{R}}_{\boldsymbol{n}} \right) \boldsymbol{b_k} \right. \right. \\
&\qquad \left. \left. -2 \left(\tilde{\boldsymbol{G}} \boldsymbol{b_k} \right)^T \left(\boldsymbol{w_k} - \tilde{\boldsymbol{X}} \boldsymbol{\beta_k} \right) \right] \right\} = 0 \\
&\Leftrightarrow N - \frac{1}{\sigma_k^2} \left\{ \left\| \boldsymbol{w_k} - \tilde{\boldsymbol{X}} \boldsymbol{\beta_k} \right\|^2 + \boldsymbol{b_k^T} \left(\sum_{n=1}^{N} \tilde{\boldsymbol{R}}_{\boldsymbol{n}} \right) \boldsymbol{b_k} - 2 \left(\tilde{\boldsymbol{G}} \boldsymbol{b_k} \right)^T \left(\boldsymbol{w_k} - \tilde{\boldsymbol{X}} \boldsymbol{\beta_k} \right) \right\} = 0 \\
&\Leftrightarrow \sigma_k^2 = \frac{1}{N} \left\{ \left\| \boldsymbol{w_k} - \tilde{\boldsymbol{X}} \boldsymbol{\beta_k} \right\|^2 + \boldsymbol{b_k^T} \left(\sum_{n=1}^{N} \tilde{\boldsymbol{R}}_{\boldsymbol{n}} \right) \boldsymbol{b_k} - 2 \left(\tilde{\boldsymbol{G}} \boldsymbol{b_k} \right)^T \left(\boldsymbol{w_k} - \tilde{\boldsymbol{X}} \boldsymbol{\beta_k} \right) \right\}.
\end{aligned}
$$

Now, we need to estimate the vector $\boldsymbol{b_k}$ under the upper triangular constraint. For each $k = 1, \ldots, J$, let $\boldsymbol{b_k^T} = (\boldsymbol{b_{1:k,k}^T}, \boldsymbol{0^T})$ be the regression parameters, where $\boldsymbol{b_{1:k,k}^T} = (b_{1k}, \ldots, b_{kk})$ is a vector of length k to be estimated and $\mathbf{0}$ is a null vector of length $(J-k)$ *a priori* fixed. In this case, we define $(\tilde{\boldsymbol{R}}_{\boldsymbol{n}})_{\boldsymbol{1:k,1:k}}$ as the sub-matrix of size $k \times k$ of $\tilde{\boldsymbol{R}}_{\boldsymbol{n}}$

and $\tilde{\boldsymbol{G}}_{\mathbf{1:}\boldsymbol{k}}$ as the matrix composed by the first k columns of $\tilde{\boldsymbol{G}}$. The maximization yields

$$
\begin{aligned}
&\nabla_{\boldsymbol{b}_{\mathbf{1:}\boldsymbol{k},\boldsymbol{k}}} \mathbb{E}[l(\boldsymbol{\Theta}; \boldsymbol{\mathcal{W}}, \boldsymbol{G}) | \boldsymbol{\mathcal{W}}, \boldsymbol{\Theta}'] = 0 \\
&\Leftrightarrow \nabla_{\boldsymbol{b}_{\mathbf{1:}\boldsymbol{k},\boldsymbol{k}}} \left\{ \boldsymbol{b}_{\mathbf{1:}\boldsymbol{k},\boldsymbol{k}}^{\boldsymbol{T}} \left[\sum_{n=1}^{N} \upsilon_{nk} \left(\tilde{\boldsymbol{R}}_{\boldsymbol{n}} \right)_{\mathbf{1:}\boldsymbol{k},\mathbf{1:}\boldsymbol{k}} \right] \boldsymbol{b}_{\mathbf{1:}\boldsymbol{k},\boldsymbol{k}} - 2 \left(\tilde{\boldsymbol{G}}_{\mathbf{1:}\boldsymbol{k}} \boldsymbol{b}_{\mathbf{1:}\boldsymbol{k},\boldsymbol{k}} \right)^{T} \boldsymbol{W}_{\boldsymbol{k}} \left(\boldsymbol{w}_{\boldsymbol{k}} - \tilde{\boldsymbol{X}} \boldsymbol{\beta}_{\boldsymbol{k}} \right) \right\} = 0 \\
&\Leftrightarrow \left(\tilde{\boldsymbol{G}}_{\mathbf{1:}\boldsymbol{k}} \right)^{T} \boldsymbol{W}_{\boldsymbol{k}} \left(\boldsymbol{w}_{\boldsymbol{k}} - \tilde{\boldsymbol{X}} \boldsymbol{\beta}_{\boldsymbol{k}} \right) - \left[\sum_{n=1}^{N} \upsilon_{nk} \left(\tilde{\boldsymbol{R}}_{\boldsymbol{n}} \right)_{\mathbf{1:}\boldsymbol{k},\mathbf{1:}\boldsymbol{k}} \right] \boldsymbol{b}_{\mathbf{1:}\boldsymbol{k},\boldsymbol{k}} = 0 \\
&\Leftrightarrow \boldsymbol{b}_{\mathbf{1:}\boldsymbol{k},\boldsymbol{k}} = \left[\sum_{n=1}^{N} \upsilon_{nk} \left(\tilde{\boldsymbol{R}}_{\boldsymbol{n}} \right)_{\mathbf{1:}\boldsymbol{k},\mathbf{1:}\boldsymbol{k}} \right]^{-1} \left(\tilde{\boldsymbol{G}}_{\mathbf{1:}\boldsymbol{k}} \right)^{T} \boldsymbol{W}_{\boldsymbol{k}} \left(\boldsymbol{w}_{\boldsymbol{k}} - \tilde{\boldsymbol{X}} \boldsymbol{\beta}_{\boldsymbol{k}} \right).
\end{aligned}
$$

Likewise, for $k = J+1, \ldots, K$, $\boldsymbol{b}_{\boldsymbol{k}}$ is given by

$$
\boldsymbol{b}_{\boldsymbol{k}} = \left[\sum_{n=1}^{N} \upsilon_{nk} \tilde{\boldsymbol{R}}_{\boldsymbol{n}} \right]^{-1} \tilde{\boldsymbol{G}}^{\boldsymbol{T}} \boldsymbol{W}_{\boldsymbol{k}} \left(\boldsymbol{w}_{\boldsymbol{k}} - \tilde{\boldsymbol{X}} \boldsymbol{\beta}_{\boldsymbol{k}} \right).
$$

4.1.4.3 The algorithm

As a result of the aforementioned developments, we shall use Algorithm 8 to estimate the parameters of the factor model.

4.1.5 The overall algorithm

Algorithm 9 consists in alternating the following steps: (i) Given the current set of parameters, calculate all the components of all the themes through the PING algorithm. (ii) Given the current components, find the working variables through the maximum likelihood of GLM. (iii) Given the working variables, estimate the factors model parameters through the EM algorithm. The method thus implemented is named F-SCGLR, for Factor SCGLR.

4.1.6 The clustering steps

The final aim of this work is to group the responses according to their mutual dependencies, conditional on the explanatory covariates. In other words, two responses having a high conditional correlation (positive or negative) should be cast to the same group. To achieve this, we propose the following strategy:

1. Calculate the conditional covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{B}^{\boldsymbol{T}} \boldsymbol{B}$.
2. Calculate the conditional correlation matrix $\boldsymbol{C}$ where $C_{ij} = \Sigma_{ij} / \sqrt{\Sigma_{ii} \Sigma_{jj}}$.
3. Calculate the dissimilarity matrix $\boldsymbol{D}$ where $D_{ij} = 2\left(1 - C_{ij}^2\right)$. The square conditional correlation is used in order to consider two responses highly positively or negatively correlated as close.

Algorithm 8: The EM algorithm applied to factor models with GLM

while *not convergence* **do**

Expectation step

for $n = 1, \ldots, N$ **do**

$$\boldsymbol{\alpha}_{\boldsymbol{n}}^{(\boldsymbol{t+1})} = \boldsymbol{B}^{(t)} \left(\boldsymbol{B}^{(t)\,T} \boldsymbol{B}^{(t)} + \boldsymbol{\Upsilon}_{\boldsymbol{n}}^{-\mathbf{1}} \right)^{-1}$$

$$\tilde{\boldsymbol{g}}_{\boldsymbol{n}}^{(\boldsymbol{t+1})} = \boldsymbol{\alpha}_{\boldsymbol{n}}^{(\boldsymbol{t+1})} \left(\boldsymbol{w}_{\boldsymbol{n}} - \boldsymbol{\Gamma}^{(t)\,T} \boldsymbol{f}_{\boldsymbol{n}} - \boldsymbol{\Delta}^{(t)\,T} \boldsymbol{a}_{\boldsymbol{n}} \right)$$

$$\tilde{\boldsymbol{R}}_{\boldsymbol{n}}^{(\boldsymbol{t+1})} = \boldsymbol{I}_{\boldsymbol{J}} - \boldsymbol{\alpha}_{\boldsymbol{n}}^{(\boldsymbol{t+1})} \boldsymbol{B}^{(t)\,T} + \tilde{\boldsymbol{g}}_{\boldsymbol{n}}^{(\boldsymbol{t+1})} \tilde{\boldsymbol{g}}_{\boldsymbol{n}}^{(\boldsymbol{t+1})\,T}$$

end

Maximization step

for $k = 1, \ldots, K$ **do**

$$\boldsymbol{\beta}_{\boldsymbol{k}}^{(\boldsymbol{t+1})} = \left(\tilde{\boldsymbol{X}}^{\boldsymbol{T}} \boldsymbol{W}_{\boldsymbol{k}} \tilde{\boldsymbol{X}} \right)^{-1} \tilde{\boldsymbol{X}}^{\boldsymbol{T}} \boldsymbol{W}_{\boldsymbol{k}} \left(\boldsymbol{w}_{\boldsymbol{k}} - \tilde{\boldsymbol{G}}^{(\boldsymbol{t+1})} \boldsymbol{b}_{\boldsymbol{k}}^{(t)} \right)$$

if *Gaussian* **then**

$$\sigma_k^{2\,(t+1)} = \frac{1}{N} \left\{ \left\| \boldsymbol{w}_{\boldsymbol{k}} - \tilde{\boldsymbol{X}} \boldsymbol{\beta}_{\boldsymbol{k}}^{(\boldsymbol{t+1})} \right\|^2 + \boldsymbol{b}_{\boldsymbol{k}}^{(t)\,T} \left(\sum_{n=1}^{N} \tilde{\boldsymbol{R}}_{\boldsymbol{n}}^{(\boldsymbol{t+1})} \right) \boldsymbol{b}_{\boldsymbol{k}}^{(t)} - 2 \left(\tilde{\boldsymbol{G}}^{(\boldsymbol{t+1})} \boldsymbol{b}_{\boldsymbol{k}}^{(t)} \right)^T \left(\boldsymbol{w}_{\boldsymbol{k}} - \tilde{\boldsymbol{X}} \boldsymbol{\beta}_{\boldsymbol{k}}^{(\boldsymbol{t+1})} \right) \right\}$$

end

if $k \leq J$ **then**

$$\boldsymbol{b}_{\boldsymbol{1:k,k}}^{(\boldsymbol{t+1})} = \left[\sum_{n=1}^{N} \upsilon_{nk} \left(\tilde{\boldsymbol{R}}_{\boldsymbol{n}}^{(\boldsymbol{t+1})} \right)_{\boldsymbol{1:k,1:k}} \right]^{-1} \left(\tilde{\boldsymbol{G}}_{\boldsymbol{1:k}}^{(\boldsymbol{t+1})} \right)^T \boldsymbol{W}_{\boldsymbol{k}} \left(\boldsymbol{w}_{\boldsymbol{k}} - \tilde{\boldsymbol{X}} \boldsymbol{\beta}_{\boldsymbol{k}}^{(\boldsymbol{t+1})} \right)$$

else

$$\boldsymbol{b}_{\boldsymbol{k}}^{(\boldsymbol{t+1})} = \left[\sum_{n=1}^{N} \upsilon_{nk} \tilde{\boldsymbol{R}}_{\boldsymbol{n}}^{(\boldsymbol{t+1})} \right]^{-1} \tilde{\boldsymbol{G}}^{(\boldsymbol{t+1})\,T} \boldsymbol{W}_{\boldsymbol{k}} \left(\boldsymbol{w}_{\boldsymbol{k}} - \tilde{\boldsymbol{X}} \boldsymbol{\beta}_{\boldsymbol{k}}^{(\boldsymbol{t+1})} \right)$$

end

end

$t \leftarrow t + 1$

end

Algorithm 9: The F-SCGLR algorithm

while *not convergence* **do**

Compute the components through the PING algorithm

$$\forall r = 1, \ldots, R, \forall h = 1, \ldots, H_r, \quad \boldsymbol{f}_r^{h\,(t+1)} = \boldsymbol{X}_r \boldsymbol{u}_r^{h\,(t+1)}$$

Compute the working variables through the IRLS algorithm

$$\boldsymbol{\eta}_k^{(t+1)} = \boldsymbol{F}^{(t+1)} \boldsymbol{\gamma}_k^{(t)} + \boldsymbol{A} \boldsymbol{\delta}_k^{(t)} + \boldsymbol{G} \boldsymbol{b}_k^{(t)}$$

$$\mu_{nk}^{(t+1)} = h_k^{-1}\left(\eta_{nk}^{(t+1)}\right), \forall n = 1, \ldots, N$$

$$w_{nk}^{(t+1)} = \eta_{nk}^{(t+1)} + h_k'\left(\mu_{nk}^{(t+1)}\right)\left(y_{nk} - \mu_{nk}^{(t+1)}\right), \forall n = 1, \ldots, N$$

$$\boldsymbol{W}_k^{(t+1)} = \text{diag}\left(\left[a_{nk}(\phi_k) v_k\left(\mu_{nk}^{(t+1)}\right) h_k'\left(\mu_{nk}^{(t+1)}\right)^2\right]^{-1}\right)_{n=1,\ldots,N}$$

Compute the model parameter through the EM algorithm

$$\boldsymbol{\Theta}^{(t+1)} = \underset{\boldsymbol{\Theta}}{\text{argmax}}\; l(\boldsymbol{\Theta}^{(t)}; \boldsymbol{\mathcal{W}})$$

Increment

$$t \leftarrow t + 1$$

end

4. Perform Multidimensional Scaling (MDS, Cox and Cox, 2008) on the matrix $\boldsymbol{D}$ to obtain a euclidean representation of the responses (i.e. a set of coordinates in a euclidean space) with respect to this distance structure. We use the function *cmdscale* of the **stats** package (R Core Team, 2021).

5. Perform a K-means algorithm (taking as a starting point the output of a hierarchical clustering procedure) on the coordinates obtained on the previous step. We use the **factoextra** package (Kassambara, 2017) where the function *hkmeans* runs the K-means and the function *fviz-nbclust* optimizes the number of clusters using the silhouette criterion.

4.2 Simulation study

Several simulation studies have been implemented to assess the performance of F-SCGLR. The first one focuses on the identification of the right combination of components and factors. The combination was calibrated across the cross-product grid $(H_1, \ldots, H_R, J) \in \{1, \ldots, 4\}^R \times \{0, \ldots, 5\}$ by minimizing the Bayesian Information Criterion (BIC, Schwarz, 1978). As shown by Chauvet et al. (2019), the hyper-parameters must be chosen to avoid the components to be too close to the principal components

($s > 0.5$) or to be drawn towards too local bundle ($l > 10$). Thus, the second simulation aims at studying the influence of the hyper-parameters $s \in \{0.1, 0.3, 0.5\}$ and $l \in \{1, 2, 3, 4, 7, 10\}$ in a situation of a more or less clearly separated cluster pattern. In this simulation, we use the Rand Index (RI, Rand, 1971) and the Adjusted Rand Index (ARI, Hubert and Arabie, 1985) to assess the correctness of the classification steps detailed in Section 4.1.6. In addition, to measure the quality of the latent variables recovery, we calculate the maximum square correlation between each latent variable $\boldsymbol{\xi}$ and the components:

$$\rho^2(\boldsymbol{\xi}, .) = \max_{r,h} \rho\left(\boldsymbol{\xi}, \boldsymbol{f}_r^h\right)^2,$$

where $\boldsymbol{f}_r^h$ denotes the hth component of theme $\boldsymbol{X}_r$. Finally, as reference values for comparison, we also calculated the RI and ARI of the partitions output by a competing R package in a context of binary data. For each simulation, one hundred samples have been generated. The R package **FactorSCGLR**, the simulation codes and the application to a real dataset are available at https://github.com/julien-gibaud/FactorSCGLR.

4.2.1 Simulation in a context of mixed distributions

4.2.1.1 Generation of the simulated data

The variables are simulated on $N = 100$ statistical units. Five latent variables $\boldsymbol{\xi_1}$, $\boldsymbol{\xi_2}$, $\boldsymbol{\xi_3}$, $\boldsymbol{\xi_4}$ and $\boldsymbol{\xi_5}$ are simulated independently. The $\boldsymbol{X}$ matrix consists in two themes: $\boldsymbol{X} = [\boldsymbol{X_1}, \boldsymbol{X_2}]$. The first theme $\boldsymbol{X_1} = [\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}, \boldsymbol{\mathcal{M}_1}]$ is made of three blocks: $\boldsymbol{\mathcal{X}_1} \in \mathbb{R}^{N\times 90}$ and $\boldsymbol{\mathcal{X}_2} \in \mathbb{R}^{N\times 60}$ are bundles of variables distributed about $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ respectively, and $\boldsymbol{\mathcal{M}_1}$ contains fifty unstructured noise variables drawn from a multivariate normal distribution. Likewise, the second theme $\boldsymbol{X_2} = [\boldsymbol{\mathcal{X}_3}, \boldsymbol{\mathcal{X}_4}, \boldsymbol{\mathcal{X}_5}, \boldsymbol{\mathcal{M}_2}]$ is made of four blocks: $\boldsymbol{\mathcal{X}_3} \in \mathbb{R}^{N\times 100}$, $\boldsymbol{\mathcal{X}_4} \in \mathbb{R}^{N\times 80}$ and $\boldsymbol{\mathcal{X}_5} \in \mathbb{R}^{N\times 60}$ are bundles of variables distributed about $\boldsymbol{\xi_3}$, $\boldsymbol{\xi_4}$ and $\boldsymbol{\xi_5}$ respectively, and $\boldsymbol{\mathcal{M}_2}$ contains sixty unstructured noise variables drawn from a multivariate normal distribution. More formally, for all $i = 1, \ldots, 5$, a variable $\boldsymbol{x_p}$ within a bundle is simulated as $\boldsymbol{x_p} = \boldsymbol{\xi_i} + \boldsymbol{\varepsilon_p}$, where $\boldsymbol{\varepsilon_p} \sim \mathcal{N}_N(0, 0.1\boldsymbol{I_N})$. This generation yields $P = 500$ explanatory variables. The N realizations of the $J = 3$ factors, simulated through $\boldsymbol{g_n} \sim \mathcal{N}_J(0, \boldsymbol{I_J})$, are stored in matrix $\boldsymbol{G} \in \mathbb{R}^{N\times J}$. The matrix $\boldsymbol{B} \in \mathbb{R}^{J\times K}$ of factor loadings is generated so as to exhibit a three-cluster pattern

$$\begin{aligned}
&\forall k = 1, \ldots, 5, \quad \boldsymbol{b_k} \sim \mathcal{N}_J\left(\boldsymbol{\mu_1}, \sigma_B^2 \boldsymbol{I_J}\right), \quad \forall k = 6, \ldots, 10, \quad \boldsymbol{b_k} \sim \mathcal{N}_J\left(-\boldsymbol{\mu_1}, \sigma_B^2 \boldsymbol{I_J}\right) \\
&\forall k = 11, \ldots, 20, \quad \boldsymbol{b_k} \sim \mathcal{N}_J\left(\boldsymbol{\mu_2}, \sigma_B^2 \boldsymbol{I_J}\right), \quad \forall k = 21, \ldots, 35, \quad \boldsymbol{b_k} \sim \mathcal{N}_J\left(-\boldsymbol{\mu_2}, \sigma_B^2 \boldsymbol{I_J}\right) \\
&\forall k = 36, \ldots, 50, \quad \boldsymbol{b_k} \sim \mathcal{N}_J\left(\boldsymbol{\mu_3}, \sigma_B^2 \boldsymbol{I_J}\right),
\end{aligned}$$

where $\sigma_B^2 = 0.1$, $\boldsymbol{\mu_1} = (2, 0, 0)^T$, $\boldsymbol{\mu_2} = (0, -1, 0)^T$ and $\boldsymbol{\mu_3} = (0, 0, 1.5)^T$. Finally, the response matrix $\boldsymbol{Y}$ is simulated as a mix of Gaussian, Poisson and Bernoulli distributions,

with

$$\begin{aligned}
\forall k = 1, \ldots, 20, \quad & \boldsymbol{y_k} \sim \mathcal{N}_N \left(\boldsymbol{\mu} = \gamma_{1k}\boldsymbol{\xi_1} + \gamma_{2k}\boldsymbol{\xi_2} + \boldsymbol{G b_k}, \boldsymbol{\Sigma} = \sigma_k^2 \boldsymbol{I_N}\right) \\
\forall k = 21, \ldots, 40, \quad & \boldsymbol{y_k} \sim \mathcal{P}\left(\boldsymbol{\lambda} = \exp\left[0.5\gamma_{1k}\boldsymbol{\xi_4} + 0.5\gamma_{2k}\boldsymbol{\xi_3} + \boldsymbol{G b_k}\right]\right) \\
\forall k = 41, \ldots, 50, \quad & \boldsymbol{y_k} \sim \mathcal{B}\left(\boldsymbol{p} = \text{logit}^{-1}\left[\gamma_{2k}\boldsymbol{\xi_3} + \gamma_{3k}\boldsymbol{\xi_2} + \boldsymbol{G b_k}\right]\right),
\end{aligned}$$

where for all k, σ_k^2, γ_{1k}, γ_{2k} and γ_{3k} are uniformly generated, with $\sigma_k^2 \in [0.1, 0.2]$, $\gamma_{1k} \in [-4, 4]$, $\gamma_{2k} \in [-2, 2]$ and $\gamma_{3k} \in [-0.5, 0.5]$. In the linear predictors, we rank the latent variables in the decreasing order of their regression parameter values.

4.2.1.2 Identification of the true model

In this simulation study, the hyper-parameters are first calibrated through the **SCGLR** package (e.g. without factors) and set to $s = 0.3$ and $l = 4$. Table 4.1 sums up the results on a cross-product grid. As expected, the combination which minimizes the BIC is given by the true combination $(H_1, H_2, J) = (2, 2, 3)$. However, several points deserve mentioning. We observe, for all component combinations, that the values of the BIC decrease dramatically when factors are involved in the model. This shows that, because mutual dependencies may generally exist, the conditional covariance should be modeled. When the model involves too many factors (when $J = 4$ and $J = 5$), the number of useful components is underestimated. Indeed, the variability of the model captured by the factors then contains a part of the variability otherwise captured by the components. In the opposite situation, when $J = 0$ or $J = 1$, the BIC leads to overestimate the number of components.

4.2.1.3 Varying the hyper-parameters and the variance within the clusters

Henceforth, keeping the true combination found by the BIC, we focus on the influence which the hyper-parameters s and l have on the clustering decision and latent variable recovery. In order to compare the results in a context of more or less distinct cluster pattern, we vary the variance within the cluster by taking $\sigma_B^2 \in \{0.1, 0.2, 0.3\}$. Figure 4.2 shows the conditional correlation matrices for the three values of σ_B^2.

Table 4.2 gives the results for $\sigma_B^2 = 0.1$. In the $s = 0.3$ and $s = 0.5$ cases, the values of RI and ARI are slightly better than in the $s = 0.1$ case. Moreover, for $s = 0.3$ and $s = 0.5$, the maximum value for RI and ARI is reached for $l = 4$. This is in accordance with the hyper-parameters calibrated through the **SCGLR** package. Table 4.3 and Table 4.4 sum up the results for respectively $\sigma_B^2 = 0.2$ and $\sigma_B^2 = 0.3$. As expected, the higher the variance within the cluster, the weaker the values of RI and ARI for all the combinations of s and l. We may also note that the difference between the values of RI and ARI across the hyper-parameters s and l tends to fade when σ_B^2 increases. The main result about the square correlations is that the variance within the cluster does not have a relevant influence on the quality of the latent variables recovery. Indeed, the search for components is related to the deterministic part of the model, while σ_B^2 is involved in the stochastic one. The

Table 4.1: Mean values of the BIC over a hundred samples for $(H_1, H_2) \in \{1, 2, 3, 4\}^2$ and J ranging from 0 to 5. The lowest values are in bold font.

$J = 0$				
H_2 \ H_1	1	2	3	4
1	79051	74909	63621	70120
2	57546	55794	46597	45896
3	54710	53330	43731	41406
4	44658	43938	42169	**39733**

$J = 1$				
H_2 \ H_1	1	2	3	4
1	34205	54295	28680	30613
2	29705	30369	26841	25463
3	31065	26012	24542	25592
4	34943	26930	**24520**	25369

$J = 2$				
H_2 \ H_1	1	2	3	4
1	21265	19833	19907	21087
2	20303	**18678**	21227	20157
3	20601	20361	22026	20565
4	20774	19022	19309	19496

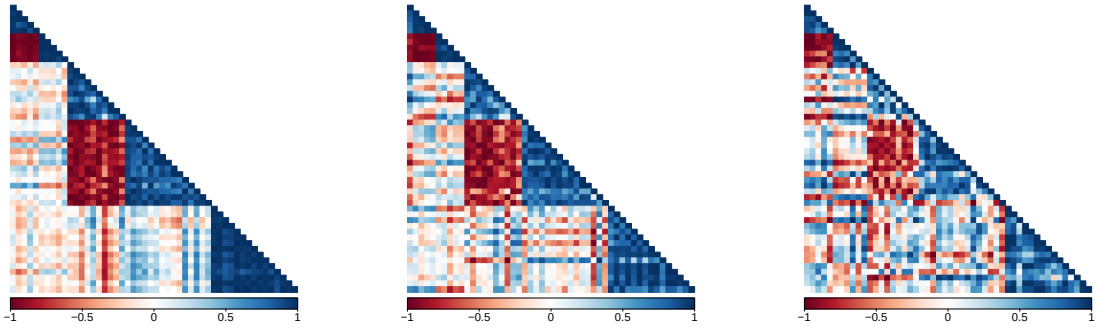
$J = 3$				
H_2 \ H_1	1	2	3	4
1	25655	17235	17474	17673
2	19150	**15915**	16197	16378
3	19050	16059	16341	16588
4	19329	16308	16640	16808

$J = 4$				
H_2 \ H_1	1	2	3	4
1	18356	**16025**	16237	16556
2	16364	16058	16287	16651
3	16534	16242	16484	16919
4	16835	16424	16761	17103

$J = 5$				
H_2 \ H_1	1	2	3	4
1	16584	**16212**	16462	16706
2	16647	16387	16628	16911
3	16852	16661	17020	17167
4	17211	16849	17279	17412

square correlations, for $s = 0.3$ and $s = 0.5$ with $l \geq 2$, are greater than for $s = 0.1$. This observation is consistent with Chauvet et al. (2019) who notice that the thinner the bundles, the greater the value of s has to be to recover the latent variables correctly. Here, indeed, the variance within the bundles is equal to 0.1 (thin bundles). However, the particular case of $l = 1$ deserves mentioning. The components calculated with $l = 1$ being close to the principal components, the two components of theme $\boldsymbol{X_2}$ settle between the three bundles and so, produce low square correlations with the latent variables. The interest of tuning the locality is shown by the gap between the results obtained for $l = 1$ and $l = 2$: in the latter case, the square correlations are dramatically improved. Furthermore, $\boldsymbol{\xi_3}$ being the less explanatory latent variable, $\rho^2(\boldsymbol{\xi_3}, .)$ is always lower than the others square correlations.

Figure 4.3 shows the correlation scatterplots in the component planes $(1, 2)$ for the first two themes. The components are almost perfectly aligned with the explanatory bundles. However, as observed in Table 4.2, Table 4.3 and Table 4.4, the bundle $\boldsymbol{\mathcal{X}_3}$ seems slightly less correlated with the component $\boldsymbol{f_2^2}$ than the other bundles with their corresponding components.



(a) Conditional correlation matrix for $\sigma_B^2 = 0.1$

(b) Conditional correlation matrix for $\sigma_B^2 = 0.2$

(c) Conditional correlation matrix for $\sigma_B^2 = 0.3$

Figure 4.2: Conditional correlation matrices for different values of σ_B^2

Table 4.2: Mean values of RI, ARI and square correlation over a hundred samples with $\sigma_B^2 = 0.1$, $s \in \{0.1, 0.3, 0.5\}$ and $l \in \{1, 2, 3, 4, 7, 10\}$.

s	l	RI	ARI	$\rho^2(\boldsymbol{\xi_1}, .)$	$\rho^2(\boldsymbol{\xi_2}, .)$	$\rho^2(\boldsymbol{\xi_3}, .)$	$\rho^2(\boldsymbol{\xi_4}, .)$
0.1	1	0.926	0.839	0.938	0.906	0.759	0.838
	2	0.927	0.839	0.980	0.959	0.810	0.935
	3	0.920	0.816	0.981	0.962	0.805	0.942
	4	0.928	0.837	0.979	0.966	0.816	0.945
	7	0.926	0.830	0.966	0.954	0.798	0.946
	10	0.927	0.835	0.967	0.955	0.792	0.945
0.3	1	0.944	0.876	0.973	0.936	0.735	0.753
	2	0.944	0.877	0.993	0.972	0.934	0.950
	3	0.946	0.881	0.987	0.974	0.938	0.965
	4	0.947	0.882	0.985	0.974	0.927	0.962
	7	0.943	0.875	0.984	0.974	0.911	0.964
	10	0.945	0.878	0.984	0.974	0.911	0.964
0.5	1	0.942	0.871	0.974	0.937	0.697	0.659
	2	0.944	0.875	0.994	0.972	0.943	0.946
	3	0.947	0.882	0.988	0.975	0.948	0.961
	4	0.948	0.884	0.986	0.975	0.946	0.967
	7	0.945	0.879	0.985	0.975	0.917	0.975
	10	0.944	0.877	0.985	0.975	0.911	0.969

Table 4.3: Mean values of RI, ARI and square correlation over a hundred samples with $\sigma_B^2 = 0.2$, $s \in \{0.1, 0.3, 0.5\}$ and $l \in \{1, 2, 3, 4, 7, 10\}$.

s	l	RI	ARI	$\rho^2(\boldsymbol{\xi_1}, .)$	$\rho^2(\boldsymbol{\xi_2}, .)$	$\rho^2(\boldsymbol{\xi_3}, .)$	$\rho^2(\boldsymbol{\xi_4}, .)$
0.1	1	0.797	0.515	0.939	0.909	0.770	0.830
	2	0.788	0.494	0.986	0.967	0.837	0.934
	3	0.789	0.505	0.985	0.969	0.848	0.948
	4	0.790	0.501	0.984	0.970	0.842	0.950
	7	0.785	0.493	0.977	0.969	0.839	0.942
	10	0.787	0.494	0.973	0.970	0.825	0.941
0.3	1	0.795	0.519	0.973	0.936	0.720	0.738
	2	0.801	0.527	0.992	0.972	0.937	0.944
	3	0.801	0.528	0.988	0.974	0.920	0.964
	4	0.800	0.532	0.985	0.975	0.908	0.966
	7	0.803	0.538	0.977	0.975	0.904	0.950
	10	0.806	0.538	0.977	0.975	0.878	0.956
0.5	1	0.796	0.524	0.974	0.937	0.703	0.657
	2	0.804	0.535	0.993	0.973	0.946	0.943
	3	0.798	0.522	0.985	0.975	0.938	0.955
	4	0.792	0.514	0.981	0.975	0.919	0.954
	7	0.799	0.527	0.978	0.976	0.919	0.957
	10	0.796	0.520	0.978	0.976	0.902	0.952

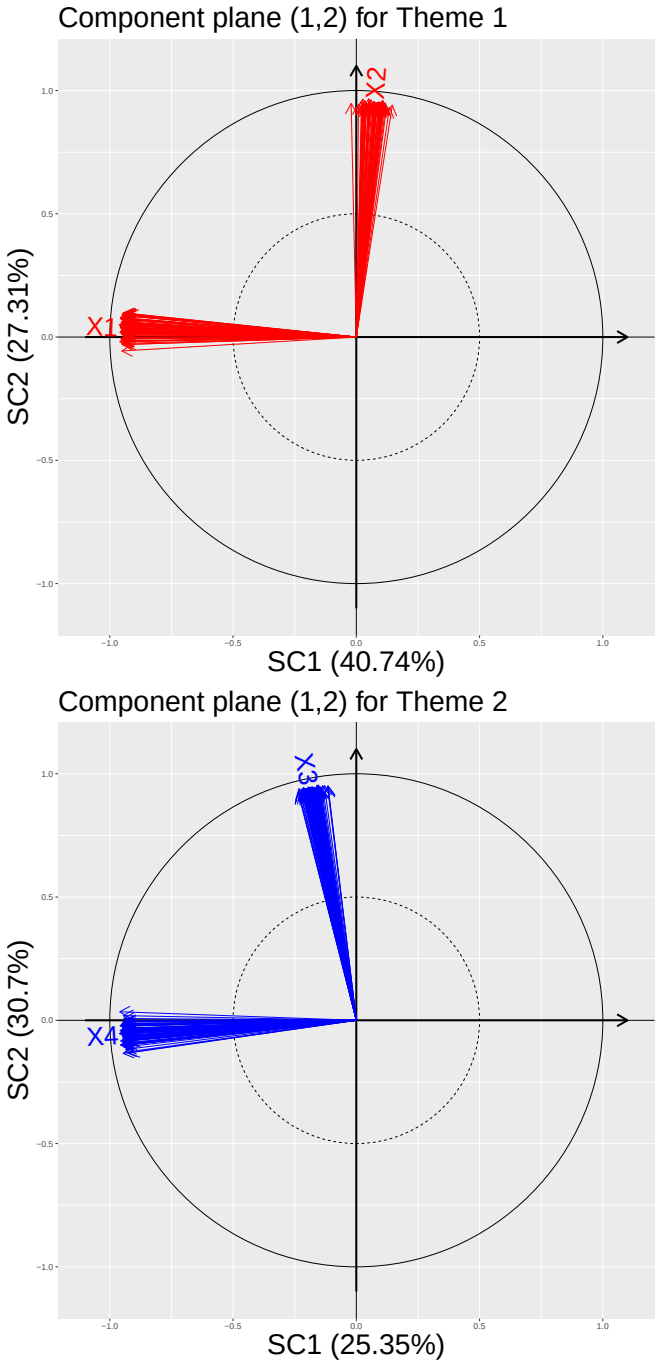


Figure 4.3: Correlation scatterplot of plane (1,2) for the two themes obtained by the F-SCGLR algorithm with $s = 0.3$ and $l = 4$. The red arrows represent the bundles $\boldsymbol{\mathcal{X}_1}$ and $\boldsymbol{\mathcal{X}_2}$ which explain the first theme. The blue ones represent the bundles $\boldsymbol{\mathcal{X}_3}$ and $\boldsymbol{\mathcal{X}_4}$ which explain the second theme. The percentage of inertia captured by each component is given in parentheses.

Table 4.4: Mean values of RI, ARI and square correlation over a hundred samples with $\sigma_B^2 = 0.3$, $s \in \{0.1, 0.3, 0.5\}$ and $l \in \{1, 2, 3, 4, 7, 10\}$.

s	l	RI	ARI	$\rho^2(\boldsymbol{\xi_1}, .)$	$\rho^2(\boldsymbol{\xi_2}, .)$	$\rho^2(\boldsymbol{\xi_3}, .)$	$\rho^2(\boldsymbol{\xi_4}, .)$
0.1	1	0.701	0.264	0.949	0.921	0.760	0.827
	2	0.706	0.288	0.987	0.970	0.838	0.921
	3	0.704	0.278	0.986	0.972	0.824	0.944
	4	0.704	0.274	0.982	0.971	0.829	0.946
	7	0.705	0.292	0.975	0.964	0.825	0.945
	10	0.709	0.302	0.967	0.966	0.817	0.944
0.3	1	0.699	0.280	0.974	0.938	0.715	0.726
	2	0.701	0.281	0.991	0.972	0.917	0.945
	3	0.698	0.289	0.988	0.975	0.921	0.965
	4	0.701	0.290	0.987	0.975	0.904	0.950
	7	0.700	0.288	0.974	0.977	0.886	0.944
	10	0.697	0.291	0.970	0.977	0.871	0.944
0.5	1	0.699	0.277	0.974	0.938	0.706	0.654
	2	0.706	0.287	0.992	0.973	0.927	0.941
	3	0.705	0.288	0.987	0.975	0.920	0.949
	4	0.704	0.287	0.983	0.975	0.913	0.938
	7	0.703	0.294	0.973	0.977	0.904	0.956
	10	0.700	0.296	0.971	0.977	0.899	0.950

4.2.2 Comparative study

To compare the different GLLVM implementations, we use the R package **gllvm** (Niku et al., 2019b). This package offers three ways to perform GLLVM estimation: using a variational approximation (VA, Hui et al., 2017), a Laplace approximation (LA, Niku et al., 2017, 2019a) or an extended variational approximation (EVA, Korhonen et al., 2023). Due to the excessive computation time of the Bayesian MCMC methods, the R packages **boral** (Hui, 2016) and **Hmsc** (Tikhonov et al., 2020) are not tested in this article. Their performances are respectively discussed by Niku et al. (2019b) and Pichler and Hartig (2021).

4.2.2.1 Generation of the simulated data

The variables are simulated on $N \in \{100, 200, 300\}$ statistical units. For the sake of simplicity, a bundle $\boldsymbol{X}$ of ten variables distributed about the latent variable $\boldsymbol{\xi}$ is generated. One categorical variable with three levels is taken as only additional covariate $\boldsymbol{A}$. In this simulation, $J = 2$ factors are simulated to model the conditional covariance of the $K \in \{10, 30, 50\}$ responses. The regression coefficients of the factors are generated in order to

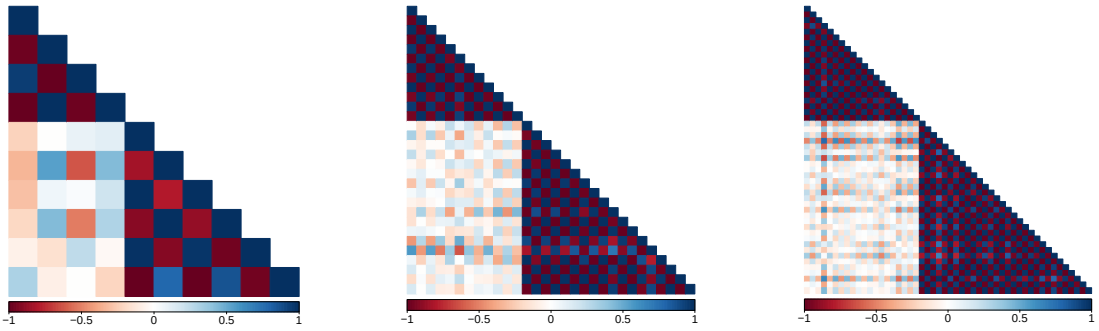
get a two-cluster design

$$\forall k = 1, \ldots, 0.4K, \quad \boldsymbol{b_k} \sim \mathcal{N}_J\left((-1)^k \boldsymbol{\mu_1}, 0.1\boldsymbol{I_J}\right)$$
$$\forall k = 0.4K + 1, \ldots, K, \quad \boldsymbol{b_k} \sim \mathcal{N}_J\left((-1)^k \boldsymbol{\mu_2}, 0.1\boldsymbol{I_J}\right),$$

where $\boldsymbol{\mu_1} = (0, 2)^T$ and $\boldsymbol{\mu_2} = (1.5, 0)^T$. The **gllvm** package not allowing to consider different distribution families for the responses, we restricted the comparison to binary outcomes

$$\forall k = 1, \ldots, K, \quad \boldsymbol{y_k} \sim \mathcal{B}\left(\boldsymbol{p} = \text{logit}^{-1}\left[\gamma_k \boldsymbol{\xi} + \boldsymbol{A\delta_k} + \boldsymbol{Gb_k}\right]\right),$$

where for all k, γ_k and δ_k are uniformly generated, with $\gamma_k \in [-4, 4]$ and $\delta_k \in [-1, 1]$. Figure 4.4 shows the conditional correlation matrices obtained for the three values of K.



(a) Conditional correlation matrix for $K = 10$

(b) Conditional correlation matrix for $K = 30$

(c) Conditional correlation matrix for $K = 50$

Figure 4.4: Conditional correlation matrices for different values of K

4.2.2.2 Compared results

In this simulation the package **SCGLR** calibrates the hyper-parameters to $s = 0.5$ and $l = 1$, while the BIC selects $H_1 = 1$ and $J = 2$. Table 4.5 sums up the RI, ARI and computation time output by, on the one hand, our package **FactorSCGLR** performing the F-SCGLR method, and, on the other hand, the package **gllvm** implementing the VA, LA and EVA approaches. We observe that, for all combinations of N and K, F-SCGLR gives the best values of RI and ARI, followed by EVA, VA and then LA. Indeed, the highest values obtained of the ARI are respectively: 0.991, 0.625, 0.556 and 0.517. Unlike our package and LA, which perform better when the number of either statistical units or responses increase, a higher number of responses may cause a deterioration of the classification correctness for EVA and VA. Across the simulations, F-SCGLR appears to be the fastest of the compared methods. The longest computation time (almost 17 seconds) occurred for $N = 300$ and $K = 50$, while EVA, VA and LA ran for 84, 74 and 373 seconds

respectively in that case. However, contrary to Korhonen et al. (2023), we did not observe that EVA ran faster than VA: in the $K = 10$ and $(N, K) = (300, 50)$ cases, the computation time of VA was lower. The conclusions of our respective works agree nevertheless that LA is relatively slow.

Table 4.5: Mean values of RI, ARI and computation time over a hundred samples with $N \in \{100, 200, 300\}$ and $K \in \{10, 30, 50\}$.

		F-SCGLR			gllvm-EVA		
N	K	RI	ARI	Time	RI	ARI	Time
100	10	0.847	0.683	1.200	0.687	0.386	1.569
	30	0.885	0.770	3.431	0.792	0.584	5.534
	50	0.933	0.867	8.148	0.806	0.611	11.54
200	10	0.922	0.839	2.307	0.721	0.456	4.963
	30	0.956	0.913	5.524	0.792	0.583	16.36
	50	0.980	0.961	11.43	0.762	0.518	39.96
300	10	0.950	0.898	2.823	0.726	0.466	8.507
	30	0.980	0.960	7.339	0.813	0.625	34.88
	50	0.996	0.991	16.99	0.767	0.529	84.52

		gllvm-VA			gllvm-LA		
N	K	RI	ARI	Time	RI	ARI	Time
100	10	0.657	0.337	1.508	0.598	0.231	32.35
	30	0.713	0.424	9.093	0.702	0.403	65.84
	50	0.716	0.426	20.74	0.709	0.414	117.3
200	10	0.727	0.468	4.355	0.652	0.323	43.84
	30	0.739	0.476	34.90	0.727	0.453	117.4
	50	0.719	0.432	41.76	0.733	0.460	256.3
300	10	0.725	0.455	8.288	0.710	0.433	50.91
	30	0.779	0.556	35.95	0.752	0.503	172.2
	50	0.770	0.534	74.22	0.761	0.517	373.8

4.3 Analysis of an agricultural ecology dataset

4.3.1 Data description

We apply F-SCGLR to the dataset available following the link https://doi.org/10.15454/AJZUQN. The sample we consider gives the observation of $K = 12$ agrobiodiversity variables over $N = 54$ winter cereal fields in the French Vallées et Coteaux de Gascogne. The agrobiodiversity is reported through three carabid beetle variables (two abundances and one Shannon index), three vascular plant variables

(one richness, one relative cover and one Shannon index) and six axes of correspondence analyses (CA) performed on presence-absence data of carabid species and plant species respectively. The three abundance and richness responses are assumed to be samples of Poisson random variables while the other responses are considered normally distributed. To model the agrobiodiversity, we have $P = 21$ variables partitioned into four themes and $Q = 1$ additional covariable. The first theme $\boldsymbol{X_1}$ characterizes the pest control through four variables. Six farming intensity variables make up the second theme $\boldsymbol{X_2}$. The third and fourth themes $\boldsymbol{X_3}$ and $\boldsymbol{X_4}$ gather six and five variables representing the landscape heterogeneity related to semi-natural covers and to the crop mosaic respectively. The binary categorical variable coding the observation year (2016 or 2017) is considered as the additional covariate put into matrix $\boldsymbol{A}$. For more information about this dataset, we refer the reader to Duflot et al. (2022).

4.3.2 Results and interpretation

As in Section 4.2, we need to calibrate the hyper-parameters. We first tune s and l through the **SCGLR** package, then we find the best combination of number of components and factors according to the BIC. However, due to the small number of explanatory variables in each theme, we only allow the number of components to reach $H_r = 3$. We thus minimize the BIC on the cross-product grid $(H_1, H_2, H_3, H_4, J) \in \{0, \ldots, 3\}^4 \times \{0, \ldots, 5\}$ with the s and l values previously found.

The **SCGLR** package recommends tuning hyper-parameters to $s = 0.5$ and $l = 1$ for this agricultural ecology dataset. Henceforth, the component combination minimizing the BIC is $(H_1, H_2, H_3, H_4) = (0, 3, 0, 0)$ meaning that only the farming intensity theme was found relevant for the prediction of the agrobiodiversity. Duflot et al. (2022) make the assumption that agrobiodiversity is predictable from the farming intensity (theme $\boldsymbol{X_2}$) and the landscape heterogeneity (themes $\boldsymbol{X_3}$ and $\boldsymbol{X_4}$). The combination found by the BIC validates this hypothesis as to the effect of the farming intensity and the non-effect of the pest control in the prediction of the agrobiodiversity. However the landscape heterogeneity themes proved here useless to this prediction.

We henceforth try to interpret the components of the second theme. The first component $\boldsymbol{f_2^1}$ is correlated ($\rho = -0.924$, $\rho = -0.794$ and $\rho = -0.738$) with a bundle of three variables, of which "TFI.total" and "TFI.h" represent a treatment frequency index of herbicides, and "nb.op" is the total number of operations conducted by the farmers. The second component $\boldsymbol{f_2^2}$ is correlated ($\rho = 0.779$) with the variable "cum.till.depth" measuring the cumulative tillage depth. The quantity of nitrogen denoted "qtyN.kg" is the most correlated explanatory variable ($\rho = -0.781$) with the last component $\boldsymbol{f_2^3}$. Figure 4.5 and Figure 4.6 represent the correlation plots of the second theme.

In this agricultural ecology dataset, three factors are recommended, according to the

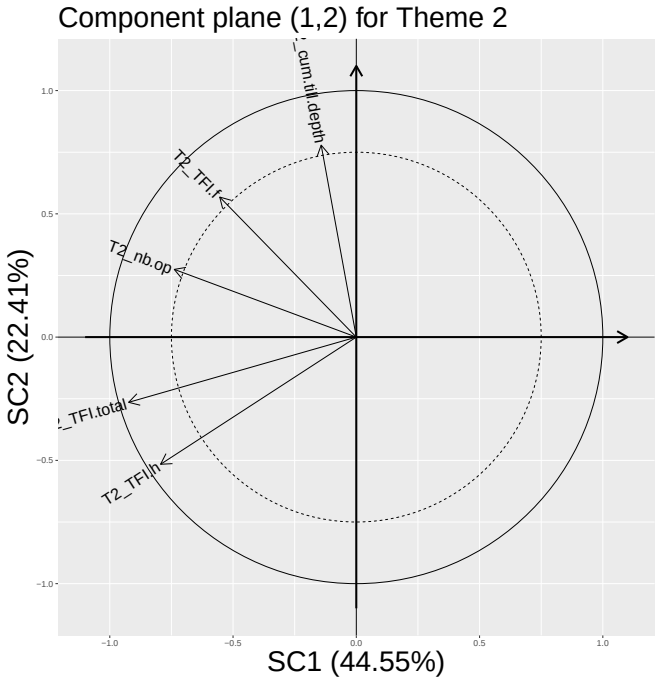


Figure 4.5: Correlation plot of F-SCGLR plane (1,2) of the second theme (farming intensity). The plot displays only variables having a cosine greater than 0.75 with the plane. The percentage of inertia captured by each component is given in parentheses.

BIC, to model the conditional variance-covariance matrix. By applying the clustering steps given in Section 4.1.6, four groups of responses are identified. The first group is composed by the three measures of the carabids. The second group gathers the first axis of the CA of the carabids, the plant richness and the plant Shannon diversity index. The carabids' second CA axis, the plant cover and the first and third plants' CA axes make up the third group. Finally, the fourth group contains the carabids' third CA axis and the plants' second CA axis. Figure 4.7 shows the conditional correlation values.

4.4 Conclusion and discussion

The original SCGLR was designed to regularize GLM estimation and reduce the explanatory dimension through components, so as to decompose the linear predictor in an interpretable way. It allowed to find strong and interpretable supervised components common to response variables, by achieving a trade-off between Goodness-of-Fit and a Structural Relevance measure. THEME-SCGLR extends SCGLR to a thematic partition of the explanatory variables, allowing to make better use of the complementary between the explanatory themes, both statistically when fitting the model, and conceptually when interpreting the components. F-SCGLR refines THEME-SCGLR in a major way: through a factor model, it models the conditional variance-covariance matrix of the responses using

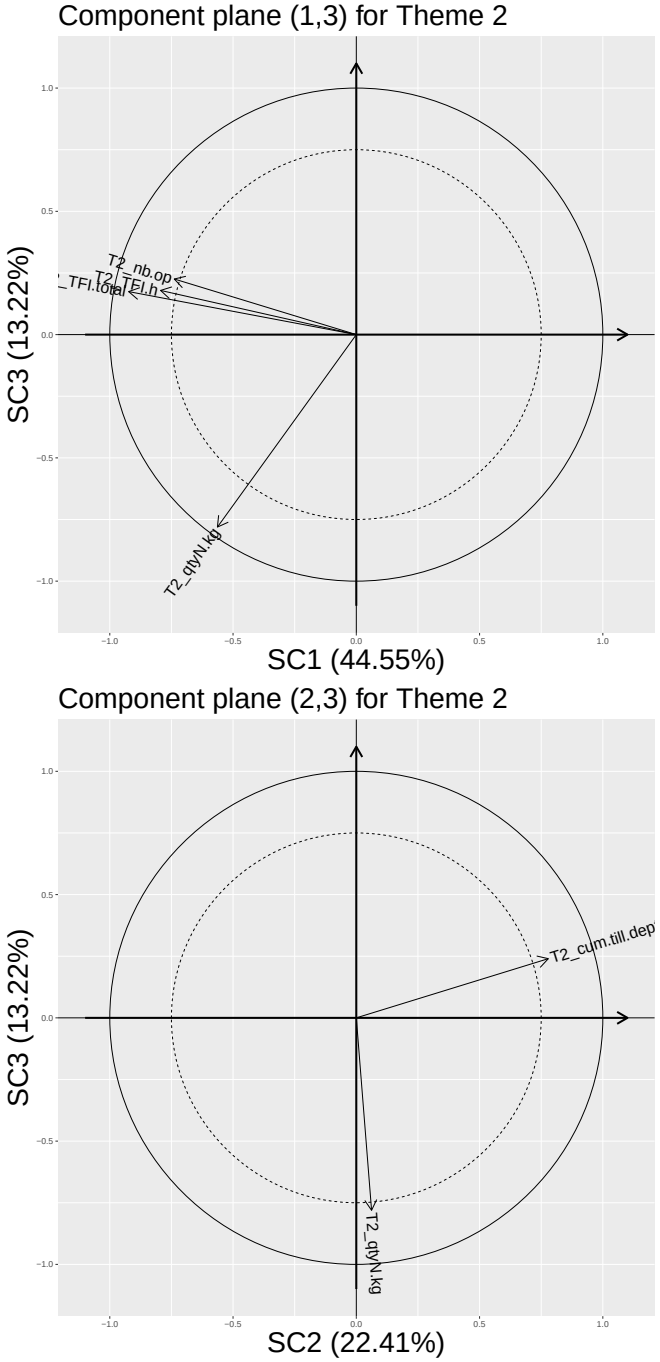


Figure 4.6: Correlation plots of F-SCGLR planes (1,3) and (2,3) of the second theme (farming intensity). The plot displays only variables having a cosine greater than 0.75 with the plane. The percentage of inertia captured by each component is given in parentheses.

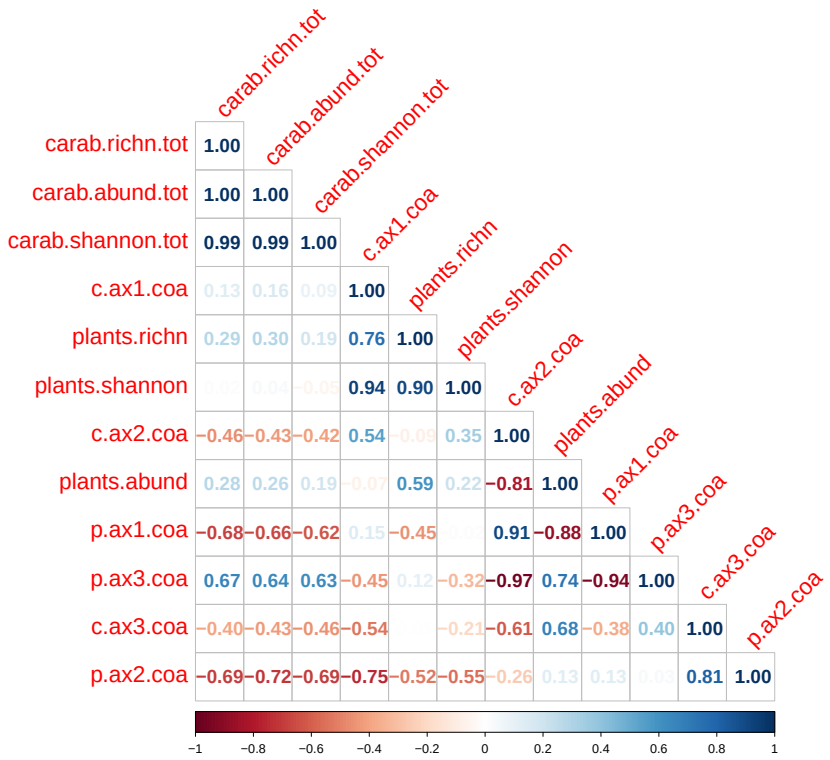


Figure 4.7: Conditional correlation matrix for the agricultural ecology dataset. The response variables are ordered by group.

a small number of latent factors. This matrix can be used as a basis for clustering, enabling to identify groups of linked responses.

In our simulation study, F-SCGLR proved to behave as expected regarding response clusters. Whenever the clusters were reasonably distinct, the original partitions were recovered. Whatever the dispersion of the regression coefficients within the clusters, it provided components aligned with the simulated latent variables. Our package outperforms the package **gllvm** in three ways: (i) The thematic model allows to find supervised components, thus reducing the dimension in a context of possibly numerous explanatory variables. (ii) Responses with different distribution families are allowed. (iii) The performances of our package are better in terms of computation time and of cluster detection. On the agricultural ecology dataset, we found four groups of responses. Due to very high conditional correlations, the first one gathers the measures of the carabids. The other groups are composed by a mix between the plant variables and the axes of the correspondence analyses. Moreover, performing F-SCGLR, we revealed that the treatment by herbicides,

the operations conducted by the farmers, the tillage depth and the quantity of nitrogen are the most involved variables in the prediction of the agrobidiversity.

In this research, some limitations have been reached. The use of EM algorithm on each step of the overall algorithm involves a high number of iterations. Due to the absence of consensus about the maximization of the log-likelihood, we think that more researches in this topics need to be effected. As mentioned in Section 3.4, SCGLR and its extensions suffer of a high number of hyper-parameters involving the use of heuristics to well calibrate the latter. Only Bernoulli, Binomial, Gaussian and Poisson distributions can currently be handled in the **FactorSCGLR** package. The package should be improved by adding different distributions as Negative Binomial, Zero Inflated Poisson, Tweedie, Gamma, Beta or Exponential, which are allowed in the **gllvm** package.

CHAPTER 5

IS THE WORK DONE? IT NEVER IS...

Contents

This chapter proposes to present a few works in progress and the perspectives for the future of supervised components.

5.1 SCGLR for statistical ecology

In statistical ecology literature, Carrascal et al. (2009) reveal that the use of supervised components built by PLS is already widespread. Their paper addresses, among other things, the analysis of the number of breeding landbird species among the Canary and Selvagem Islands using a bundle of six highly correlated explanatory variables. The authors highlight the need to perform supervised component-based model by interpreting the first PLS component as an "island syndrome" affecting bird species richness. However, as detailed by Chauvet (2019, pages 43–44), in a multiple predictive bundles framework, PLS fails to detect their presence, entailing a decrease of the interpretative power. We thus daresay that SCGLR and its extensions could be useful methodologies for modelers. In this vein, Mortier et al. (2022) propose a tutorial for the statistical ecologists. However, this tutorial only presenting the former versions of SCGLR, the approaches developed in this thesis complete these former versions and should prove useful in future applied work.

Some work remains to be done in the wake of ours. We hereafter suggest two possible directions, which are of course not restrictive.

5.2 Combining the extensions of SCGLR

First, we may integrate the formerly proposed extensions of SCGLR to our group-oriented version. For instance, Chauvet et al. (2019) propose to extend SCGLR to Generalized Linear Mixed Models (GLMMs), i.e. GLMs with random effects, in order to model responses with a repeated measure design. The independence assumption of statistical units is then no longer valid. The random effect being assumed different across the K responses, we have K vectors of random effects supposed independent and drawn from a multivariate normal distribution

$$\forall k = 1, \ldots, K, \quad \boldsymbol{\nu_k} \sim \mathcal{N}_L\left(0, \sigma^2_{\boldsymbol{\nu_k}} \boldsymbol{I_L}\right),$$

where $\sigma^2_{\boldsymbol{\nu_k}}$ is the variance component and L the number of realizations of the random effect observed in the data. Denoting $\boldsymbol{U} \in \mathbb{R}^{N \times L}$ the random effect design matrix, the linear predictor associated with response $\boldsymbol{y_k}$ becomes

$$\boldsymbol{\eta_k} = (\boldsymbol{Xu})\,\gamma_k + \boldsymbol{A\delta_k} + \boldsymbol{U\nu_k}.$$

In order to estimate the model parameters, Chauvet et al. (2019) adapt Schall (1991)'s algorithm to deal with a component-based model. As in Section 4.1.3, the model is first linearized, and then, the parameters of the linearized model are estimated under a Gaussian assumption. Let $\boldsymbol{w_k}$ be the working variable associated with response $\boldsymbol{y_k}$ and $\boldsymbol{W_k^{-1}}$ be its variance matrix, the linearized model is defined by

$$\boldsymbol{w_k} = (\boldsymbol{Xu})\,\gamma_k + \boldsymbol{A\delta_k} + \boldsymbol{U\nu_k} + \boldsymbol{\zeta_k},$$

where $\mathbb{E}[\boldsymbol{w_k}|\boldsymbol{\nu_k}] = (\boldsymbol{Xu})\,\gamma_k + \boldsymbol{A\delta_k} + \boldsymbol{U\nu_k}$ and $\mathbb{V}[\boldsymbol{w_k}|\boldsymbol{\nu_k}] = \mathbb{V}[\boldsymbol{\zeta_k}|\boldsymbol{\nu_k}] = \boldsymbol{W_k^{-1}}$. Given a component $\boldsymbol{f} = \boldsymbol{Xu}$, the regression parameters γ_k and $\boldsymbol{\delta_k}$ as well as vectors of random effect $\boldsymbol{\nu_k}$ are solution of the system proposed by Henderson (1975)

$$\begin{pmatrix} \boldsymbol{f^T W_k f} & \boldsymbol{f^T W_k A} & \boldsymbol{f^T W_k U} \\ \boldsymbol{A^T W_k f} & \boldsymbol{A^T W_k A} & \boldsymbol{A^T W_k U} \\ \boldsymbol{U^T W_k f} & \boldsymbol{U^T W_k A} & \boldsymbol{U^T W_k U} + \frac{1}{\sigma^2_{\nu_k}} \boldsymbol{I_L} \end{pmatrix} \begin{pmatrix} \gamma_k \\ \boldsymbol{\delta_k} \\ \boldsymbol{\nu_k} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f^T W_k w_k} \\ \boldsymbol{A^T W_k w_k} \\ \boldsymbol{U^T W_k w_k} \end{pmatrix}.$$

Finally, the maximum of pseudo-likelihood for variance components yields

$$\sigma^2_{\nu_k} = \frac{\boldsymbol{\nu_k^T \nu_k}}{L - \frac{1}{\sigma^2_{\nu_k}} \mathrm{Tr}\left[\left(\boldsymbol{U^T W_k U} + \frac{1}{\sigma^2_{\nu_k}} \boldsymbol{I_L}\right)^{-1}\right]}.$$

Using these developments, rmSCGLR could be refined to deal with a mixed model. In a finite response mixture framework, where the groups are defined by specific components, the linear predictor for response $\boldsymbol{y_k}$ writes

$$\boldsymbol{\eta_{kg}} = (\boldsymbol{Xu_g})\,\gamma_{kg} + \boldsymbol{A\delta_{kg}} + \boldsymbol{U\nu_{kg}}.$$

Likewise, in order to relax the independence assumption of both the responses and statistical units, F-SCGLR could be extended by integrating random effects in the linear predictor associated with response $\boldsymbol{y_k}$

$$\boldsymbol{\eta_k} = \boldsymbol{F\gamma_k} + \boldsymbol{A\delta_k} + \boldsymbol{Gb_k} + \boldsymbol{U\nu_k}.$$

We think that the development of these approaches would allow to make a better use of the information in the data whenever the independence hypothesis of the statistical units is unrealistic.

5.3 THEME sparse SCGLR

As pointed out in Section 3.4, SCGLR could be improved by constructing sparse loading vectors, whose coordinates are required to be null for covariates that are irrelevant to explain the response. In the PLS framework, following the Lasso principle (Tibshirani, 1996) where the shrinkage to zero is performed through a l_1 norm, Durif et al. (2018) propose to solve the following optimization program

$$\underset{\boldsymbol{u^T u}=1}{\mathrm{argmin}} \; -\mathrm{cov}\,(\boldsymbol{Xu}, \boldsymbol{w}) + \lambda \|\boldsymbol{u}\|_1,$$

where $\boldsymbol{w}$ is the working variable associated with univariate response $\boldsymbol{y}$. An equivalent program could improve the maximization of the specific criterion dedicated to SCGLR

$$\max_{\boldsymbol{u^T M^{-1} u}=1} s \ln(\phi(\boldsymbol{u})) + (1-s)\ln(\psi_{\boldsymbol{A}}(\boldsymbol{u})) - \lambda\|\boldsymbol{u}\|_1.$$

In the spirit of Yuan and Lin (2006) and Simon et al. (2013), another way of refining our method would be to select the themes in order to remove the non-explanatory ones. Besides, as in Liquet et al. (2016), the model could be directed towards solutions that are sparse at both theme level and component level, with intent to perform variable selection. In this case, the optimization program should be solved by iteratively maximizing in turn on every theme $\boldsymbol{X}_r$

$$\forall r, \quad \max_{\boldsymbol{u}_r^T \boldsymbol{M}^{-1} \boldsymbol{u}_r = 1} s \ln(\phi(\boldsymbol{u}_r)) + (1-s) \ln(\psi_{\boldsymbol{A}_r}(\boldsymbol{u}_r)) - \alpha\lambda \|\boldsymbol{u}_r\|_1 - (1-\alpha)\lambda \|\boldsymbol{u}_r\|_2,$$

where $\alpha \in [0, 1]$ is a hyper-parameter used to effect a convex combination of the lasso (l_1 norm) and group lasso (l_2 norm) penalties. The themes reported as relevant in the prediction of the response matrix, by this THEME-sparse-SCGLR should be compared with the original THEME-SCGLR where the themes kept for the prediction are those with a non-null number of components.

Note that, the criterion we aim to optimize not being differentiable, we should calculate the directional derivatives. A coordinate descent step therefore needs to be added in the PING algorithm.

5.4 The whole picture

In this manuscript we have extended SCGLR to address the possible existence of response groups in two ways. In a first proposal, the response mixture SCGLR was designed to identify groups of responses that can be predicted from group-specific supervised components. In order to specify sub-spaces which components need to keep away from, a separation sub-criterion was introduced. In a second work, we introduced random latent variables into the model to account for a conditional variance-covariance of the responses in which groups could be searched. Thus, across this manuscript, we have explored two aspects of the "response group" fuzzy concept. It is important to note that the groups we have been looking for are only defined through statistical links as common explanatory dimensions or conditional correlations. Many other accepting of the group concept could be investigated by integrating ecological or biological dynamics. For instance Favrichon (1998) clusters tropical forest species according to their growth behavior or Bellwood and Wainwright (2001) group fishes in the great barrier reef through their morphology.

CHAPTER 6

SUPPLEMENTARY MATERIAL

Contents

This chapter is dedicated to present the supplementary materials we need in this work.

6.1 The PING algorithm

The Projected Iterated Normed Gradient (PING) algorithm is an extension of the Power Iteration algorithm. To find the hth component, we use the PING algorithm which aims at solving any optimization program of the form

$$\begin{cases} \max\limits_{\boldsymbol{u}} \quad J_h(\boldsymbol{u}), \\ \text{s.t.} \quad \boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{M}^{-\mathbf{1}}\boldsymbol{u} = 1 \quad \text{and} \quad \boldsymbol{\Delta}_{\boldsymbol{h}}^{\boldsymbol{T}}\boldsymbol{u} = 0, \end{cases} \tag{6.1}$$

where J_h is a function of $\boldsymbol{u}$ to maximize and $\boldsymbol{\Delta_h}$ an additional constraint matrix. In the SCGLR context, $J_h(\boldsymbol{u})$ is the specific criterion and $\boldsymbol{\Delta_h}$ the orthogonal constraint matrix. We rewrite this optimization program by posing $\boldsymbol{v} = \boldsymbol{M}^{\mathbf{1/2}}\boldsymbol{u}$, $G_h(\boldsymbol{v}) = J_h(\boldsymbol{M}^{\mathbf{1/2}}\boldsymbol{v})$ and $\boldsymbol{E_h} = \boldsymbol{M}^{\mathbf{1/2}}\boldsymbol{\Delta_h}$.

$$\begin{cases} \max\limits_{\boldsymbol{v}} \quad G_h(\boldsymbol{v}), \\ \text{s.t.} \quad \boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{v} = 1 \quad \text{and} \quad \boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\boldsymbol{v} = 0. \end{cases} \tag{6.2}$$

6.1.1 The basic iteration

To solve (6.2), we must equate to zero the gradient of the following Lagrangian

$$\mathcal{L}(\boldsymbol{v}, \lambda, \boldsymbol{\eta}) = G_h(\boldsymbol{v}) - \lambda(\boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{v} - 1) - \boldsymbol{\eta}^{\boldsymbol{T}}\boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\boldsymbol{v}.$$

Setting $\Gamma_h(\boldsymbol{v}) = \nabla_{\boldsymbol{v}} G_h(\boldsymbol{v})$, we have

$$\nabla_{\boldsymbol{v}}\mathcal{L}(\boldsymbol{v}, \lambda, \boldsymbol{\eta}) = 0 \Leftrightarrow \Gamma_h(\boldsymbol{v}) - 2\lambda\boldsymbol{v} - \boldsymbol{E_h}\boldsymbol{\eta} = 0 \tag{6.3}$$

$$\Leftrightarrow \boldsymbol{v} = \frac{1}{2\lambda}\left(\Gamma_h(\boldsymbol{v}) - \boldsymbol{E_h}\boldsymbol{\eta}\right). \tag{6.4}$$

Multiplying (6.3) by $\boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}$

$$2\lambda \underbrace{\boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\boldsymbol{v}}_{=0} = \boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\Gamma_h(\boldsymbol{v}) - \boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\boldsymbol{E_h}\boldsymbol{\eta} \Leftrightarrow \boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\Gamma_h(\boldsymbol{v}) = \boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\boldsymbol{E_h}\boldsymbol{\eta}$$

$$\Leftrightarrow \boldsymbol{\eta} = \left(\boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\boldsymbol{E_h}\right)^{-1}\boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\Gamma_h(\boldsymbol{v}). \tag{6.5}$$

Substituting (6.5) in (6.4), we get

$$\begin{aligned} \boldsymbol{v} &= \frac{1}{2\lambda}\left(\Gamma_h(\boldsymbol{v}) - \boldsymbol{E_h}\left(\boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\boldsymbol{E_h}\right)^{-1}\boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\Gamma_h(\boldsymbol{v})\right) \\ &= \frac{1}{2\lambda}\left(\boldsymbol{I} - \boldsymbol{E_h}\left(\boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\boldsymbol{E_h}\right)^{-1}\boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\right)\Gamma_h(\boldsymbol{v}) \\ &= \frac{1}{2\lambda}\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^{\perp}}\Gamma_h(\boldsymbol{v}), \end{aligned}$$

where $\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^{\perp}} = \boldsymbol{I} - \boldsymbol{E_h}\left(\boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}\boldsymbol{E_h}\right)^{-1}\boldsymbol{E}_{\boldsymbol{h}}^{\boldsymbol{T}}$. Finally, the constraint $\|\boldsymbol{v}\|^2 = 1$ gives

$$\boldsymbol{v} = \frac{\frac{1}{2\lambda}\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^{\perp}}\Gamma_h(\boldsymbol{v})}{\left\|\frac{1}{2\lambda}\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^{\perp}}\Gamma_h(\boldsymbol{v})\right\|} = \frac{\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^{\perp}}\Gamma_h(\boldsymbol{v})}{\left\|\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^{\perp}}\Gamma_h(\boldsymbol{v})\right\|},$$

which suggests the basic iteration of the PING algorithm

$$\boldsymbol{v}^{(t+1)} = \frac{\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp}\Gamma_h(\boldsymbol{v}^{(t)})}{\left\|\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp}\Gamma_h(\boldsymbol{v}^{(t)})\right\|}. \tag{6.6}$$

6.1.2 Direction of ascent

Let us show that the basic iteration of the PING algorithm follows a direction of ascent. One way to do this is to show that the direction given by the arc $(\boldsymbol{v}^{(t)}, \boldsymbol{v}^{(t+1)})$ is a direction of ascent. In other words, show that

$$\left\langle \boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)}, \Gamma_h\left(\boldsymbol{v}^{(t)}\right)\right\rangle \geq 0.$$

By construction, we know that on every iteration t of the algorithm, $\boldsymbol{v}^{(t)}$ is orthogonal to span$[\boldsymbol{E_h}]$. Thus, since for all t, $\boldsymbol{v}^{(t)} = \boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp}\boldsymbol{v}^{(t)}$, we have

$$\begin{aligned}\left\langle \boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)}, \Gamma_h\left(\boldsymbol{v}^{(t)}\right)\right\rangle &= \left\langle \boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp}\left(\boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)}\right), \Gamma_h\left(\boldsymbol{v}^{(t)}\right)\right\rangle \\ &= \left\langle \boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)}, \boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp}\Gamma_h\left(\boldsymbol{v}^{(t)}\right)\right\rangle.\end{aligned}$$

Now, Equation (6.6) implies that

$$\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp}\Gamma_h(\boldsymbol{v}^{(t)}) = \boldsymbol{v}^{(t+1)}\left\|\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp}\Gamma_h(\boldsymbol{v}^{(t)})\right\|.$$

So,

$$\begin{aligned}\text{sgn}\left(\left\langle \boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)}, \Gamma_h\left(\boldsymbol{v}^{(t)}\right)\right\rangle\right) &= \text{sgn}\left(\left\langle \boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)}, \boldsymbol{v}^{(t+1)}\right\rangle\right) \\ &= \text{sgn}\left(\left\|\boldsymbol{v}^{(t+1)}\right\|^2 - \left\langle \boldsymbol{v}^{(t)}, \boldsymbol{v}^{(t+1)}\right\rangle\right) \\ &= \text{sgn}\left(1 - \cos\left(\boldsymbol{v}^{(t)}, \boldsymbol{v}^{(t+1)}\right)\right).\end{aligned}$$

Finally,

$$\left\langle \boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)}, \Gamma_h\left(\boldsymbol{v}^{(t)}\right)\right\rangle \geq 0.$$

6.1.3 Staying close enough to the current starting point

Although iteration (6.6) follows a direction of ascent, it does not guarantee that function G_h actually increases on every step. Indeed, we may go too far in such a direction, and overshoot the maximum. However, let us consider

$$\boldsymbol{\kappa}^{(t)} = \frac{\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp}\Gamma_h(\boldsymbol{v}^{(t)})}{\left\|\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp}\Gamma_h(\boldsymbol{v}^{(t)})\right\|}.$$

Staying close enough to the current starting point on the arc $(\boldsymbol{v}^{(t)}, \boldsymbol{\kappa}^{(t)})$ ensures that function G_h increases on every iteration. Indeed, let $\boldsymbol{\varpi}$ be the plane tangent to the unit sphere on $\boldsymbol{v}^{(t)}$ and let $\boldsymbol{w}$ denote the unit–vector tangent to arc $(\boldsymbol{v}^{(t)}, \boldsymbol{\kappa}^{(t)})$ on $\boldsymbol{v}^{(t)}$. Then, there exists $\tau > 0$ such that, $\boldsymbol{w} = \tau\boldsymbol{\Pi}_{\boldsymbol{\varpi}}\boldsymbol{\kappa}^{(t)}$, and

$$\left\langle \boldsymbol{w}, \boldsymbol{\kappa}^{(t)}\right\rangle = \tau\left\langle \boldsymbol{\Pi}_{\boldsymbol{\varpi}}\boldsymbol{\kappa}^{(t)}, \boldsymbol{\kappa}^{(t)}\right\rangle = \tau\cos^2(\boldsymbol{\kappa}^{(t)}, \boldsymbol{\varpi}) > 0.$$

6.1.4 The generic iteration

However, staying too close to the current starting point can impact the convergence speed of the algorithm to reach the maximum. We avoid that by using a one dimensional maximization function (e.g. Gauss-Newton type) to find the maximum of G_h on the arc $(\boldsymbol{v}^{(t)}, \boldsymbol{\kappa}^{(t)})$, and take it as $\boldsymbol{v}^{(t+1)}$. Therefore, we propose two possible generic iterations for the PING algorithm, which deal with this problem. Algorithm 10 and Algorithm 11 present these alternatives. The first one should be preferred, but is less easy to program.

Algorithm 10: The PING algorithm

while *not convergence* **do**

$$\boldsymbol{\kappa}^{(t)} \leftarrow \frac{\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp} \Gamma_h(\boldsymbol{v}^{(t)})}{\left\| \boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp} \Gamma_h(\boldsymbol{v}^{(t)}) \right\|}$$

Use a Gauss-Newton unidimensional maximization procedure to find the maximum of $G_h(\boldsymbol{v})$ on the arc $(\boldsymbol{v}^{(t)}, \boldsymbol{\kappa}^{(t)})$ and take it as $\boldsymbol{v}^{(t+1)}$

$t \leftarrow t + 1$

end

Algorithm 11: The alternative PING algorithm

while *not convergence* **do**

$$\boldsymbol{\kappa} \leftarrow \frac{\boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp} \Gamma_h(\boldsymbol{v}^{(t)})}{\left\| \boldsymbol{\Pi}_{\text{span}[\boldsymbol{E_h}]^\perp} \Gamma_h(\boldsymbol{v}^{(t)}) \right\|}$$

while $G_h(\boldsymbol{\kappa}) < G_h(\boldsymbol{v}^{(t)})$ **do**

$$\boldsymbol{\kappa} \leftarrow \frac{\boldsymbol{v}^{(t)} + \boldsymbol{\kappa}}{\left\| \boldsymbol{v}^{(t)} + \boldsymbol{\kappa} \right\|}$$

end

$\boldsymbol{v}^{(t+1)} \leftarrow \boldsymbol{\kappa}$

$t \leftarrow t + 1$

end

6.2 Analytical expression of the SCGLR specific criterion

The specific criteria which SCGLR needs to maximize for computing the $(h+1)$-th loading-vector write

$$\phi(\boldsymbol{u}) = \left(\sum_{j=1}^{p} \omega_j (\boldsymbol{u}^{\boldsymbol{T}} \boldsymbol{N_j} \boldsymbol{u})^l \right)^{1/l},$$

$$\psi_{\boldsymbol{A_h}}(\boldsymbol{u}) = \sum_{k=1}^{K} \|\boldsymbol{w_k}\|^2_{\boldsymbol{W_k}} \cos^2_{\boldsymbol{W_k}} \left(\boldsymbol{w_k},\, \text{span}[\boldsymbol{X}\boldsymbol{u}, \boldsymbol{A}^{\boldsymbol{h}}] \right)$$

and

$$\varphi_{\boldsymbol{F_{-g}}} \left(\boldsymbol{u_g^{h+1}} \right) = 1 - \frac{1}{G-1} \sum_{r \neq g} \left\langle \boldsymbol{\varpi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F_g^{h+1}}]}, \boldsymbol{\varpi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F_r^{Hr}}]} \right\rangle_{\text{Frob}}.$$

To facilitate the computation of the loading-vector, we give below an analytical expression of each sub-criterion and its gradient.

6.2.1 The structural relevance measure

In practice, we take either the variance component or the variable power inertia (VPI). In the first case, the SR and its gradient are easily given by

$$\phi(\boldsymbol{u}) = \|\boldsymbol{X}\boldsymbol{u}\|^2_{\boldsymbol{W}} \quad \text{and} \quad \nabla_{\boldsymbol{u}} \phi(\boldsymbol{u}) = 2\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{W} \boldsymbol{X} \boldsymbol{u}.$$

The explicit expression of VPI is

$$\phi(\boldsymbol{u}) = \left(\frac{1}{p} \sum_{j=1}^{p} \langle \boldsymbol{X}\boldsymbol{u}, \boldsymbol{x_j} \rangle^{2l}_{\boldsymbol{W}} \right)^{1/l}.$$

To calculate the gradient we use the classical rules of derivation

$$\begin{aligned}
\nabla_{\boldsymbol{u}} \phi(\boldsymbol{u}) &= \frac{1}{l} \left[\nabla_{\boldsymbol{u}} \left(\frac{1}{p} \sum_{j=1}^{p} \langle \boldsymbol{X}\boldsymbol{u}, \boldsymbol{x_j} \rangle^{2l}_{\boldsymbol{W}} \right) \right] \left[\frac{1}{p} \sum_{j=1}^{p} \langle \boldsymbol{X}\boldsymbol{u}, \boldsymbol{x_j} \rangle^{2l}_{\boldsymbol{W}} \right]^{1/l-1} \\
&= \frac{1}{l} \left[\frac{1}{p} \sum_{j=1}^{p} 2l \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{W} \boldsymbol{x_j} \langle \boldsymbol{X}\boldsymbol{u}, \boldsymbol{x_j} \rangle^{2l-1}_{\boldsymbol{W}} \right] \phi(\boldsymbol{u})^{1-l} \\
&= \frac{2}{p} \phi(\boldsymbol{u})^{1-l} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{W} \sum_{j=1}^{p} \langle \boldsymbol{X}\boldsymbol{u}, \boldsymbol{x_j} \rangle^{2l-1}_{\boldsymbol{W}} \boldsymbol{x_j}.
\end{aligned}$$

6.2.2 The goodness of fit measure

We aim at expressing $\psi_{\boldsymbol{A_h}}(\boldsymbol{u})$ as a function of quadratic forms. To achieve that, we decompose the projection on the regression space as follows

$$\text{span}[\boldsymbol{X}\boldsymbol{u}, \boldsymbol{A_h}] = \text{span}[\boldsymbol{\mathcal{X}_k^h}\boldsymbol{u}, \boldsymbol{A_h}] \quad \text{with} \quad \boldsymbol{\mathcal{X}_k^h} = \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{A_h}]^\perp} \boldsymbol{X}.$$

Since span$[\boldsymbol{\mathcal{X}}_k^h]$ is orthogonal to span$[\boldsymbol{A}_h]$,

$$\boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}_h]} = \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u},\boldsymbol{A}_h]} = \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}]} + \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{A}_h]}.$$

Consequently, by classical Euclidean statistical concepts, we have

$$\begin{aligned}
&\cos^2_{\boldsymbol{W_k}}\left(\boldsymbol{w_k},\ \text{span}[\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}_h]\right)\\
&= \cos_{\boldsymbol{W_k}}\left(\boldsymbol{w_k},\ \text{span}[\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}_h]\right)\cos_{\boldsymbol{W_k}}\left(\boldsymbol{w_k},\ \text{span}[\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}_h]\right)\\
&= \left[\frac{\left\|\boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}_h]}\boldsymbol{w_k}\right\|_{\boldsymbol{W_k}}}{\|\boldsymbol{w_k}\|_{\boldsymbol{W_k}}}\right]\left[\frac{\left\langle \boldsymbol{w_k}, \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}_h]}\boldsymbol{w_k}\right\rangle_{\boldsymbol{W_k}}}{\|\boldsymbol{w_k}\|_{\boldsymbol{W_k}}\left\|\boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}_h]}\boldsymbol{w_k}\right\|_{\boldsymbol{W_k}}}\right]\\
&= \frac{\left\langle \boldsymbol{w_k}, \left(\boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}]} + \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{A}_h]}\right)\boldsymbol{w_k}\right\rangle_{\boldsymbol{W_k}}}{\|\boldsymbol{w_k}\|^2_{\boldsymbol{W_k}}}\\
&= \frac{\left\langle \boldsymbol{w_k}, \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}]}\boldsymbol{w_k}\right\rangle_{\boldsymbol{W_k}}}{\|\boldsymbol{w_k}\|^2_{\boldsymbol{W_k}}} + \frac{\left\langle \boldsymbol{w_k}, \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{A}_h]}\boldsymbol{w_k}\right\rangle_{\boldsymbol{W_k}}}{\|\boldsymbol{w_k}\|^2_{\boldsymbol{W_k}}}.
\end{aligned}$$

The goodness of fit measure $\psi_{\boldsymbol{A}_h}(\boldsymbol{u})$ then writes more explicitly

$$\begin{aligned}
\psi_{\boldsymbol{A}_h}(\boldsymbol{u}) &= \sum_{k=1}^{K}\|\boldsymbol{w_k}\|^2_{\boldsymbol{W_k}}\cos^2_{\boldsymbol{W_k}}\left(\boldsymbol{w_k},\ \text{span}[\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}_h]\right)\\
&= \sum_{k=1}^{K}\left(\left\langle \boldsymbol{w_k}, \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}]}\boldsymbol{w_k}\right\rangle_{\boldsymbol{W_k}} + \left\langle \boldsymbol{w_k}, \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{A}_h]}\boldsymbol{w_k}\right\rangle_{\boldsymbol{W_k}}\right).
\end{aligned}$$

Now,

$$\begin{aligned}
\left\langle \boldsymbol{w_k}, \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}]}\boldsymbol{w_k}\right\rangle_{\boldsymbol{W_k}} &= \boldsymbol{w_k}^{\boldsymbol{T}}\boldsymbol{W_k}\boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}]}\boldsymbol{w_k}\\
&= \boldsymbol{w_k}^{\boldsymbol{T}}\boldsymbol{W_k}\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\left(\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{\mathcal{X}}_k^{h\,\boldsymbol{T}}\boldsymbol{W_k}\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right)^{-1}\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{\mathcal{X}}_k^{h\,\boldsymbol{T}}\boldsymbol{W_k}\boldsymbol{w_k}\\
&= \frac{\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{\mathcal{X}}_k^{h\,\boldsymbol{T}}\boldsymbol{W_k}\boldsymbol{w_k}\boldsymbol{w_k}^{\boldsymbol{T}}\boldsymbol{W_k}\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}}{\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{\mathcal{X}}_k^{h\,\boldsymbol{T}}\boldsymbol{W_k}\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}}.
\end{aligned}$$

Let,

$$\boldsymbol{a_k} := \boldsymbol{\mathcal{X}}_k^{h\,\boldsymbol{T}}\boldsymbol{W_k}\boldsymbol{w_k}\boldsymbol{w_k}^{\boldsymbol{T}}\boldsymbol{W_k}\boldsymbol{\mathcal{X}}_k^h, \quad \boldsymbol{b_k} := \boldsymbol{\mathcal{X}}_k^{h\,\boldsymbol{T}}\boldsymbol{W_k}\boldsymbol{\mathcal{X}}_k^h$$

and

$$c_k := \left\langle \boldsymbol{w_k}, \boldsymbol{\Pi}^{\boldsymbol{W_k}}_{\text{span}[\boldsymbol{A}_h]}\boldsymbol{w_k}\right\rangle_{\boldsymbol{W_k}}.$$

Finally, we have

$$\psi_{\boldsymbol{A}_h}(\boldsymbol{u}) = \sum_{k=1}^{K}\left(\frac{\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{a_k}\boldsymbol{u}}{\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{b_k}\boldsymbol{u}} + c_k\right)$$

and

$$\nabla_{\boldsymbol{u}}\psi_{\boldsymbol{A}_h}(\boldsymbol{u}) = 2\sum_{k=1}^{K}\frac{\left(\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{b_k}\boldsymbol{u}\right)\boldsymbol{a_k}\boldsymbol{u} - \left(\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{a_k}\boldsymbol{u}\right)\boldsymbol{b_k}\boldsymbol{u}}{\left(\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{b_k}\boldsymbol{u}\right)^2}.$$

6.2.3 The separation sub-criterion

We want to separate $\boldsymbol{F_g}$ of $\boldsymbol{F_r}$ for all $r \neq g$. The sub-criterion writes

$$\begin{aligned}
\varphi_{\boldsymbol{F}_{-g}}(\boldsymbol{u}_g^{h+1}) &= 1 - \frac{1}{G-1}\sum_{r\neq g}\left\langle \boldsymbol{\varpi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F}_g^h, \boldsymbol{X}\boldsymbol{u}_g^{h+1}]}, \boldsymbol{\varpi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F}_r^{H_r}]}\right\rangle_{\text{Frob}} \\
&= 1 - \frac{1}{G-1}\sum_{r\neq g}\left\langle \frac{\boldsymbol{\Pi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F}_g^h, \boldsymbol{X}\boldsymbol{u}_g^{h+1}]}}{\sqrt{h+1}}, \frac{\boldsymbol{\Pi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F}_r^{H_r}]}}{\sqrt{H_r}}\right\rangle_{\text{Frob}} \\
&= 1 - \frac{1}{G-1}\sum_{r\neq g}\frac{1}{\sqrt{H_r(h+1)}}\text{Tr}\left\{\boldsymbol{\Pi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F}_g^h, \boldsymbol{X}\boldsymbol{u}_g^{h+1}]}\boldsymbol{\Pi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F}_r^{H_r}]}\right\}.
\end{aligned}$$

Since $\text{span}[\boldsymbol{F}_g^h, \boldsymbol{X}\boldsymbol{u}_g^{h+1}] = \text{span}[\boldsymbol{f}_g^1, \ldots, \boldsymbol{f}_g^{h+1}]$ and $\text{span}[\boldsymbol{F}_r^{H_r}] = \text{span}[\boldsymbol{f}_r^1, \ldots, \boldsymbol{f}_r^{H_r}]$, we have

$$\begin{aligned}
&\text{Tr}\left\{\boldsymbol{\Pi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F}_g^h, \boldsymbol{X}\boldsymbol{u}_g^{h+1}]}\boldsymbol{\Pi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F}_r^{H_r}]}\right\} = \\
&\text{Tr}\left\{\left[\boldsymbol{f}_g^1, \ldots, \boldsymbol{f}_g^{h+1}\right]\left(\left[\boldsymbol{f}_g^1, \ldots, \boldsymbol{f}_g^{h+1}\right]^T\boldsymbol{W}\left[\boldsymbol{f}_g^1, \ldots, \boldsymbol{f}_g^{h+1}\right]\right)^{-1}\left[\boldsymbol{f}_g^1, \ldots, \boldsymbol{f}_g^{h+1}\right]^T\boldsymbol{W}\right. \\
&\quad\left.\left[\boldsymbol{f}_r^1, \ldots, \boldsymbol{f}_r^{H_r}\right]\left(\left[\boldsymbol{f}_r^1, \ldots, \boldsymbol{f}_r^{H_r}\right]^T\boldsymbol{W}\left[\boldsymbol{f}_r^1, \ldots, \boldsymbol{f}_r^{H_r}\right]\right)^{-1}\left[\boldsymbol{f}_r^1, \ldots, \boldsymbol{f}_r^{H_r}\right]^T\boldsymbol{W}\right\}.
\end{aligned}$$

Now, thanks to the orthogonality between the components, we obtain

$$\begin{aligned}
&\text{Tr}\left\{\boldsymbol{\Pi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F}_g^h, \boldsymbol{X}\boldsymbol{u}_g^{h+1}]}\boldsymbol{\Pi}^{\boldsymbol{W}}_{\text{span}[\boldsymbol{F}_r^{H_r}]}\right\} \\
&= \text{Tr}\left\{\left[\frac{\boldsymbol{f}_g^1}{\|\boldsymbol{f}_g^1\|_{\boldsymbol{W}}}, \ldots, \frac{\boldsymbol{f}_g^{h+1}}{\|\boldsymbol{f}_g^{h+1}\|_{\boldsymbol{W}}}\right]\left[\frac{\boldsymbol{f}_g^1}{\|\boldsymbol{f}_g^1\|_{\boldsymbol{W}}}, \ldots, \frac{\boldsymbol{f}_g^{h+1}}{\|\boldsymbol{f}_g^{h+1}\|_{\boldsymbol{W}}}\right]^T\boldsymbol{W}\right. \\
&\quad\left.\left[\frac{\boldsymbol{f}_r^1}{\|\boldsymbol{f}_r^1\|_{\boldsymbol{W}}}, \ldots, \frac{\boldsymbol{f}_r^{H_r}}{\|\boldsymbol{f}_r^{H_r}\|_{\boldsymbol{W}}}\right]\left[\frac{\boldsymbol{f}_r^1}{\|\boldsymbol{f}_r^1\|_{\boldsymbol{W}}}, \ldots, \frac{\boldsymbol{f}_r^{H_r}}{\|\boldsymbol{f}_r^{H_r}\|_{\boldsymbol{W}}}\right]^T\boldsymbol{W}\right\} \\
&= \text{Tr}\left\{\left[\frac{\boldsymbol{f}_g^1}{\|\boldsymbol{f}_g^1\|_{\boldsymbol{W}}}, \ldots, \frac{\boldsymbol{f}_g^{h+1}}{\|\boldsymbol{f}_g^{h+1}\|_{\boldsymbol{W}}}\right]^T\boldsymbol{W}\left[\frac{\boldsymbol{f}_r^1}{\|\boldsymbol{f}_r^1\|_{\boldsymbol{W}}}, \ldots, \frac{\boldsymbol{f}_r^{H_r}}{\|\boldsymbol{f}_r^{H_r}\|_{\boldsymbol{W}}}\right]\right. \\
&\quad\left.\left[\frac{\boldsymbol{f}_r^1}{\|\boldsymbol{f}_r^1\|_{\boldsymbol{W}}}, \ldots, \frac{\boldsymbol{f}_r^{H_r}}{\|\boldsymbol{f}_r^{H_r}\|_{\boldsymbol{W}}}\right]^T\boldsymbol{W}\left[\frac{\boldsymbol{f}_g^1}{\|\boldsymbol{f}_g^1\|_{\boldsymbol{W}}}, \ldots, \frac{\boldsymbol{f}_g^{h+1}}{\|\boldsymbol{f}_g^{h+1}\|_{\boldsymbol{W}}}\right]\right\} \\
&= \text{Tr}\{\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A}\},
\end{aligned}$$

where $A_{ij} = \dfrac{\left\langle \boldsymbol{f}_r^i, \boldsymbol{f}_g^j\right\rangle_{\boldsymbol{W}}}{\|\boldsymbol{f}_r^i\|_{\boldsymbol{W}}\left\|\boldsymbol{f}_g^j\right\|_{\boldsymbol{W}}}$, with $(i,j) \in \{1, \ldots, H_r\} \times \{1, \ldots, h+1\}$. This development leads to the explicit expression of $\varphi_{\boldsymbol{F}_{-g}}$

$$\varphi_{\boldsymbol{F}_{-g}}\left(\boldsymbol{u}_g^{h+1}\right) = 1 - \frac{1}{G-1}\sum_{r\neq g}\frac{1}{\sqrt{H_r(h+1)}}\sum_{i=1}^{H_r}\sum_{j=1}^{h+1}\frac{\left\langle \boldsymbol{X}\boldsymbol{u}_r^i, \boldsymbol{X}\boldsymbol{u}_g^j\right\rangle_{\boldsymbol{W}}^2}{\|\boldsymbol{X}\boldsymbol{u}_r^i\|_{\boldsymbol{W}}^2\left\|\boldsymbol{X}\boldsymbol{u}_g^j\right\|_{\boldsymbol{W}}^2}.$$

Let,

$$\begin{cases} \boldsymbol{d_{rgi}} := 2\left\langle \boldsymbol{X u_r^i}, \boldsymbol{X u_g^{h+1}} \right\rangle_{\boldsymbol{W}} \|\boldsymbol{X u_g^{h+1}}\|_{\boldsymbol{W}}^2 \boldsymbol{X^T W X u_r^i} \\ \boldsymbol{e_{rgi}} := 2\left\langle \boldsymbol{X u_r^i}, \boldsymbol{X u_g^{h+1}} \right\rangle_{\boldsymbol{W}}^2 \boldsymbol{X^T W X u_g^{h+1}} \\ f_{rgi} := \left(\left\|\boldsymbol{X u_g^{h+1}}\right\|_{\boldsymbol{W}}^2\right)^2 \left\|\boldsymbol{X u_r^i}\right\|_{\boldsymbol{W}}^2 \end{cases}$$

The gradient of the quotient becomes

$$\nabla_{\boldsymbol{u_g^{h+1}}} \left(\frac{\left\langle \boldsymbol{X u_r^i}, \boldsymbol{X u_g^{h+1}} \right\rangle_{\boldsymbol{W}}^2}{\|\boldsymbol{X u_r^i}\|_{\boldsymbol{W}}^2 \left\|\boldsymbol{X u_g^{h+1}}\right\|_{\boldsymbol{W}}^2} \right) = \frac{\boldsymbol{d_{rgi}} - \boldsymbol{e_{rgi}}}{f_{rgi}}$$

Then, we compute the gradient of $\varphi_{\boldsymbol{F_{-g}}}$

$$\nabla_{\boldsymbol{u_g^{h+1}}} \varphi_{\boldsymbol{F_{-g}}} \left(\boldsymbol{u_g^{h+1}}\right) = \frac{-1}{G-1} \sum_{r \neq g} \frac{1}{\sqrt{H_r(h+1)}} \sum_{i=1}^{H_r} \frac{\boldsymbol{d_{rgi}} - \boldsymbol{e_{rgi}}}{f_{rgi}}.$$

BIBLIOGRAPHY

Bartholomew, D. J. (1995). Spearman and the origin and development of factor analysis. *British Journal of Mathematical and Statistical Psychology*, 48(2):211–220.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach, Third Edition*. John Wiley & Sons.

Bastien, P., Vinzi, V. E., and Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics & data analysis*, 48(1):17–46.

Beale, C. M., Lennon, J. J., and Gimona, A. (2008). Opening the climate envelope reveals no macroscale associations with climate in European birds. *Proceedings of the National Academy of Sciences*, 105(39):14908–14912.

Bellwood, D. and Wainwright, P. (2001). Locomotion in labrid fishes: implications for habitat use and cross-shelf biogeography on the great barrier reef. *Coral reefs*, 20:139–150.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

Bougeard, S., Niang, N., Verron, T., and Bry, X. (2018). Current multiblock methods: competition or complementarity? A comparative study in a unified framework. *Chemometrics and Intelligent Laboratory Systems*, 182:131–148.

Bry, X., Redont, P., Verron, T., and Cazes, P. (2012). THEME-SEER: a multidimensional exploratory technique to analyze a structural model using an extended covariance criterion. *Journal of Chemometrics*, 26(5):158–169.

Bry, X., Simac, T., El Ghachi, S. E., and Antoine, P. (2020a). Bridging data exploration and modeling in event-history analysis: the supervised-component Cox regression. *Mathematical Population Studies*, 27(3):139–174.

Bry, X., Trottier, C., Mortier, F., and Cornu, G. (2020b). Component-based regularization of a multivariate GLM with a thematic partitioning of the explanatory variables. *Statistical Modelling*, 20(1):96–119.

Bry, X., Trottier, C., Verron, T., and Mortier, F. (2013). Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119:47–60.

Bry, X. and Verron, T. (2015). THEME: THEmatic Model Exploration through multiple co-structure maximization. *Journal of Chemometrics*, 29(12):637–647.

Bry, X., Verron, T., and Cazes, P. (2009). Exploring a physico-chemical multi-array explanatory model with a new multiple covariance-based technique: Structural equation exploratory regression. *Analytica chimica acta*, 642(1-2):45–58.

Carrascal, L. M., Galván, I., and Gordo, O. (2009). Partial least squares regression as an alternative to current regression methods used in ecology. *Oikos*, 118(5):681–690.

Chauvet, J. (2019). *Introducing complex dependency structures into supervised components-based models*. PhD thesis, Université Montpellier.

Chauvet, J., Trottier, C., and Bry, X. (2019). Component-Based Regularization of Multivariate Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, 28(4):909–920.

Chavent, M., Simonet, V. K., Liquet, B., and Saracco, J. (2012). ClustOfVar: An R Package for the Clustering of Variables. *Journal of Statistical Software*, 50(13):1–16.

Colder, C. R., Campbell, R. T., Ruel, E., Richardson, J. L., and Flay, B. R. (2002). A finite mixture model of growth trajectories of adolescent alcohol use: predictors and consequences. *Journal of consulting and clinical psychology*, 70(4):976.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory, Second Edition*. John Wiley & Sons.

Cox, M. A. A. and Cox, T. F. (2008). Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):463–474.

De Cáceres, M., Legendre, P., and Moretti, M. (2010). Improving indicator species analysis by combining groups of sites. *Oikos*, 119(10):1674–1684.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Duflot, R., San-Cristobal, M., Andrieu, E., Choisis, J.-P., Esquerré, D., Ladet, S., Ouin, A., Rivers-Moore, J., Sheeren, D., Sirami, C., Fauvel, M., and Vialatte, A. (2022). Farming intensity indirectly reduces crop yield through negative effects on agrobiodiversity and key ecological functions. *Agriculture, Ecosystems & Environment*, 326:107810.

Dufrêne, M. and Legendre, P. (1997). Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecological monographs*, 67(3):345–366.

Dunstan, P. K., Foster, S. D., and Darnell, R. (2011). Model based grouping of species across environmental gradients. *Ecological Modelling*, 222(4):955–963.

Dunstan, P. K., Foster, S. D., Hui, F. K., and Warton, D. I. (2013). Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *Journal of agricultural, biological, and environmental statistics*, 18(3):357–375.

Durif, G., Modolo, L., Michaelsson, J., Mold, J. E., Lambert-Lacroix, S., and Picard, F. (2018). High dimensional classification with combined adaptive sparse PLS and logistic regression. *Bioinformatics*, 34(3):485–493.

El Attar, A. (2012). *Estimation robuste des modeles de melange sur des donnees distribuees*. PhD thesis, Université de Nantes.

Escofier, B. and Pagès, J. (1984). L'analyse factorielle multiple. *Cahiers du Bureau universitaire de recherche opérationnelle Série Recherche*, 42:3–68.

Escofier, B. and Pagès, J. (1998). *Analyses factorielles simples et multiples*. Dunod, Paris.

Fahrmeir, L. (1994). Multivariate stastistical modelling based on generalized linear models. Technical report.

Fahrmeir, L. and Tutz, G. (2013). *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media.

Favrichon, V. (1998). Modeling the dynamics and species composition of a tropical mixed-species uneven-aged natural forest: effects of alternative cutting regimes. *Forest science*, 44(1):113–124.

Fonville, J. M., Richards, S. E., Barton, R. H., Boulange, C. L., Ebbels, T. M., Nicholson, J. K., Holmes, E., and Dumas, M.-E. (2010). The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *Journal of Chemometrics*, 24(11-12):636–649.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2):557–587.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2):149–170.

Greenacre, M. and Blasius, J. (2006). *Multiple correspondence analysis and related methods*. Chapman and Hall/CRC.

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology letters*, 8(9):993–1009.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2):423–447.

Hill, N., Woolley, S. N. C., Foster, S., Dunstan, P. K., McKinlay, J., Ovaskainen, O., and Johnson, C. (2020). Determining marine bioregions: A comparison of quantitative approaches. *Methods in Ecology and Evolution*, 11(10):1258–1272.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Hui, F. K. (2016). boral–Bayesian ordination and regression analysis of multivariate abundance data in R. *Methods in Ecology and Evolution*, 7(6):744–750.

Hui, F. K., Taskinen, S., Pledger, S., Foster, S. D., and Warton, D. I. (2015). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, 6(4):399–411.

Hui, F. K., 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(1):35–43.

Hunt, L. and Jorgensen, M. (2003). Mixture model clustering for mixed data with missing information. *Computational Statistics & Data Analysis*, 41(3-4):429–440.

Hutter, F., Lücke, J., and Schmidt-Thieme, L. (2015). Beyond manual tuning of hyperparameters. *KI-Künstliche Intelligenz*, 29(4):329–337.

Hwang, H. (2009). Regularized generalized structured component analysis. *Psychometrika*, 74(3):517–530.

Hwang, H. and Takane, Y. (2004). Generalized structured component analysis. *Psychometrika*, 69(1):81–99.

Jamshidian, M. (1997). An EM algorithm for ML Factor Analysis with Missing Data. In Berkane, M., editor, *Latent Variable Modeling and Applications to Causality*, volume 120, pages 247–258, New York, NY. Springer New York.

Jones, B. L., Nagin, D. S., and Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological methods & research*, 29(3):374–393.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32(4):443–482.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2):183–202.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Kassambara, A. (2017). Package 'factoextra'. *http://www.sthda.com/english/rpkgs/ factoextra*.

Kendall, M. G. and Stuart, A. (1961). *The advanced theory of statistics, Volume 2: Inference and Relationship*. Hafner.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 62(1):49–66.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451.

Korhonen, P., Hui, F. K., Niku, J., and Taskinen, S. (2023). Fast and universal estimation of latent variable models using extended variational approximations. *Statistics and Computing*, 33(26).

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19(2):191–201.

Lee, K., Guillemot, L., Yue, Y., Kramer, M., and Champion, D. (2012). Application of the Gaussian mixture model in pulsar astronomy-pulsar classification and candidates ranking for the Fermi 2FGL catalogue. *Monthly Notices of the Royal Astronomical Society*, 424(4):2832–2840.

Lee, S., Chugh, P. E., Shen, H., Eberle, R., and Dittmer, D. P. (2013). Poisson factor models with applications to non-normalized microRNA profiling. *Bioinformatics*, 29(9):1105–1111.

Lin, T. H. (2010). A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Quality & quantity*, 44(2):277–287.

Liquet, B., Lafaye De Micheaux, P., Hejblum, B. P., and Thiébaut, R. (2016). Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics*, 32(1):35–42.

Marx, B. D. (1996). Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, 38(4):374–381.

Marx, B. D. and Smith, E. P. (1990). Principal component estimation for generalized linear regression. *Biometrika*, 77(1):23–31.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Mevik, B.-H. and Wehrens, R. (2007). The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, 18(2):1–23.

Meyer, K. (2009). Factor-analytic models for genotype $\times$ environment type problems and structured covariance matrices. *Genetics Selection Evolution*, 41(21).

Monni, S. and Tadesse, M. G. (2009). A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Analysis*, 4(3):413–436.

Mortier, F., Chauvet, J., Trottier, C., Cornu, G., and Bry, X. (2022). Supervised component-based generalized linear regression: Method and extensions. In *Statistical Approaches for Hidden Variables in Ecology*, pages 181–202. John Wiley & Sons.

Mortier, F., Ouédraogo, D.-Y., Claeys, F., Tadesse, M. G., Cornu, G., Baya, F., Benedet, F., Freycon, V., Gourlet-Fleury, S., and Picard, N. (2015). Mixture of inhomogeneous matrix models for species-rich ecosystems. *Environmetrics*, 26(1):39–51.

Mortier, F., Trottier, C., Cornu, G., and Bry, X. (2016). SCGLR - An R Package for Supervised Component Generalized Linear Regression.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. John Wiley & Sons.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55(2):463–469.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K., Taskinen, S., and Warton, D. I. (2019a). Efficient estimation of generalized linear latent variable models. *PloS one*, 14(5):e0216129.

Niku, J., Hui, F. K., Taskinen, S., and Warton, D. I. (2019b). gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution*, 10(12):2173–2182.

Niku, J., Warton, D. I., Hui, F. K., 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(4):498–522.

Osborne, M. R. (1992). Fisher's method of scoring. *International Statistical Review / Revue Internationale de Statistique*, 60(1):99–117.

Ovaskainen, O. and Soininen, J. (2011). Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, 92(2):289–295.

Pagès, J. (2021). *Analyse factorielle multiple avec R*. EDP sciences.

Pichler, M. and Hartig, F. (2021). A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods in Ecology and Evolution*, 12(11):2159–2173.

Pledger, S. and Arnold, R. (2014). Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. *Computational Statistics & Data Analysis*, 71:241–261.

Pledger, S. and Phillpot, P. (2008). Using mixtures to model heterogeneity in ecological capture-recapture studies. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(6):1022–1034.

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J. S., and Thuiller, W. (2021). On the interpretations of joint modeling in community ecology. *Trends in Ecology & Evolution*, 36(5):391–401.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., and McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, 5(5):397–406.

Pörtner, H., Roberts, D., Adams, H., Adler, C., Aldunce, P., Ali, E., Ara Begum, R., Betts, R., Bezner Kerr, R., Biesbroek, R., Birkmann, J., Bowen, K., Castellanos, E., Cissé, G., Constable, A., Cramer, W., Dodman, D., Eriksen, S., Fischlin, A., Garschagen, M., Glavovic, B., Gilmore, E., Haasnoot, M., Harper, S., Hasegawa, T., Hayward, B., Hirabayashi, Y., Howden, M., Kalaba, K., Kiessling, W., Lasco, R., Lawrence, J., Lemos, M., Lempert, R., Ley, D., Lissner, T., Lluch-Cota, S., Loeschke, S., Lucatello, S., Luo, Y., Mackey, B., Maharaj, S., Mendez, C., Mintenbeck, K., Moncassim Vale, M., Morecroft, M., Mukherji, A., Mycoo, M., Mustonen, T., Nalau, J., Okem, A., Ometto, J., Parmesan, C., Pelling, M., Pinho, P., Poloczanska, E., Racault, M.-F., Reckien, D.,

Pereira, J., Revi, A., Rose, S., Sanchez-Rodriguez, R., Schipper, E., Schmidt, D., Schoeman, D., Shaw, R., Singh, C., Solecki, W., Stringer, L., Thomas, A., Totin, E., Trisos, C., Viner, D., van Aalst, M., Wairiu, M., Warren, R., Yanda, P., and Zaiton Ibrahim, Z. (2022). *Climate change 2022: Impacts, Adaptation and Vulnerability*. IPCC.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1):1–21.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239.

Réjou-Méchain, M., Mortier, F., Bastin, J.-F., Cornu, G., Barbier, N., Bayol, N., Bénédet, F., Bry, X., Dauby, G., Deblauwe, V., et al. (2021). Unveiling African rainforest composition and vulnerability to global change. *Nature*, 593:90–94.

Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76.

Rubin, D. B. and Thayer, D. T. (1983). More on EM for ML factor analysis. *Psychometrika*, 48(2):253–257.

Saidane, M. (2006). *Modèles à facteurs conditionnellement hétéroscédastiques et à structure markovienne cachée pour les séries financières*. PhD thesis, Université Montpellier II-Sciences et Techniques du Languedoc.

Saidane, M., Bry, X., and Lavergne, C. (2013). Generalized linear factor models: A new local EM estimation algorithm. *Communications in Statistics-Theory and Methods*, 42(16):2944–2958.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–727.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Sohn, M. B. and Li, H. (2018). A GLM-based latent variable ordination method for microbiome samples. *Biometrics*, 74(2):448–457.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–292.

Tami, M. (2016). *Approche EM pour modèles multi-blocs à facteurs à une équation structurelle*. PhD thesis, Université de Montpellier.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76(2):257–284.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M., and Lauro, C. (2005). PLS path modeling. *Computational statistics & data analysis*, 48(1):159–205.

Thomson, G. H. (1916). A hierarchy without a general factor. *British Journal of Psychology*, 8(3):271.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological review*, 38(5):406.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M., Oksanen, J., and Ovaskainen, O. (2020). Joint species distribution modelling with the R-package Hmsc. *Methods in ecology and evolution*, 11(3):442–447.

Trottier, C. (1998). *Estimation dans les modèles linéaires généralisés à effets aléatoires*. PhD thesis, Institut National Polytechnique de Grenoble-INPG.

Wangen, L. and Kowalski, B. (1989). A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of chemometrics*, 3(1):3–20.

Warton, D. I., Foster, S. D., De'ath, G., Stoklosa, J., and Dunstan, P. K. (2015). Model-based thinking for community ecology. *Plant Ecology*, 216(5):669–682.

Westerhuis, J. A., Kourti, T., and MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of chemometrics*, 12(5):301–321.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, pages 391–420.

Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate behavioral research*, 5(3):329–350.

Xu, T., Demmer, R. T., and Li, G. (2021). Zero-inflated Poisson factor model with application to microbiome read counts. *Biometrics*, 77(1):91–101.

Yee, T. W. and Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical modelling*, 3(1):15–41.

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(1):49–67.

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(2):301–320.

Abstract: In this thesis, a response matrix is assumed to depend on a set of explanatory variables, and a set of additional covariates. Explanatory variables are supposed many and redundant, thus demanding dimension reduction and regularization. By contrast, additional covariates contain few selected variables which are forced into the regression model, as they demand no regularization. Originally, the Supervised Component-based Generalized Linear Regression (SCGLR), a Partial Least Squares-type method, and its extension to multiple explanatory variable-blocks, THEME-SCGLR, are designed to extract from the explanatory variables several components jointly supervised by the set of responses. However, this methodology still has some limitations we aim to overcome in this thesis. The first limitation comes from the assumption that all the responses are predicted by the same explanatory space. However, in many practical situations, large sets of responses are not likely to depend exactly on the same explanatory dimensions. As a second limitation, the previous works involving SCGLR assume the responses independent conditional on the explanatory variables. Again, this is not very likely in practice, especially in situations like those in ecology, where a non-negligible part of the explanatory variables could not be measured. To overcome the first limitation, we assume that the responses are partitioned into several unknown groups. We suppose that the responses in each group are predictable from an appropriate number of specific orthogonal supervised components of the explanatory variables. We develop an extension of SCGLR based on a finite mixture model of the responses. The second work relaxes the conditional independence assumption. As in THEME-SCGLR, the response matrix is modeled by a thematic partitioning of the explanatory variables, named "themes". Thus, regularization is performed searching each theme for an appropriate number of components that both contribute to predict the response matrix and capture relevant structural information in themes. A set of few latent factors models the "residual" covariance matrix of the responses conditional on the components. The approaches presented in this work are tested on many simulation schemes, and then applied on ecology datasets.
Keywords: EM algorithm; factor model; latent variables; response mixture model; supervised components

Résumé : Dans cette thèse, une matrice réponse est supposée dépendre d'un ensemble de variables explicatives et d'un ensemble de covariables additionnelles. Les variables explicatives sont supposées nombreuses et redondantes, demandant ainsi réduction de dimension et régularisation. Au contraire, les covariables additionnelles contiennent quelques variables sélectionnées qui sont forcées dans le modèle de régression sans subir de régularisation. À l'origine, la Régression Linéaire Généralisée sur Composantes Supervisées (SCGLR) et son extension au multi-tableaux, THEME-SCGLR, sont créés pour extraire dans les variables explicatives plusieurs composantes conjointement supervisées par l'ensemble des réponses. Cependant, cette méthodologie a toujours des limitations que nous proposons de surpasser dans cette thèse. La première limitation vient de l'hypothèse que toutes les réponses sont prédites par le même espace explicatif. Cependant, dans de nombreuses situations pratiques, il est peu probable que de grands ensembles de réponses dépendent exactement des mêmes dimensions explicatives. Comme deuxième limitation, les précédents travaux impliquant SCGLR supposent que les réponses sont indépendantes conditionnellement aux variables explicatives. Encore une fois, cela est peu probable dans la pratique, spécialement dans des situations telles que l'écologie où une part non-négligeable des variables explicatives ne peuvent pas être mesurées. Pour surpasser la première limitation, nous supposons que les réponses sont partitionnées en plusieurs groupes inconnus. Nous supposons que les réponses dans chaque groupe sont prédites par un nombre approprié de composantes supervisées orthogonales spécifiques dans les variables explicatives. Nous développons une extension de SCGLR basée sur un modèle de mélange fini des réponses. Le deuxième travail relâche l'hypothèse d'indépendance conditionnelle. Comme pour THEME-SCGLR, la matrice réponse est modélisée par un partitionnement thématique des variables explicatives, nommés "thèmes". Ainsi, la régularisation est effectuée afin de chercher, dans chacun des thèmes, un nombre approprié de composantes qui contribuent à la fois à la prédiction de la matrice réponse et à la capture d'informations pertinentes des thèmes. Un ensemble de quelques facteurs latents modélise la covariance "résiduelle" des réponses conditionnellement aux composantes. Les approches présentées dans ce travail sont testées sur de nombreux schémas de simulation et ensuite appliquées à des jeux de données issus de l'écologie.
Mots clefs : algorithme EM ; composantes supervisées ; modèle à facteurs ; modèle de mélange sur les réponses ; variables latentes