On Optimization and Regularization for Grouped Dirichlet-multinomial Regression

by
Catherine Crea

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Doctor of Philosophy
in
Statistics

Guelph, Ontario, Canada

ABSTRACT

ON OPTIMIZATION AND REGULARIZATION FOR GROUPED DIRICHLET-MULTINOMIAL REGRESSION

Catherine Crea
University of Guelph, 2017

Advisor:
Dr. R. Ayesha Ali

This thesis focuses on developing the grouped Dirichlet-multinomial (DM) regression model for ecological applications with particular attention to optimization (for parameter estimation) and regularization (for variable selection). We adapt the grouped DM regression model for discrete choice behaviour to the analysis of mutualistic interactions between plant and pollinator species within a given ecosystem. The DM model provides a flexible approach to modelling over-dispersed grouped data and is fully parametric, but has not been well studied. The first part of this thesis focuses on establishing the DM model as a viable approach for analyzing pollination networks that can provide insights into the mechanisms driving ecological processes. Next, we study the behaviour of various parameterizations of the DM likelihood and identify non-convex regions that are either flat or non-smooth. Correspondingly, we evaluate the performance of three optimization methods (derivative and derivative-free) and assess their robustness to misspecification of dispersion structure. The last part of this thesis implements regularized regression for most parameterizations of the grouped Dirichlet-multinomial model using standard and adaptive lasso methods. Tuning

parameters are selected using an information criterion while optimization is achieved via the fast iterative shrinkage-thresholding algorithm. All the proposed methods are evaluated via simulated and empirical data sets and all implementations of the standard and regularized grouped DM regression model are publicly available as routines in `R`.

Dedication

This thesis is dedicated to my late brother Vince, who taught me that a strong mind combined with a strong will can overcome almost any challenge. Your strength, fearlessness and resiliency will forever be an inspiration.

Acknowledgements

I would like to thank my advisor, Dr. Ayesha Ali, for her expertise, support, and patience during my Ph.D. Her optimism and creativity are an inspiration and I am lucky to have been able to work with her for so many years. She always had a way of calming me down and putting me back on track during stressful times. She pushed me outside of my comfort zone and, as a result, I am a more confident scientist. I am forever grateful to have been able to have her by my side during this immensely challenging, yet rewarding, experience.

I would like to thank my advisory committee, Dr. Gary Umphrey, Dr. Tony Desmond and Dr. Peter Kevan. Our discussions/meetings were always insightful, encouraging, and amusing. I appreciate your time and contributions to the final preparation of my thesis.

Special thanks to Susan McCormick, Carrie Tanti and my student colleagues in the Department of Mathematics and Statistics for their support and laughter over the years.

It is difficult to overstate my gratitude to my employer, Geosyntec Consultants, for supporting me both financially and professionally during my Ph.D. A special thank you goes out to my greatest mentor and advocate, Julie Konzuk. Her hard work, dedication, and resiliency continually impresses me. I am also thankful for my close colleagues, Cathy Garvin, Noam Bar-Nahoum and James Rayner, for taking on extra responsibilities so that I could take time

off to focus on my research.

Lastly, this thesis would not have been possible without the love and support of my family and friends. My sister Mary, in particular, continually reminded me that I can achieve anything in spite of adversity and self-doubt. My friends kept me sane and laughing during all the stages of my Ph.D.

Contents

List of Tables

List of Figures

Chapter 1

General Introduction

This thesis contributes to the areas of multivariate statistics, unconstrained non-linear optimization, and regularized regression with applications to mutualistic ecological networks. In particular, grouped Dirichlet-multinomial (DM) regression is used to model the interaction probabilities of plant-pollinator mutualistic networks to gain insights into the mechanisms driving network structure. Further, a comprehensive study of the behaviour of the DM likelihood under various parameterizations, including an evaluation of the robustness to model misspecification, was conducted to facilitate the selection of the most appropriate optimization method for estimating the DM model parameters. Finally, regularized regression was developed for the grouped DM model using standard and adaptive lasso methods to facilitate simultaneous model selection and parameter estimation.

Mutualistic networks arise out of several ecological processes including pollination, seed dispersal, and host-parasite relationships, though the focus of this thesis is on pollination networks. Studies on these networks have revealed common structural patterns, such as the nested organization of pairwise interactions and the skewed distribution of links per species (Bascompte and Jordano 2007; Jordano et al. 2006; Vázquez et al. 2009a; Montoya et al. 2006). It is believed that these structural patterns are being driven by both evolutionary and ecological processes, which are summarized by the theories of neutrality (random interactions) and linkage rules (functional trait matching) (Blüthgen, 2006; Vázquez et al., 2009a; Santamaría and Rodríguez-Gironés, 2007; Allesina et al., 2008; Stang et al., 2009; Blüthgen, 2010; Olesen et al., 2011; Bartomeus, 2013; Junker et al., 2013; Eklöf et al., 2013; Rohr et al., 2010; Gravel et al., 2013). Few studies, however, have quantified the extent to which functional traits affect the probability of plants and pollinators interacting with each other.

Further, it is only within the past few years that field ecologists have collected detailed data to better understand the mechanisms that drive pollination, but there is a lack of statistical models to analyze these networks comprehensively. Therefore, DM regression was adapted from econometrics to quantify the contributions of species traits and/or linkage rules to pollination, thereby providing insights into the mechanisms that drive this ecological process. Additionally, the characteristics pollinators seek in plant species may be better anticipated if species interactions are modelled by the functional traits that drive them.

DM regression (Guimarães and Lindrooth, 2007) is a multinomial logistic regression model that accounts for over-dispersion, which occurs when the counts in the network are greater than what is predicted by the multinomial distribution. The counts in an interaction network can be thought of as (multinomial) responses such that for each pollinator species, a count represents the number of individual pollinators of a given species that select/visit one of the plant species. Using an econometric approach to modelling individual choice behaviour (Hausman et al., 1984; Shonkwiler and Hanley, 2003; Guimarães and Lindrooth, 2007), consider a pollinator being faced with a number of choices, where the number of choices is the number of plant species in the network. Within this framework, it is assumed that the pollinator assigns a level of utility to each plant species and then selects the one with the maximum utility (McFadden, 1974). These "utilities" can be modelled as a function of plant species attributes/traits, pollinator species characteristics/traits, or as interactions between the two (e.g. linkage rules). Therefore, the probability of a given pollinator species interacting with a given plant species can be modelled as a function of these traits. As such, the DM regression model also estimates a dispersion parameter which accounts for this extra-multinomial variation (Mosimann, 1962). This dispersion parameter can be a constant (i.e., the same value for all pollinator species) or it can also be modelled as a function of pollinator-specific covariates (i.e., the value of the dispersion parameter will vary between pollinators). In particular, we consider three parameterizations of the DM model: (i) no dispersion (standard multinomial model), (ii) δ (constant or a function of covariates), or (iii) ρ or intra-group correlation coefficient (constant or a function of covariates). The estimates of the regression coefficients quantify the relative contribution of each trait/linkage rule to the interaction probabilities, while the dispersion parameter accounts for heterogeneity in the data that is not explained by the covariates. Model parameters are estimated using standard maximum likelihood methods.

Unlike the multinomial likelihood, the DM likelihood can be complex and is sensitive to the choice of optimization/maximization methods. In particular, the DM likelihood is not part of the exponential family of distributions and, in general, is non-convex with non-smooth regions (Zhang et al., 2016). These features can result in convergence issues (both local and global) and Hessian instability. Guimarães and Lindrooth (2007) introduced this (econometric) model and implemented a modified Newton-Raphson for the maximization of the DM regression model, but they did not evaluate the behaviour of the DM likelihood under the various parameterizations or the robustness to misspecification of dispersion structure. In fact, the Newton-Raphson can be limited in practice because of its poor global convergence properties and the need for second derivative information; hence, lower order optimization methods (i.e., quasi-Newton or direct search) may be better conditioned to handle the complexities of the DM likelihood (Ypma, 1995). Likewise, for empirical data sets, the true underlying dispersion structure is unknown; thus, robustness to model misspecification can provide valuable insights when fitting these models to real-world data sets.

Another practical problem when analyzing empirical data sets is variable selection. When there are several available covariates, it is important to be able to discern which, if any, of the covariates are important predictors of interaction. Standard methods for variable selection use forward/backward strategies; however, these methods are computationally expensive and can yield models that are too small with biased estimates and can become unstable in high dimensional settings (Hastie et al., 2009). Regularized regression is a popular alternative that can conduct variable selection and parameter estimation simultaneously. These methods typically apply a penalty term, based on a function of the model parameters, to the log-likelihood and then maximization is done on the penalized log-likelihood. One of the most prominent examples is the least absolute shrinkage and soft-thresholding operator (lasso) of Tibshirani (1996), which induces sparsity by using an $l1$ norm penalty term. Other sparsity-inducing penalties include the adaptive lasso (Zou, 2006), the fused lasso (Tibshirani et al., 2005), the group lasso (Yuan and Lin, 2006; Meier et al., 2008), and the sparse group lasso (Yuan and Lin, 2007). Alternative regularized estimators include ridge regression (Tikhonov, 1977), the elastic net (Zou and Hastie, 2005), and the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001). Although these methods have been well developed for many linear and nonlinear regression models, particularly those with a univariate response, regularized regression has not been developed for this implementation of the grouped DM model. Also, most algorithms for penalized likelihoods assume that the

log-likelihood is convex and use coordinate gradient descent methods for optimization. In this thesis, we focus on implementing lasso methods to the grouped DM regression model and we implement accelerated proximal gradient methods (Nesterov, 2007; Beck and Teboulle, 2009) for optimizing the penalized likelihood.

The objectives of this thesis are three-fold. First and foremost, this thesis aims to provide ecologists with robust statistical methods to conduct community-level analyses to better understand the mechanisms that drive mutualistic networks. This is accomplished through an evaluation of the performance of DM regression for ecological networks and an exploration of the model parameter space. The second objective is to study the behaviour of various parameterizations of the DM likelihood and its robustness to model misspecification to facilitate the selection of the appropriate combination of parameterization and optimization method. This is achieved through an evaluation of the robustness of derivative and derivative-free optimization methods for model parameter estimation. The final objective is to extend the DM model to perform model selection via regularized regression. This extension includes an evaluation of both standard and adaptive lasso methods, tuning parameter selection criteria, and the implementation of the fast iterative shrinkage thresholding algorithm (FISTA), which is an accelerated version of the proximal gradient method (Beck and Teboulle, 2009).

The remainder of this thesis is organized as follows. Chapter 2 provides general background on the various implementations of DM regression, pollination networks and network metrics, and details on derivative and derivative-free optimization methods. Chapter 3 presents an evaluation of the performance of DM regression and an exploration of the model parameter space (Crea et al., 2016). Chapter 4 presents a comprehensive evaluation of the behaviour of the DM likelihood under various parameterizations and the robustness to misspecification of dispersion structure. Chapter 5 contains the implementation of regularized DM regression, using lasso-type penalties and a proximal gradient algorithm for optimization. Conclusions and future works are discussed in Chapter 6.

Chapter 2

Background

2.1 Relation to Other DM Models and Thesis Notation

The most common implementation of the DM regression model focuses on applications in which counts are at the individual level and the log odds of selecting categories are modelled as a function of the individuals' characteristics. This parallels the structure of the standard multinomial logit model, where one of the J response categories is set as the baseline and the log odds for all other categories are calculated relative to the baseline. Effectively the model estimates $J-1$ regression coefficients for each covariate, and each β_j can be interpreted as the effect of the covariates on the odds of making a given choice over another. In essence, the MNL aims to explain how an individual's characteristics affect the likelihood of falling in a particular response category. This type of model frequently arises in bioinformatics, where interest lies in the association between microbiome composition and environmental factors. Chen and Li (2013) use the DM regression model to link nutrient intake to the human gut microbiome. In topic modelling, DM regression is used within a mixture model, in which a Dirichlet distribution over topics is modelled as a function of document features (Mimno and McCallum, 2008). In these applications, the DM model is derived from a probability framework, where the multivariate counts are assumed to follow a multinomial distribution and the multinomial parameters follow a Dirichlet distribution, which results in a DM distribution (Mosimann, 1962). Zhang et al. (2016) give a comprehensive evaluation of these types of multinomial regression models, motivated from high-throughput data in genomics, where they compare models that accommodate varying correlation structures, e.g., a generalized Dirichlet distribution is used for the distribution of the multinomial parameters.

Alternatively, a multinomial logit which incorporates category-specific variables is commonly referred to as a conditional logit (CL) model. CLs are often used in econometrics to study discrete choice behaviour. In a pure CL, there is only one regression coefficient for each covariate, and the log odds can be interpreted as the weighted difference between the choice characteristics, where the weight is the regression coefficient. Inherently, the CL aims to explain how characteristics of the choices affect the likelihood of an individual selecting a choice. Note that conditional logits can also incorporate individual level covariates and interactions between individual and choice covariates. Additionally, CLs can be applied to data sets where vectors of counts across the choices are aggregated into distinct groups, hence the data are at the group level and not the individual level, and will be referred to as the grouped CL (GCL). Within this context, the DM regression model parallels the structure of the GCL, i.e., log odds are modelled as a function of choice and/or group-specific characteristics, while accommodating for overdispersed grouped count data. In these applications, the DM model is derived from a Random Utility Model (RUM) framework, where groups of individuals ascribe a level of utility, which is a linear combination of group/choice characteristics, to each choice and select the one with maximum utility (McFadden, 1974). The logit formulation for the multinomial probabilities is the result of the Type I Extreme Value (or standard Gumbel) distribution placed on the utility function. When additional random effect terms are included in the utility function, heterogeneity in the counts (e.g., overdispersion) and/or varying correlation structures among choices provide a more flexible framework for discrete choice models. In the econometrics literature, the mixed logit model is the most popular choice (McFadden and Train, 2000); however, the grouped DM model is a fully parametric alternative (Guimarães and Lindrooth, 2007) and it is this implementation of DM regression that is the focus of this thesis.

Given the number of parameters involved in the grouped DM regression setup and for ease of reference, the notation used throughout this thesis is summarized in Table 2.1.

2.2 Pollination Networks and Network Metrics

A plant-pollinator network can be represented by a matrix of counts, in which each cell corresponds to the number of interactions between a particular plant-pollinator species pair. Every species represented by the matrix corresponds to a node in the network, and every non-zero entry corresponds to a link between a plant species and pollinator species. These

Table 2.1: DM Regression Notation

Parameter	Meaning
$i = 1, \ldots, I$	index for an individual pollinator
$g = 1, \ldots, G$	index for pollinator species
$j = 1, \ldots, J$	index for plant species
$k = 1, \ldots, K$	index for plant covariates and/or linkage rules
$l = 1, \ldots, L$	index for pollinator covariates
$N = G \times J$	size of the network
$M = K + L$	total number of model parameters
U_{igj}	latent utility assigned by an individual pollinator
$\mathbf{Y}$	$\mathbf{G} \times \mathbf{J}$ matrix of interactions or counts
y_{gj}	gj^{th} count in $\mathbf{Y}$
n_g	row sums of $\mathbf{Y}$
$\mathbf{P}$	$\mathbf{G} \times \mathbf{J}$ matrix of interaction probabilities
p_{gj}	gj^{th} probability in $\mathbf{P}$
p_g	the row probability vectors of $\mathbf{P}$
α_g	Dirichlet parameter vector associated with p_g
η_{gj}	Gamma distributed random group (i.e., pollinator species) effect
$\mathbf{X}$	$\mathbf{G} \times \mathbf{J} \times \mathbf{K}$ array of plant species covariates and/or linkage rules
x_{gjk}	gjk^{th} entry of $\mathbf{X}$
x_{gj}	$\mathbf{K}$-length vector of covariates in $\mathbf{X}$
β	$\mathbf{K}$-length vector of regression coefficients associated with $\mathbf{X}$
$\mathbf{Z}$	$\mathbf{G} \times \mathbf{L}$ matrix of pollinator species covariates
z_g	$\mathbf{L}$-length vector of pollinator covariates in $\mathbf{Z}$
γ	$\mathbf{L}$-length vector of regression coefficients associated with $\mathbf{Z}$
δ_g	overdispersion parameter
ρ_g	intra-group correlation coefficient
β^*	$\mathbf{M}$-length vector containing all model parameters

networks can also be represented by a bipartite graph, where the nodes represent plant and pollinator species and the undirected edges represent the mutualistic interactions between them. These networks may consist of hundreds of species that form highly complex and heterogeneous ecosystems (Bascompte and Jordano, 2007).

Studies have revealed that these networks share common structural patterns and that these structural patterns are being driven by both evolutionary and ecological processes, which can be explained by two opposing theories: neutrality and linkage rules. Neutrality states that all individuals interact randomly such that all plant-pollinator pairs have the same probability of interacting. As a result, individuals interact proportionately to their relative

abundance, i.e., more abundant species interact more frequently than rare species (Vázquez et al., 2009b). On the other hand, the linkage rule hypothesis posits that interaction patterns result from morphological, physiological, behavioural or evolutionary constraints that influence interaction probabilities between plants and pollinators (González-Castro et al., 2015; Jordano et al., 2003; Rezende et al., 2007; Santamaría and Rodríguez-Gironés, 2007; Dupont and Olesen, 2009; Mello et al., 2011; Olesen et al., 2011). For example, among phenotypic traits, the pollinator proboscis length and the plant tubal length will determine whether or not a pollinator can reach the reproductive organs of the plant for pollination. If these traits do not match, then an interaction cannot occur (also known as a forbidden link).

The structure of these networks can be described and quantified by calculating aggregate network metrics. The most common metrics include:

- Connectance - proportion of links that are actually realized (Jordano, 1987);
- Nestedness Metric based on Overlap and Decreasing Fill (NODF) - a measure of nestedness, where 0 indicates non-nestedness and 100 indicates perfect nesting. Nestedness is defined as the degree to which interactions can be arranged into subsets of the larger community;
- Relative Diversity - ratio of effective number of links (or interacting number of plant-pollinator species pairs) and the number of links in a network in which all potential links are equally common/strong;
- Linkage Density - marginal totals-weighted diversity of interactions per species (quantitative) (Bersier et al. 2002);
- Species Degree - number of different species to which a specific

  species is linked (Jordano et al., 2003); and
- Interaction strength or dependence - an estimate of the extent to which one species depends on another species, typically approximated by interaction frequency (Vázquez et al., 2005; Bascompte et al., 2006).

In Chapter 3, networks were generated according to the grouped DM model using relative species abundance and simple linkage rules as covariates. Network metrics were then calculated to demonstrate that the data generated from the DM model exhibit the patterns in observed plant-pollinator networks.

2.3 Non-linear Optimization Methods

In Chapter 4, three optimization methods are considered to estimate the DM model parameters: a modified NR, the BFGS, and the NM algorithm. What follows is a brief overview of each of these methods.

Modified Newton-Raphson
The NR method is a widely used method for finding the roots of a function, but can also be generalized to the problem of finding solutions to a system of nonlinear equations (Ypma, 1995). It is considered a higher-order optimization method since both the first and second derivative information of the objective function (i.e., log-likelihood function) is needed for each iteration of the algorithm. This property makes the NR method a very powerful technique because its order of convergence is quadratic, resulting in one of the fastest convergence rates among all methods. However, the disadvantages of using this method include: derivatives may not be available analytically; it is computationally expensive; and it has poor global convergence properties.

For maximum likelihood estimation, the NR method is used to find the solution to the problem[1]:

$$\ln L(\hat{\boldsymbol{\theta}}) = \underset{\theta \in \Theta}{\operatorname{argmax}} \ \ln L(\boldsymbol{\theta}),$$

where $\hat{\boldsymbol{\theta}}$ is a vector of maximum likelihood estimates (MLE) of the model parameters, $\ln L(\cdot)$ is the log-likelihood function, and $\boldsymbol{\theta}$ is the vector of unknown model parameters. Beginning at an initial value $\boldsymbol{\theta}_0$, the NR algorithm generates a sequence of iterates $\{\boldsymbol{\theta}_i\}_{i=0}^{\infty}$ and terminates when the difference between two successive iterates is less than some predefined value, e.g., $\epsilon = 1\text{e}{-8}$, indicating that a solution has been approximated with sufficient accuracy (Wright and Nocedal, 1999). More specifically, the NR algorithm uses both the first and second partial derivatives of the log-likelihood function, $g(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$[2], respectively, to form the descent direction which is used in each iteration update as follows:

1. Start with initial $\boldsymbol{\theta}_i$.

[1]Optimizers typically search for minima rather than maxima. As such, we convert to minimizing the negative log-likelihood function, i.e., $-\ln L(\hat{\boldsymbol{\theta}}) = \min_{\theta \in \Theta} -\ln L(\boldsymbol{\theta})$.

[2]First and second order optimality conditions $g(\boldsymbol{\theta}) = 0$ and $H(\boldsymbol{\theta}) \succeq 0$, respectively, must be satisfied.

2. Calculate a direction vector $\boldsymbol{d} = \{-H(\boldsymbol{\theta}_i)\}^{-1} g(\boldsymbol{\theta}_i)$.

3. Calculate a new guess $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \alpha_i \boldsymbol{d}$, where α_i is a scalar step length.[3]

4. Repeat $2 - 4$ until convergence.

The standard NR algorithm uses a step length of one, i.e., $\alpha_i = 1$; however, α_i is rarely fixed and is adjusted to produce a substantial reduction in $-\ln L(\boldsymbol{\theta})$. Another common modification to the standard NR is to use numerical derivatives for the direction vector d_i by finite-difference approximation (Jordán et al., 2000), which can be inaccurate particularly for non-smooth functions (Kolda et al., 2003). Consequently one may encounter instability and/or imprecision in parameter estimates. In `R`, the modified NR was implemented via the `nlm` function and second derivatives were approximated numerically for all parameterizations of the DM model.

Broyden-Fletcher-Goldfarb-Shanno

The BFGS method (suggested independently by Broyden, Fletcher, Goldfarb, and Shanno in 1970) is the most commonly used quasi-Newton technique for unconstrained nonlinear optimization. It builds on the variable metric technique known as the DFP method, a modification of Davidson's method by Fletcher and Powell (Davidson, 1959; Fletcher and Powell, 1963). Variants of the BFGS include the limited memory BFGS, or L-BFGS, suited for larger problems, and the BFGS-B, which handles simple box constraints. Only the standard BFGS is discussed and implemented in this paper.

BFGS approximates the NR method by replacing the Hessian, the matrix of partial second derivatives $H(\boldsymbol{\theta})$, by an approximation $B(\boldsymbol{\theta})$, which is updated iteratively from changes in the gradient, $g(\boldsymbol{\theta})$. The BFGS method has some advantages over the NR method because the Hessian does not need to be evaluated directly and the approximation matrix $B(\boldsymbol{\theta})$ is ensured, by explicit conditions, to be a symmetric positive definite matrix. However, the BFGS algorithm converges only superlinearly and may not be as fast as the quadratic convergence of the NR method. The BFGS algorithm is as follows:

1. Start with initial $B(\boldsymbol{\theta}_i)$ and $\boldsymbol{\theta}_i$.

[3] In computing the step length, α_i, we choose α_i to give a substantial reduction of $f(\boldsymbol{\theta}) = -\ln L(\boldsymbol{\theta})$, i.e., $\text{argmin}_{\alpha_i > 0} f(\boldsymbol{\theta}_i + \alpha_i \boldsymbol{d})$. The ideal choice is the global minimizer of this univariate function, which can be expensive to identify; therefore, both exact and inexact methods exist to identify a step length that achieves adequate reductions in f at minimal cost (Wright and Nocedal, 1999).

2. Solve $B(\boldsymbol{\theta}_i)\boldsymbol{d}_i = g(\boldsymbol{\theta}_i)$ for the direction vector $\boldsymbol{d}_i$.

3. Use a line search to determine the step length α_i.

4. Calculate a new guess $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \alpha_i \boldsymbol{d_i}$.

5. Compute $B(\boldsymbol{\theta}_{i+1})$ using the BFGS update:

   $B_{i+1} = \frac{B_i - (B_i s_i)(B_i s_i)^T}{s_i^T B_i s_i} + \frac{y_i y_i^T}{y_i^T s_i}$, where $s_t = \boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i$ and $y_i = g(\boldsymbol{\theta}_{i+1}) - g(\boldsymbol{\theta}_i)$.

6. Repeat $2 - 5$ until convergence is met.

The standard implementation of the BFGS algorithm, available in the `optim` function in `R`, was used to estimate the parameters of the DM model.

Nelder-Mead

The NM method (Nelder and Mead, 1965) is a heuristic search procedure used to find a minimum (or maximum) of an objective function in a multi-dimensional space. It is one of the most commonly used methods for unconstrained nonlinear optimization problems and falls in the class of direct search methods. The main advantage of the NM method is that it does not require any derivative information, so it can handle complicated functions for which first and second derivatives are tedious to derive analytically or are analytically intractable. However, a weakness of the method is that it may take an unnecessarily large number of function evaluations with negligible improvement in function value to locate a solution (Nash, 1990).

The method seeks to minimize f, where $f : \mathbb{R}^n \to \mathbb{R}$ is the (nonlinear) objective function and n is the dimension. It uses the concept of a simplex, which is a special polytope of $n + 1$ vertices in n dimensions. Simplex S in $\mathbb{R}^n$ is, therefore, defined as the convex hull of $n + 1$ vertices, $x_1, \ldots, x_{n+1} \in \mathbb{R}^n$. For example, a simplex in $\mathbb{R}^2$ is a triangle and a simplex in $\mathbb{R}^3$ is a tetrahedron. The method iteratively generates a sequence of $n + 1$ test points arranged in a simplex and the test point having the highest value is replaced by a new point with a lower function value. There are four main operations which are made on the simplex to determine a new test point: reflection (α), expansion (γ), contraction (β), and shrink (δ). The most common values chosen for these parameters in the standard NM algorithm are $\{\alpha, \gamma, \beta, \delta\} = \{1, 2, 0.5, 0.5\}$. One iteration of the NM algorithm is as follows:

1. **Sort**. Evaluate f at the $n+1$ vertices and sort the vertices such that $f(x_1) \leq f(x_2) \leq \ldots \leq f(x_{n+1})$, where x_1 is the best vertex and x_{n+1} is the worst vertex.

2. **Reflect**. Compute the reflection point x_r from

$$x_r = \bar{x} + \alpha(\bar{x} - x_{n+1}),$$

where $\bar{x} = \frac{1}{n}\sum_{n=1}^{n} x_i$ is the centroid of the n best vertices. Evaluate $f_r = f(x_r)$. If $f_1 \leq f_r < f_n$, replace x_{n+1} with x_r.

3. **Expand**. If $f_r < f_1$, then compute the expansion point x_e from

$$x_e = \bar{x} + \gamma(x_r - \bar{x}),$$

and evaluate $f_e = f(x_e)$. If $f_e < f_r$, replace x_{n+1} with x_e, otherwise replace x_{n+1} with x_r.

4. **Contract**. If $f_r \geq f_n$, perform a contraction between $\bar{x}$ and the better of x_{n+1} and x_r.

   (a) Outside Contraction. If $f_n \leq f_r < f_{n+1}$, compute the outside contraction point

$$x_{oc} = \bar{x} + \beta(x_r - \bar{x}),$$

   and evaluate $f_{oc} = f(x_{oc})$. If $f_{oc} \leq f_r$, replace x_{n+1} with x_{oc}, otherwise go to step 5.

   (b) Inside Contraction. If $f_r \geq f_{n+1}$, compute the inside contraction point

$$x_{ic} = \bar{x} + \beta(\bar{x} - x_{n+1}),$$

   and evaluate $f_{ic} = f(x_{ic})$. If $f_{ic} < f_{n+1}$, replace x_{n+1} with x_{ic}, otherwise go to step 5.

5. **Shrink**. Evaluate f at the n points

$$v_i = x_1 + \delta(x_i - x_1),$$

where $i = 2, \ldots, n+1$. The vertices of the simplex at the next iteration consist of $x_1, v_2, \ldots, v_{n+1}$.

The standard implementation of the NM algorithm available in the `optim` function in `R` was used to estimate the parameters of the DM model.

Chapter 3

DM Regression for Ecological Networks

A new model for ecological networks using species level traits[1]

Catherine Crea, R. Ayesha Ali, and Romina Rader

Abstract

1. Recent studies on plant-pollinator networks have focused on explaining network structure through linkage rules, including spatio-temporal overlap, and phenotypic trait or phylogenetic signal complementarity. Few studies, however, have quantified the extent to which functional traits affect the probability of plants and pollinators interacting with each other.

2. Dirichlet-multinomial (DM) regression is a consumer-resource model for the interaction probabilities in a mutualistic network. This flexible model accommodates network heterogeneity through random effects and over-dispersion, and can estimate the contribution of species level traits to plant-pollinator interactions.

3. Using artificial networks based on linkage rules and neutrality, we evaluate the performance of DM regression and explore the model's parameter space. We also analyze an empirical network in which the interaction probabilities are modeled by species characteristics.

4. Study results show that such random effects models can provide good fits to observed data. The characteristics pollinators seek in plant species may be better anticipated if species interactions are modelled by the functional traits that drive them.

Keywords: *pollination webs, pollen transfer network, Dirichlet-multinomial regression, interaction probabilities, linkage rules, forbidden links, complementary traits, network structure.*

[1]This chapter has been published in Methods in Ecology and Evolution and is cited as Crea et al., 2016 throughout this thesis. See bibliography for complete citation.

3.1 Introduction

Species and their interactions assemble into large, complex networks that shape and maintain ecological systems. Hence, effects upon one species may (in-)directly impact other species and processes such as pollination, seed dispersal, and host-parasitoid relationships (Montoya and Sole 2002, Montoya et al. 2006). Using adjacency matrices that represent these interactions, interest often lies in describing which species interact with each other; quantifying the frequency of interaction; and understanding the driving mechanisms of why species interact. However, several knowledge gaps need to be overcome to better predict which species are connected to each other. Although we focus on plant and pollinator systems, these methods can be easily adapted to other ecological systems.

Studies suggest that ecological networks are nested organizations of pairwise interactions (Bascompte and Jordano 2007; Jordano et al. 2006; Vázquez et al. 2009a; Montoya et al. 2006) that may be driven by both evolutionary and ecological processes. Functional traits, such as the varied morphological and behavioural characteristics of pollinator taxa, have been identified as strong candidates to quantify ecosystem service delivery (Kremen, 2005; Díaz et al., 2011; Lavorel and Grigulis, 2012) because of their effects on quantity and quality of pollination services. For example, body size measures correlate with pollination efficiency (Larsen et al., 2005), foraging duration (Stone and Willmer 1989; Stone, 1994), foraging distance in some bees (Greenleaf et al., 2007) and susceptibility to land-use change (Winfree et al., 2009; Williams et al., 2010). Pollinator preferences in plant characteristics, detectability of plant-pollinator links and the development of plant-pollinator linkage rules may thus be better anticipated if interactions are modeled as a function of these important functional attributes.

In fact, using functional traits or linkage rules as a proxy to infer biotic interactions has progressed our understanding about the dynamics and organisation of communities (Santamaría and Rodríguez-Gironés, 2007; Allesina et al., 2008; Stang et al., 2009; Blüthgen, 2010; Olesen et al., 2011; Bartomeus, 2013; Junker et al., 2013; Eklöf et al., 2013; Rohr et al., 2010; Gravel et al., 2013). Santamaría and Rodríguez-Gironés (2007) demonstrated that two to four linkage rules were sufficient in reproducing much of the structure in empirical networks. Eklöf et al. (2013) found that only a few traits were required to represent the structure of 200 bipartite networks and that three-trait models were often similar to all-trait models. For

food webs, body size tends to dominate species interactions (Gravel et al., 2013), though latent traits could increase the number of predicted links from 20% to 73% (Rohr et al., 2010).

Complementarily, empirical and theoretical works have supported neutral models, in which the probability of species interacting is proportional to their relative abundances. As such, network properties are the direct result of frequency distributions of different guilds. Several aspects of network structure could be explained by neutrality (Blüthgen, 2006; Vázquez et al., 2009a); however, the determinants of abundances themselves are not accounted for and are likely non-random. Olito and Fox (2015) used the multinomial model of Vázquez et al. (2009b) to evaluate the contribution of neutrality to network structure and found that not only are the best predictive models network metric dependent, but all failed to predict interaction frequencies. Poisot (2015) discussed the dynamic nature of interaction networks and noted that by neutrality, local variations in abundance can affect the overall network structure. However, even if two species meet, local trait distributions affect which interactions will be realized. Which theory dominates in the real world is unclear - it is likely that both theories contribute but few models incorporate them simultaneously. The model we propose here can accommodate both theories.

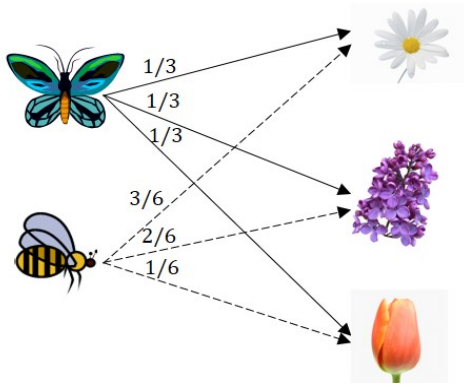

Figure 3.1: Pollination network with 2 pollinator and 3 plant species. Interaction probabilities appear above arrows.

The motivation of our research is to model interaction probabilities as a function of multiple factors much the way econometricians model consumers choosing among a set of brand products (Mosimann, 1962; McFadden, 1974; Hausman et al., 1984; Shonkwiler and Hanley, 2003; Guimarães and Lindrooth, 2007). In particular, we adapt Dirichlet-multinomial (DM)

regression (Guimarães and Lindrooth, 2007) to analyzing ecological networks such as pollination webs. We assume that pollinators assign a level of utility to every plant species it is faced with, based on the plant species attributes and random effects, and choose the ones that provide maximum utility. These utilities are then translated into interaction probabilities (see Figure 3.1). Individual pollinators of the same species choose plant species according to the same probabilities as each other, but they may ultimately select different plant species.

The model is called "Dirichlet-multinomial regression" because it is based on assuming that pollinator species choose plants species according to a multinomial distribution with multinomial (interaction) probabilities that are themselves random. These probabilities follow what is known as a Dirichlet distribution and are related to the species traits and/or linkage rules through the utilities mentioned above (Guimarães and Lindrooth, 2007). The unconditional distribution for the number of interactions is known as the Dirichlet-multinomial distribution (Mosimann, 1962). However, by Bayes Theorem, the posterior distribution for the interaction probabilities is itself a Dirichlet distribution. If there were only one pollinator species and two plant species in the system, then the multinomial distribution reduces to a binomial distribution, the Dirichlet distribution is equivalent to a beta distribution, and the (unconditional) distribution for the number of interactions is the beta-binomial distribution.

A benefit of DM regression is that it can account for pollinator species heterogeneity through an over-dispersion parameter. Since plant-pollinator networks are known to be heterogeneous (Bascompte and Jordano, 2007), the interactions of a given pollinator species may be observed more frequently than predicted by the multinomial distribution. Accordingly, the over-dispersion parameter is pollinator specific (possibly a function of pollinator traits) and accounts for this extra-multinomial variation. Ultimately, DM regression can be used to estimate the relative contribution of the factors that affect the interaction probabilities, which in turn provides insights into the mechanisms driving the observed network structure.

Unlike previous models, DM regression accommodates all of the following features and advantages: (i) extra-multinomial variation in observed counts can be modelled using an over-dispersion parameter; (ii) interaction probabilities can be modeled as a function of plant and pollinator characteristics and/or linkage rules; (iii) only detailed information at the species level is needed to supplement the quantitative observed network; and (iv) it is easy to fit the model, calculate standard errors and confidence intervals, and interpret the results. It

should also be noted that if the model is fit using only relative abundances, then the model is equivalent to assuming neutrality.

Our specific questions of interest are: (1) Is the DM model capable of simulating networks with diverse network structures? (2) Can the interaction probabilities be well estimated, given count data and detailed trait information? (3) Is the model robust to misspecifications of over-dispersion? (4) Even when linkage rule information is not available, can DM regression provide useful insights into which species characteristics contribute to network structure?

We address (1) to (3) by generating artificial plant-pollinator networks, driven by three linkage rules and relative species abundances, with varying dispersion structures (none, constant, function of traits, intra-correlation) and fitting all networks using DM regression assuming each of these dispersion structures. We address (4) by analyzing a real world network in which the factors that affect the interaction probabilities are unknown. Note that here we focus on the simplest scenario, in which there is only a single network of interest, so that the reader can better understand DM regression. Understanding the factors driving interactions in a particular ecosystem can facilitate development of improved species management strategies.

3.2 Materials and Methods

3.2.1 Dirichlet-multinomial regression

Consider a plant-pollinator network represented by a count matrix Y with G pollinator species (rows) and J plant species (columns). Interaction probabilities are modeled as a function of K plant species attributes or linkage rules similar to Vázquez et al. (2009b). In particular, we use X to represent the $G \times J \times K$ array of covariates (species attributes and/or linkage rules), with entries x_{gjk}, and use x_{gj} to represent the K-length vector of covariates associated with each pollinator species g and plant species j. We assume that the counts for pollinator species g follow a multinomial distribution with interaction probabilities π_{gj}, summarized as follows:

$$y_{g1}, \ldots, y_{gJ} \sim \text{multinomial}(\text{n}_{g1}, \ldots, \text{n}_{gJ}; \pi_{g1}, \ldots, \pi_{gJ}), \text{ for } g = 1, \ldots, G,$$

where n_{gj} is the number of interactions made between pollinator species g and plant species j, and the sum of the probabilities for pollinator g sums to one; $\sum_j \pi_{gj} = 1$. Further, we assume that each individual pollinator i of species g assigns a utility U_{igj} to plant species j as follows,

$$U_{igj} = \beta' x_{gj} + \eta_{gj} + \epsilon_{igj},$$

for $g = 1, \ldots, G$; $j = 1, \ldots, J$; and $i = 1, \ldots, I$, where $\beta' = (\beta_1, \beta_2, \beta_3, \ldots, \beta_K)$ is a K-length vector of unknown regression coefficients associated with the covariates in x_{gj}, η_g is a scalar random group effect, and ϵ_{igj} is a random error term for the i^{th} pollinator of species g that visits plant species j. Since pollinators seek plant species that maximize their utility, the errors follow independent (Type I) extreme value distributions; see Appendix A. It is these utilities that relate the interaction probabilities to the covariates using random utility theory.

To simplify the discussion, let us consider the special case in which there is only one pollinator species and two plant species. The above scenario simplifies considerably because we can omit the subscripts g and j, the number of visits across the two plant species reduces to a binomial distribution and there is only one interaction probability to estimate (by letting $\pi_1 = \pi$ and noting that $\pi_2 = 1 - \pi$). Further, the log-odds of visiting plant species 1 can be modeled by the average (expected) utility assigned to plant species 1,

$$\log \frac{\pi}{1-\pi} = \beta' x + \eta. \tag{3.1}$$

With a little re-arranging of terms, (3.1) can be shown to give interaction probabilities by:

$$\pi = \frac{\exp(\beta' x + \eta)}{1 + \exp(\beta' x + \eta)} \tag{3.2}$$

In the general case, in which there is more than one pollinator species and more than two plant species, equation (3.2) extends naturally and can be re-written as,

$$\pi_{gj} = \frac{\exp(\beta' x_{gj} + \eta_{gj})}{\sum_{j=1}^{J} \exp(\beta' x_{gj} + \eta_{gj})} = \frac{\lambda_{gj} + \exp(\eta_{gj})}{\sum_{j=1}^{J} \lambda_{gj} + \exp(\eta_{gj})}, \text{ for } g = 1, \ldots, G; \text{ and } j = 1, \ldots, J, \tag{3.3}$$

where $\lambda_{gj} = \exp(\beta' x_{gj})$. We further make the assumption that the exponential of the ran-

dom group effects, $\exp(\eta_{gj})$, follow independent gamma distributions with both shape and scale (rate) parameters $\delta_g \lambda_{gj}$, for some $\delta_g > 0$. Under these assumptions, Mosimann (1962) showed that the interaction probabilities for pollinator g follow a Dirichlet distribution (multivariate version of a beta distribution) with parameters $(\delta_g \lambda_{g1}, \delta_g \lambda_{g2}, \ldots, \delta_g \lambda_{gJ})$. The term δ_g is used to quantify the over-dispersion in the interaction counts.

The model has two main components: (a) the structural model, $\beta' x_{gj}$, that relates the plant attributes and/or (binary) linkage rules directly to the interaction probabilities, and (b) the random effects, η_g, that relate the pollinator traits with the model over-dispersion. The coefficient associated with the k^{th} covariate, can be interpreted as follows: $\exp(\beta_k)$ is the odds ratio of a pollinator species interacting with a plant species relative to a second plant species with the same traits as the first, but for which the value of the k^{th} covariate differs by one. The goal of fitting a DM regression is to unbiasedly estimate the regression coefficients in β and the dispersion parameter δ_g using maximum likelihood estimation.

3.2.2 Dispersion Structure

The over-dispersion associated with pollinator species g can be represented in several ways. The simplest case is constant over-dispersion across all pollinator species: $\delta_g = \delta$. However, sometimes it may be reasonable to assume that over-dispersion is a function of pollinator-specific traits: $\delta_g = \mathrm{f}(z_g) = \exp(\gamma_0 + \gamma_1 z_g)$, where γ contains coefficients to be estimated by the model and z_g is an L-length vector of pollinator specific characteristics (e.g., mean body size). Alternatively, over-dispersion can be modelled as a function of the intra-class correlation coefficient, ρ_g, which represents the correlation among individual pollinators within species g (Guimarães and Lindrooth, 2007). See Appendix A. In the absence of dispersion, DM regression is equivalent to a standard multinomial logistic regression model.

3.2.3 Simulation Design

We generated networks assuming one of four dispersion structures: none, constant ($\delta_g = \delta$), function of pollinator-specific covariates ($\delta_g = \exp(\gamma_0 + \gamma_1 z_g)$), and constant intra-correlation ($\rho_g = \rho$). There were four simulated covariates; three were Boolean operators that linked plant traits with pollinator traits per Santamaría and Rodríguez-Gironés (2007), giving $X = (x_1$: barrier trait, x_2 : complementarity trait with narrow range variability, x_3 : complementarity trait with medium range variability, x_4 : plant species relative abundances),

where x_k is a $G \times J$ matrix with entries x_{jgk}, $k = 1, \ldots, 4$. The associated regression coefficients used in the DM models were β_1, β_2, β_3, and β_4, respectively, which were then used to generate artificial networks. We analyzed these networks using DM regression to facilitate evaluation of its performance for settings in which the true parameter values were known.

3.2.4 Parameter Settings

The number of pollinator species was regressed on the number of plant species (on the square-root scale; Santamaría and Rodríguez-Gironés, 2007) for pollination networks in the Interaction Web Database (Guimarães et al., 2011); see Appendix F. Using this regression formula, artificial networks were generated with $J = 20,\ 40,$ or 120 plant species and $G = 53,\ 78,$ and 127 pollinator species, respectively. The values of β, δ, γ, and ρ were selected to allow exploration of the parameter space, using parameter ranges based on an ad-hoc pre-analysis (results not shown) that suggested the following. Coefficient values (β_k) between $(-1,\ 5)$, δ^{-1} values between (0, 12), dispersion coefficient values (γ) between $(-4,\ 2)$, and ρ values between (0, 1) generated networks that were neither too sparse nor too highly populated. Accordingly, parameter values were sampled from independent uniform distributions over low, medium and high sub-intervals of these respective ranges: $\{(-1, 1)\,(1, 3),\ (3, 5)\}$ for each β_k; $\{(0, 4),\ (4, 8),\ (8, 12)\}$ for δ^{-1}; $\{(-4, -2)\,(-2, 0),\ (0, 2)\}$ for each of γ_0 and γ_1; and $\{(0, 0.33),\ (0.33, 0.67),\ (0.67, 1)\}$ for ρ.

For each dispersion structure, networks were generated using parameter values set at the above levels according to a factorial design. Consider the no dispersion scenario. Low, medium and high values for each of the four β_k's were sampled and then all combinations of the coefficients at the different levels were used to generate $3^4 = 81$ networks according to a DM regression. This design was then replicated 10 times resulting in 810 networks in total. The constant and intra-correlation dispersion (δ^{-1} and ρ) scenarios included an additional dispersion parameter, resulting in $3^5 \times 10 = 2430$ networks each. Similarly, when dispersion was a function of pollinator covariates, there were two additional dispersion coefficients, resulting in $3^6 \times 10 = 7290$ networks in total. All scenarios were repeated for each of the three network sizes.

3.2.5 Covariate and Count Generation

Covariates x_1, x_2, and x_3 represented linkage rules per Santamaría and Rodríguez-Gironés (2007). Let V_g and W_j be the mean trait values for the gth pollinator species and j^{th} plant species, respectively, and ψV_g and ψW_j be their respective ranges in variability. V_g and W_j were generated from independent uniform distributions over (0, 1) and ψV_g and ψW_j were generated from independent uniform distributions over (0, 0.25) or (0, 0.5) for the narrow- and medium-range complementarity traits, respectively. Variability was ignored for the barrier trait. As such, the complementarity and barrier trait values were constructed using the following Boolean operators, respectively:

$$x_{gjk} = \begin{cases} 1, & \text{if } |V_g - W_j| < 0.5(\psi V_g + \psi W_j), k = 1, 2 \\ 0, & \text{otherwise} \end{cases} \qquad x_{gj3} = \begin{cases} 1, & \text{if } V_g > W_j \\ 0, & \text{otherwise} \end{cases}$$

where x_{gjk} is the k^{th} covariate for the g-j^{th} pollinator-plant pair. Covariates x_4 and z_1 represented relative species abundances for plant and pollinator species, respectively, and were generated using the inverse cumulative distribution function method (Devroye, 1986) to sample from the species abundance distribution (Ravasz et al., 2005). Conditional on the row sums, cell counts y_{gj} were generated from a Poisson distribution with rate λ_{gj} =exp($\beta' x_{gj}$). See Appendix C.

All networks were generated in R (R Development Core Team, 2011) and the following network metrics were calculated: connectance, links per species, mean number of shared partners, NODF, weighted NODF, interaction strength asymmetry, specialisation asymmetry, generality, vulnerability, linkage density, interaction evenness, relative diversity[2], H_2', and mean paired difference index (PDI) for both species levels (Dormann et al., 2009; Poisot et al., 2012). DM regression was also fit in R (Crea, 2014) to each network and Monte Carlo estimates (Robert and Casella, 2010) of relative biases (RB) were computed as,

$$RB = \frac{100}{R} \times \sum_{r=1}^{R} \frac{(\hat{\beta}_k^{(r)} - \beta_k)}{\beta_k},$$

where $\hat{\beta}_k^{(r)}$ is the estimate of the k^{th} regression coefficient from the r^{th} generated network

[2]Relative diversity = exp(H_2)/($G \times J$), where H_2 is Shannon's interaction diversity.

and β_k is the true parameter value used to generate the r^{th} network. Percent coefficients of variation were computed as,

$$CV = \frac{100}{\beta_k} \times \sqrt{\sum_{r=1}^{R} \frac{(\hat{\beta}_k^{(r)} - \beta_k)^2}{R-1}}.$$

Percent relative biases and percent coefficients of variation less than 100 are associated with good performance. Pearson χ^2 goodness-of-fit statistics were also used to evaluate DM regression and its robustness to misspecification of dispersion structure.

3.2.6 Description of Canterbury Data

Insect pollinator sampling was carried out using flight intercept and pan traps over 5-day periods every month for one year in the Canterbury plains region, on the South Island of New Zealand. Six replicate sites were selected within four different land use types. Data were pooled across land use type and time because the primary interest was in the range of plant-pollinator interactions across the landscape. All insects captured were stored in a laboratory freezer ($-80°$C) until further processing. Pollen-transport networks quantified the number and identity of pollen grains on the insects' bodies. Pollinator effectiveness was established via published literature or unpublished studies conducted in the same geographic area.

Plant and pollinator traits were compiled using existing published and unpublished datasets from the Canterbury region (Eklöf et al., 2013; Hudson et al., 2014). Plant traits included: (1) life span: short (annual/biennial) or long (perennial); (2) flower type: single flower or small cluster; inflorescence spike, raceme, panicle/cyme/thyrse, umbel/corymb; inflorescence capitulum/head; or inflorescence catkin; (3) ease of access to nectar: openly available or partly hidden; and (4) flowers per inflorescence: tens, hundreds or thousands. Pollinator traits included: (1) average body length (mm); (2) average body width (mm); (3) average body breadth (mm); (4) species behavior: social or solitary; and (5) larval feeding preference: nectar/pollen or decaying vegetation/animal/dung.

3.2.7 Construction of Canterbury Pollen Transfer Network

In total, there were $G = 16$ pollinator species, $J = 15$ plant species, and $N = 485$ interactions collected from 337 individual pollinator specimens. A plant-pollinator species pair was deemed to have interacted if at least one pollen grain from that plant species was detected on the body of an individual from that pollinator species. For example, if there were ten individuals of *Apis mellifera* that carried any pollen grains from *Raphanus sp.*, then we recorded ten interactions between *Apis mellifera* and *Raphanus sp.* in our network, regardless of how many grains were found on each individual. Counts in the resulting 16×15 pollen transfer network represent the minimum number of interactions required to explain the observed transfer of pollen to the pollinator species (Appendix D). A stepwise selection method, based on examination of p-values from Pearson χ^2 goodness-of-fit statistics, and Akaike's Information Criterion (AIC), were used to determine which plant traits and pollinator traits (for over-dispersion) were most compatible with the data.

3.3 Results

3.3.1 Simulation Study: Network Statistics

In general, simulated networks exhibited a diverse range of network structures, regardless of the assumed dispersion. Under no dispersion, or dispersion in terms of δ: increasing trends in median were observed for vulnerability, generality, linkage density, links per species, and mean number of shared partners (both levels) while little to no trend was observed for the median of other metrics. Interestingly, under constant intra-correlation: median connectance, NODF, and weighted NODF tended to decrease and the median values were smaller overall relative to other dispersion structures; other metrics showed little to no trend. For brevity, only four metrics are shown in Figure 3.2; see Appendix E for all other metrics. Connectance ranged from 0.02 to 0.97 across all generated networks, with the highest values and least variability arising from networks generated assuming no dispersion. These findings are compatible with Dormann et al. (2009), who reported connectance between 0.37 to 0.55 for similar sized networks, and with Vázquez et al. (2009a) who found low connectance in published empirical networks. NODF ranged from 0 to 92, with a slight increasing trend with network size. Relative diversity ranged from 0.00 to 0.74 across all generated networks, with little to no trend as network size increased. Note that in the empirical networks available in the bipartite package in R, NODF ranged from 7.7 to 84.9 and relative diversity ranged

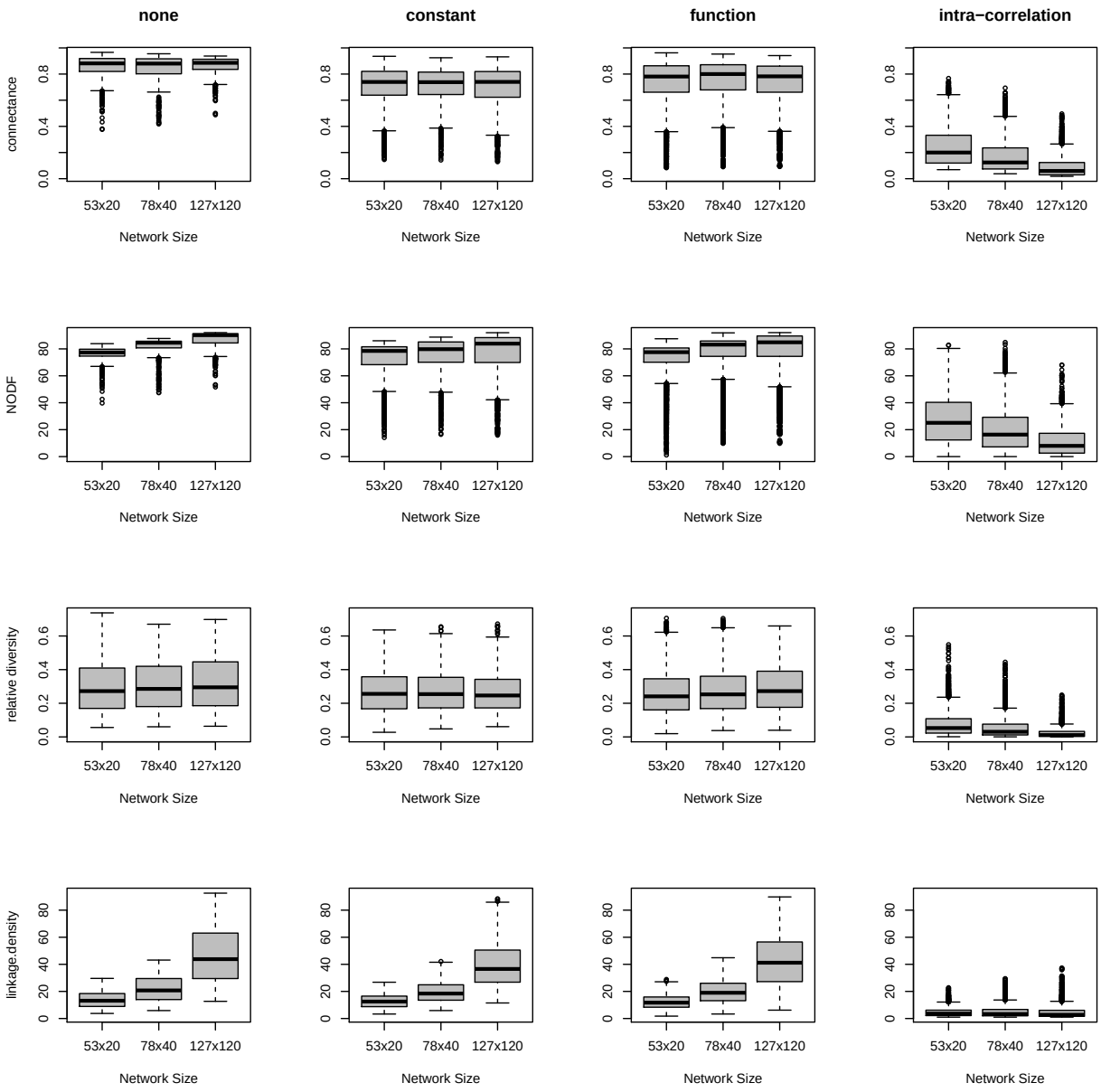

Figure 3.2: Network metric distributions of simulated networks by network size and dispersion structure. None: $\delta_g = 0$; Constant: $\delta_g = \delta$; Function: $\delta_g = f(z_g)$; Intra-correlation: $\rho_g = \rho$.

from 0.01 to 0.17. Finally, linkage density ranged from 1 to 93, with increasing values and variability seen with increasing network size. This metric is the average of vulnerability and generality and is known to be low in empirical networks (Bersier et al. 2002; Tylianakis et al., 2007). Dormann et al. (2009) suggested that linkage density slightly increases with

network size and reported values between 0 to 40 for similar sized networks.

3.3.2 Simulation Study: DM Models

In general, when networks were analyzed assuming the correct dispersion structure, DM regression could recover the regression coefficients with low bias (Tables 3.1 and 3.2), as evidenced by most of the percent relative biases being far below 100%. A closer look at the biases showed that most of the highly biased estimates were associated with either coefficients that were close to zero ($-0.5 < \hat{\beta}_k < 0.5$) and/or high intra-correlation ($\rho > 0.67$). When data was generated with no over-dispersion, the estimates obtained by assuming otherwise tended to produce small relative biases for the binary covariates (linkage rules). Under constant intra-correlation, the average relative biases tended to be larger, but decreased for the binary covariates as network size increased. The relative bias of $\hat{\beta}_4$ (continuous covariate for plant species RA) was greater than that of the binary covariates and tended to increase with network size under constant intra-correlation.

Similar trends were seen in the percent coefficients of variation for $\hat{\beta}$. If data were generated with no over-dispersion, then all methods seemed competitive, though models assuming over-dispersion tended to have smaller coefficients of variation. Further, the percent coefficient of variation values were often close to or much greater than 100% when the true β was close to zero or ρ was high. Overall, modeling dispersion in terms of δ_g or no dispersion seemed more robust to model misspecification compared to modeling dispersion in terms of ρ.

Less than 10% of the time, the χ^2-statistics indicated a poor fit to the data (p-values < 0.05) when networks generated with no over-dispersion and correctly fit assuming no over-dispersion. When those same networks were analyzed assuming some over-dispersion the model fit was quite good (p-values > 0.05). On the other hand, when networks were generated with some over-dispersion but were analyzed assuming no over-dispersion, the χ^2-statistics indicated poor model fits more than 85% of the time. However, the χ^2-statistics indicated poor model fits only 1% to 25% of the time when analyzed correctly assuming some over-dispersion. These findings suggest that models assuming some over-dispersion are statistically more robust to misspecification of dispersion structure.

Table 3.1: (Percent Relative Bias) and [Percent Coefficient of Variation] of $\hat{\beta}$.

Network	**True**	**Modeled Dispersion***		
Size	**Dispersion**	**None**[1]	**Constant**[2]	**Intra-correlation**[3]
Percent Relative Bias				
53×20	None	(6.8, 3.4, 3.41, 39.96)	(7.17, 3.6, 3.62, 42.71)	(6.98, 4.48, 3.79, 41.38)
	Constant	(9.86, 5.66, 8.47, 15.14)	(7.93, 4.86, 7.01, 12.92)	(8.69, 5.35, 7.82, 14.44)
	Intra-correlation	(1241.05, 137.34, 100.25, 747.69)	(523.99, 59.05, 42.98, 139.41)	(522.35, 58.66, 41.67, 188.43)
78×40	None	(1.22, 1.19, 3.85, 8.74)	(1.26, 1.24, 4.14, 9.17)	(1.29, 1.29, 4.23, 9.34)
	Constant	(3.64, 12.07, 2.39, 32.62)	(2.86, 10.07, 2.05, 30.18)	(3.47, 10.85, 2.23, 32.84)
	Intra-correlation	(115.89, 62.73, 68.69, 1419.86)	(50.31, 29.41, 30.22, 469.91)	(36.63, 24.84, 26.53, 366.75)
127×120	None	(2.52, 1.06, 0.68, 6.6)	(2.46, 1.07, 0.67, 6.99)	(3.01, 1.04, 0.78, 7.55)
	Constant	(3.62, 1.95, 1.52, 27.44)	(2.72, 1.62, 1.22, 27.25)	(3.51, 1.72, 1.41, 25.92)
	Intra-correlation	(38.32, 37.02, 45.35, 5481.41)	(31.39, 14.75, 19.05, 4710.98)	(13.19, 11.13, 15.01, 2221.9)
Percent Coefficient of Variation				
53×20	None	[29.59, 11.55, 14.59, 325.23]	[30.51, 11.95, 15.54, 337.56]	[29.95, 23.21, 15.95, 329.48]
	Constant	[35.27, 14.13, 29.31, 42.15]	[26.02, 10.74, 23.68, 32.33]	[25.42, 11.05, 23.77, 33.91]
	Intra-correlation	[6311.23, 509.79, 360.9, 12142.99]	[2420.52, 217.02, 133.62, 381.55]	[2497.91, 262.98, 148.36, 1976.09]
78×40	None	[2.58, 3.87, 21.88, 35.73]	[2.62, 3.98, 22.35, 35.35]	[2.78, 4.34, 23.04, 39.61]
	Constant	[8.8, 78.65, 5.33, 117]	[6.31, 62.95, 4.37, 102.43]	[7.08, 68.72, 4.41, 103.74]
	Intra-correlation	[517.83, 206.06, 198.58, 14640.54]	[105.21, 163.37, 70.61, 2309.82]	[120.03, 148.92, 72.7, 1801.92]
127×120	None	[10.3, 5.33, 2.28, 14.92]	[10.19, 5.44, 2.23, 18.46]	[14.49, 4.05, 2.64, 17.92]
	Constant	[20.97, 6.44, 4.21, 96.21]	[13.49, 4.61, 2.84, 97.44]	[14.35, 4.63, 3.11, 77.4]
	Intra-correlation	[101.99, 108.03, 101.36, 46392.42]	[34.74, 30.76, 40.05, 40616.03]	[38.82, 32.12, 40.04, 17022.77]

* None: $\delta_g = 0$; Constant: $\delta_g = \delta$; Intra-correlation: $\rho_g = \rho$

[1] For true dispersion Intra-correlation: based on the 99.5% (53×20) and 99.6% (78×40) networks for which parameter estimates converged.

[2] For true dispersion None: based on the 93% (53×20), 94% (78×40) and 96% (127×120) networks for which parameter estimates converged. For true dispersion Intra-correlation: based on the 99.7% (53×20 and 78×40) networks for which parameter estimates converged.

[3] For true dispersion None: based on the 97% (53×20 and 78×40) and 96% (127×120) networks for which parameter estimates converged. For true dispersion Intra-correlation: based on the 99.9% (53×20) networks for which parameter estimates converged.

Table 3.2: (Percent Relative Bias) and [Percent Coefficient of Variation] of $\hat{\beta}$ for data generated with $\delta_g = \mathrm{f}(z_g)$.

Network	**Modeled Dispersion***			
Size	**None**	**Constant**	**Intra-correlation**	**Function**[1]
Percent Relative Bias				
53×20	(7.39, 9.48, 18.42, 16.3)	(5.85, 8.15, 14.91, 14.34)	(6.85, 8.86, 16.46, 15.98)	(5.83, 8.05, 15.33, 14.48)
78×40	(5.85, 32.99, 722.19, 16.04)	(4.83, 22.6, 458.84, 13.7)	(5.67, 24.71, 502.77, 14.83)	(5.11, 21.68, 485.65, 14.02)
127×120	(1.99, 1.92, 3.76, 20.23)	(1.58, 1.63, 2.82, 21.08)	(2.17, 1.81, 3.09, 19.68)	(2.36, 2.13, 13.1, 19.47)
Percent Coefficient of Variation				
53×20	[25.03, 35.71, 142.11, 55.89]	[16.41, 30.74, 108.41, 47.99]	[17.97, 30.62, 115.62, 50.71]	[16.22, 30.72, 110.9, 47.56]
78×40	[21.11, 351.7, 7546.71, 66.27]	[16.74, 198.81, 3848.9, 51.61]	[18.49, 209.99, 4222.44, 52.14]	[25.06, 195.98, 4008.67, 56.62]
127×120	[6.61, 8.12, 30.03, 62.34]	[4.64, 6.13, 16.14, 67.42]	[5.32, 6.26, 16.61, 51.5]	[52.95, 28.03, 806.3, 58.87]

* None: $\delta_g = 0$; Constant: $\delta_g = \delta$; Intra-correlation: $\rho_g = \rho$; Function: $\delta_g = \mathrm{f}(z_g)$

[1] Statistics based on the 96%, 90%, and 90% networks, respectively, for which parameter estimates converged.

3.3.3 Empirical Study

We found that, for this system, all plant traits but lifespan were important for modeling plant and pollinator interactions (Table 3.3). Most markedly, the estimated odds ratio of a pollinator interacting with two floral species having the same attributes, except that one has an inflorescence with a spike, raceme, panicle/cyme/ thyrsi and the other a single or small cluster of flowers, was 3.48 ($p < 0.001$). Interestingly, after adjustment for constant dispersion, no pollinator traits were statistically significant. We re-fit the model assuming no dispersion and found that the odds ratios associated with having thousands of flowers (0.49; $p = 0.001$), with inflorescence capitulum/head (2.93; $p < 0.001$) and with inflorescence catkin (0.17; $p < 0.005$) were more statistically significant compared to the respective odds ratios assuming constant dispersion. Further, the χ^2-statistic suggested a good fit of the constant dispersion model to the data ($p = 0.797$), but a poor fit of the no dispersion model ($p < 0.001$).

Table 3.3: Odds ratios of plant-pollinator interactions for Canterbury data.

	Constant Dispersion			**No Dispersion**		
Variable	**Odds Ratio**	**95% CI**	***p*-value**	**Odds Ratio**	**95% CI**	***p*-value**
Type of Flower:						
Single flower or small cluster	1			1		
Inflorescence spike, raceme, etc.	3.48	(1.69, 7.18)	0.001	5.02	(3.14, 8.02)	< 0.001
Inflorescence captitulum/head	2.31	(0.89, 6.02)	0.086	2.93	(1.62, 5.29)	< 0.001
Inflorescence catkin	0.24	(0.06, 0.97)	0.045	0.17	(0.05, 0.57)	0.005
Ease of Access to Pollen/Nectar:						
Easy access	1			1		
Partly hidden nectar	2.15	(1.32, 3.51)	0.002	2.35	(1.81, 3.05)	< 0.001
Flower Density per Floral Unit:						
Tens of flowers	1			1		
Hundreds of flowers	1.88	(1.12, 3.13)	0.016	1.6	(1.20, 2.14)	0.002
Thousands of flowers	0.85	(0.46, 1.58)	0.611	0.49	(0.32, 0.75)	0.001
Dispersion:						
Constant	2.95	(1.23, 6.72)	0.01	NA	NA	NA

The final model predicted slightly higher values for most network metrics (Table 3.4), particularly for connectance, NODF, and linkage density. However, predicted values of other network metrics, including vulnerability, interaction evenness, relative diversity, and mean

Table 3.4: Network metrics for Canterbury data.

Network Metric	**Observed**	**Predicted**
Connectance	0.463	0.622
NODF (nestedness)	74.487	83.556
$H2'$	0.0999	0.000
Linkage Density	6.454	8.495
Vulnerability	5.409	5.929
Interaction evenness	0.702	0.775
Relative diversity	0.195	0.278
Modularity (Q)	0.126	0.047
Mean PDI (higher level)	0.833	0.574
Mean PDI (lower level)	0.910	0.943

PDI (lower level), were similar to the observed values of metrics. The predicted metrics were calculated assuming a network with the same number of interactions observed for each pollinator species as that actually observed in the study.

3.4 Discussion and Conclusions

We modeled the relative contributions of linkage rules (simulation study) and of functional traits (empirical study) to the interaction probabilities that determine the topological features of plant-pollinator networks using DM regression (Guimarães and Lindrooth, 2007). Although the linkage rules we used were rather simplistic, more than one fixed effect could be used to describe a single linkage rule. Alternatively, higher level interaction terms could be included in the model, corresponding to the product of two rules (e.g. include a covariate for $(x_1 \times x_3)$ in the model, where x_1 and x_3 represent two different rules). Regardless, the DM model was still able to generate networks that exhibited a diversity of network structures, with some very similar to those observed in real world networks. However, the simulated networks were not designed to account for sampling bias, which is a known causal effect of observed network structure (Vázquez et al., 2009a). Since the goal of the simulation study was to evaluate the performance of DM regression in a setting where the true parameter values are known, the effect of these discrepancies on the results drawn from the study are expected to be minimal.

The bias and standard errors of parameter estimates were typically low, but increased when

the values of β_k were too close to zero or intra-correlation was high. Models that assumed no or constant over-dispersion were statistically more robust to model misspecification relative to assuming other dispersion structures. However, the χ^2-statistics suggested that assuming some over-dispersion was less prone to overfitting. Estimates associated with continuous covariates showed larger bias, but this may not pose a problem in practice if covariates correspond to binary linkage rules.

Note that we did not encounter problems recovering structural zeroes in the generated networks because their presence was driven by the linkage rules included in our model. In practice, it is anticipated that either an important trait is left out of the analysis (e.g. too difficult to measure), or it is encoded incorrectly when constructing a covariate. Accordingly, one may expect to find data that is zero-inflated because the fitted model does not account for all observed zeroes. Understanding the behavior of DM regression in the presence of zero-inflation or sampling bias are important modeling issues, but are outside the scope of this paper and the focus of future research. Nonetheless, the results of the simulation study provide useful information about the performance of DM regression as a statistical model, and how to use the analysis results to infer an appropriate dispersion structure.

In the empirical analysis, the actual number of interactions made to collect the amount of observed pollen was unknown. Because the pollen from a given plant species found on an individual pollinator was assumed to be the result of a single plant interaction, the pollen transfer network likely underestimated the true number of interactions. However, since pollen grains were physically found on the body of the pollinator, it is more likely that the recorded counts would have led to a pollination event relative to counts in a visitation web (Alarcón, 2010; Popic et al., 2013).

Since an individual pollinator may have had more than one type of pollen on its body, multiple interactions in the pollen transfer network were from the same individual pollinator. Visitation webs have a similar drawback since multiple visits in a visitation web may have been made by the same individual pollinator. If visitation web data were available for these data, then a pollen-transport network (Alarcón, 2010) could have been constructed, but the cost of collecting such data was beyond the financial scope of the empirical study.

Pollinators tended to forage on plant species with inflorescences and/or many flowers, likely

due to increased visual attraction and greater floral rewards of larger floral displays (Torres and Galetto, 1998). They were also more likely to visit flowers with partly hidden floral resources, perhaps reflecting the spatio-temporal resource availability within the study area and/or the composition, abundance and resource preferences of the pollinator community (Heinrich, 1979; Willmer, 2011). For example, taxa with long tongue lengths or unique flower handling behaviours may access hidden resources more effectively than others. In the absence of nectar-accessible flowers, pollinator preferences may shift to those that require greater handling time.

Eklöf et al. (2013) analyzed a binary version of the Canterbury network and found that pollinator body width was the single trait that explained the largest fraction of links in the network. Other important traits included amount of nectar and flower type (see Table 1), resulting in 91% of the links in the binary network being explained by their model. The plant attributes we identified as significant are in line with Eklöf et al. (2013). Interestingly, we found flower type, not body width, to be the single trait with highest contribution to the interaction probabilities in the network. In fact, we found that only constant dispersion was needed to accommodate pollinator heterogeneity, rather than a pollinator species body size measure. However, it should be noted that Eklöf et al. (2013) did not consider over-dispersion in their model, which may explain this discrepancy.

Analogous to the simulation study, the model assuming over-dispersion provided a better fit to the data. Note that the coefficients associated with inflorescence catkin and with floral density 'thousands of flowers' were based on very few floral species (≤ 2) in each category, so corresponding confidence intervals for these parameter estimates under the constant dispersion model were wider than their no dispersion counterparts, resulting in less significant estimates. Hence, the constant dispersion model appears to be more robust to potential outliers.

The properties of the network predicted by DM regression were also in line with trends found in the literature. For example, Nielsen and Bascompte (2007) found that nestedness is relatively robust to sampling effort and Chacoff et al. (2012) found that the number of interacting species was typically underestimated, estimated to be as much as 55% for a visitation web based on two study sites observed over 4 flowering periods. Our predicted weighted nestedness was close to the observed nestedness, while the predicted connectance

was approximately 1.5 times the observed connectance.

3.4.1 Relation to Other Methods

Santamaría and Rodríguez-Gironés (2007) assumed that all plant and pollinator species that could interact did interact, and that each rule contributed equally to the process. Although our covariates were similar to theirs, we applied differential weights (coefficients) to each rule using our probabilistic approach. We also included species relative abundances. The dimension-search method of Eklöf et al. (2013) plots the lower trophic species in trait-dimensions and seeks the fewest number of traits to explain all links in the network. Rohr et al. (2010) use logistic regression to model the odds of a link using body size ratios. When latent traits are included, Markov Chain Monte Carlo (MCMC) methods are needed. Gravel et al. (2013) also use the allometric scaling relationship between predator and prey body sizes. All of these models use binary networks; dimension-search and logistic regression do not accommodate over-dispersion.

Vázquez et al. (2009b) modeled network interactions assuming counts follow a multinomial distribution with probabilities based on relative abundances, spatio-temporal overlap, phenotypic traits, phylogenetic signal, and sampling effects. Sorensen et al. (2012) used the Dirichlet-multinomial distribution to estimate standard deviations of interaction frequencies, and Wells and O'Hara (2012) incorporated both individual level and species level data in a Poisson random effects model for interaction probabilities. The former model does not incorporate any covariate information, while the latter model requires Gibbs sampling to fit the model.

The focus of several researchers has been to study interaction networks at the macro level (Eklöf et al., 2013; Bartomeus, 2013) and identify common trends in network structure across several networks. However, our focus here has been on individual networks. To use DM regression to model several networks, the relevant factors for each network would need to be determined and measured in order to fit the model. Then, the predicted values of the network metrics, or the contributions of the common factors, could be compared across the networks.

In summary, we have presented a novel approach to modeling interaction probabilities in

a plant-pollinator network. This approach can quantify the contributions of diverse traits and/or linkage rules to pollination, thereby providing insights into the mechanisms that drive this ecological process. However, DM regression could also be used to model other ecological networks such as seeder-disperser or host-parasite relationships.

Acknowledgements

We thank Peter Kevan, Thomas Woodcock and Stefano Allesina for insightful discussions; Ignasi Bartomeus, Timothé Poisot, Sébastian Ibanez and an anonymous reviewer for constructive comments on the manuscript; and Anna Eklöf for access to data owners. This paper is contribution #98 of NSERC-CANPOLIN.

Data Accessibility

Code to generate and analyze artificial plant-pollinator networks in R, per the simulation study, is available on GitHub (Crea, 2014). The empirical Canterbury data has been archived in the PREDICTS database (Projecting Responses of Ecological Diversity In Changing Terrestrial Systems - www.predicts.org.uk) as of December, 2014 (http://dx.doi.org/10.5519/0018993).

Chapter 4

Optimizing the grouped Dirichlet-multinomial regression model

On Optimization Methods for Grouped Dirichlet-multinomial Regression

Catherine Crea and R. Ayesha Ali

Abstract

Dirichlet-multinomial (DM) regression for grouped data is becoming more popular in econometrics and ecology. It provides a flexible approach to modelling overdispersed grouped data and is fully parametric, but has not been well studied. In this paper, we study the behaviour of various parameterizations of the DM likelihood, evaluate the performance of three optimization methods (derivative and derivative-free), and assess the robustness to misspecification of dispersion structure. Through a comprehensive simulation study, we show that the DM likelihood becomes flatter with increased dispersion and that the it can have regions of non-smoothness. We also found that the (derivative-free) Nelder-Mead method is more robust to non-smoothness and poor starting values for the dispersion parameters. Although, the quasi-Newton BFGS method performs better for the parameterizations in terms of the intragroup correlation coefficient. Finally, the parameterizations in terms of the intragroup correlation coefficient are the least robust to misspecification of dispersion structure. These findings can simultaneously facilitate the selection of the appropriate optimization method and the appropriate parameterization which are both practical issues faced by scientists. We provide our implementation of DM regression for grouped data as a publicly available routine in `R` and demonstrate its performance in the analysis of several empirical plant-pollinator networks. Other supplemental information is available online.

Keywords: *overdispersion, plant-pollinator networks, likelihood slices, quasi-Newton methods.*

4.1 Introduction

Dirichlet-multinomial (DM) regression has been gaining popularity for modelling categorical count or compositional data from diverse fields, including econometrics (Hausman et al., 1984; Shonkwiler and Hanley, 2003; Guimarães and Lindrooth, 2007), bioinformatics (Zhou and Lange, 2010; Zhou and Zhang, 2012; Chen and Li, 2013; Zhang et al., 2017), topic modelling (Mimno and McCallum, 2012; Klami et al., 2015), and ecology (Crea et al., 2016; de Valpine and Harmon-Threatt, 2013). Scientists have turned to the DM model because it can accommodate overdispersion and maximum likelihood estimation (MLE) can be achieved via unconstrained nonlinear optimization methods, such as the Newton-Raphson or Quasi-Newton methods. This paper provides an evaluation and comparison of the performance of such optimization methods under various parameterizations of the DM likelihood for grouped data.

A multinomial logit model that incorporates category-specific variables, which may include interactions with individual-specific variables, is commonly referred to as a conditional logit model and is used in econometrics to study discrete choice behaviour. The model is derived from a random utility model (RUM) framework (McFadden, 1974), which assumes that individuals ascribe a level of utility to each choice and select the one with maximum utility. When these models are applied to data aggregated into distinct groups, with each group being represented by a vector of counts across the choices, the data are analyzed at the group level, not the individual level, and are known as the grouped conditional logit. The DM regression model parallels the structure of the grouped conditional logit in that the log odds are modelled as a function of choice- and/or group-specific characteristics, but the DM model accommodates overdispersion in the grouped count data.

The logit formulation for the multinomial probabilities is the result of the Type I extreme value (or standard Gumbel) distribution being placed on the utility function errors. When additional random effect terms are included in the utility function, heterogeneity in the counts (i.e., overdispersion) and/or varying correlation structures among choices provide a more flexible framework for discrete choice models. In the econometrics literature, the mixed logit model is the most popular choice (McFadden and Train, 2000); however, the DM model is a fully parametric alternative when (i) true variation is at the group level; (ii) groups of individuals are faced with the same choices; and (iii) there are a large number of choices

(Guimarães and Lindrooth, 2007).

Early implementations of the DM regression model for choice data date back to Shonkwiler and Hanley (2003) in which they show that accounting for overdispersion in the choice of popular rock climbing sites improves the precision of parameter estimates. Guimarães and Lindrooth (2007) derived the DM model for overdispersed grouped data under several dispersion structures and applied them to a patient hospital choice data set. Crea et al. (2016) applied DM regression to ecological networks, where plant and pollinator species interactions were modelled as a function of species level traits, and performed an exploration of the model's parameter space.

Zhang et al. (2017) noted that the DM model is not part of the natural exponential family and shows regions of non-convexity in the negative log-likelihood. In highly non-convex regions, the positive semidefinite property may be violated, thereby leading to Hessian non-invertibility. As such, the standard Newton-Raphson algorithm can become unstable and alternative optimization methods that are derivative-free or Quasi-Newton methods may be preferable. Guimarães and Lindrooth (2007) have implemented grouped DM regression in Stata using numerically approximated derivatives within the context of a modified NR algorithm. However, to date an in-depth study of optimization methods for grouped DM regression is not available.

Here, we provide a comprehensive evaluation of the DM likelihood under various parameterizations and identify a suite of optimizers for DM regression. In particular, our objectives are three-fold: (i) to study the behaviour of the DM model and its associated likelihood under various parameterizations; (ii) to compare the performance of three optimization methods: Broyden-Fletcher-Goldfarb-Shanno (BFGS), Nelder-Mead (NM), and the modified Newton-Raphson (MNR), with respect to rate of convergence and Hessian stability; and (iii) to assess the robustness of the DM regression model to model misspecification. Section 2 provides a formal outline of several parameterizations of the grouped DM regression model and briefly reviews the candidate optimization methods. Section 3 studies the behaviour of the DM likelihood and evaluates the performance of the optimizers via simulation. Section 4 applies DM regression to ecological data comprised of twelve empirical plant-pollinator networks from the Mediterranean region. Finally, Section 5 highlights the relation of other work to our grouped implementations of DM regression.

4.2 Methods

4.2.1 Dirichlet-multinomial distribution

The DM distribution is a compound probability distribution that results from assuming that the parameters of a multinomial distribution are themselves random and follow a Dirichlet distribution. This distribution is particularly useful for overdispersed count variables that exhibit more variability than that of the standard multinomial distribution and arises from marginalizing over the multinomial parameters (Mosimann, 1962). More formally, consider a matrix of categorical counts Y, with rows representing G groups and columns representing J choices. Each cell $y_{gj},\ g = 1, \ldots, G$ and $j = 1, \ldots, J$, represents the number of times a member of group g selects choice j and the row sums are $\boldsymbol{n}_g = \sum_{j=1}^{J} y_{gj}$. It is assumed that the count data follow a multinomial distribution with probability matrix P, which consists of row probability vectors $\boldsymbol{p}_g = \{p_{g1} \ldots p_{gJ}\}$ and probability mass functions:

$$f_{MN}(\boldsymbol{n}_g; \boldsymbol{p}_g) = \frac{n_g!}{\prod_{j=1}^{J} y_{gj}!} \prod_{j=1}^{J} p_{gj}^{y_{gj}},\ g = 1, \ldots, G. \tag{4.1}$$

Further, it is assumed that the probability vectors $\boldsymbol{p}_g$ follow a Dirichlet, or multivariate beta, distribution with corresponding parameter vectors $\boldsymbol{\alpha}_g = \{\alpha_{g1} \ldots \alpha_{gJ}\}$ and probability density functions:

$$f_D(\boldsymbol{p}_g; \boldsymbol{\alpha}_g) = \frac{\Gamma(\alpha_g)}{\prod_{j=1}^{J} \Gamma(\alpha_{gj})} \prod_{j=1}^{J} p_{gj}^{\alpha_{gj}-1},\ g = 1, \ldots, G, \tag{4.2}$$

where $\alpha_g = \sum_{j=1}^{J} \alpha_{gj}$. Accordingly, the DM probability mass functions are obtained by:

$$\begin{aligned} f_{DM} &= \int_{\boldsymbol{p}} f_{MN}(\boldsymbol{n}_g; \boldsymbol{p}_g) f_D(\boldsymbol{p}_g; \boldsymbol{\alpha}_g) \mathrm{d}\boldsymbol{p} \\ &= \frac{n_g! \Gamma(\alpha_g)}{\Gamma(\alpha_g + n_g)} \prod_{j=1}^{J} \frac{\Gamma(\alpha_{gj} + y_{gj})}{\Gamma(\alpha_{gj}) y_{gj}!},\ g = 1, \ldots, G. \end{aligned} \tag{4.3}$$

The mean and variance of each count variable, y_{gj}, conditional on the row sums n_g, are given by:

$$\begin{aligned} E(y_{gj}|n_g) &= n_g \mathrm{E}(p_{gj}), \text{ and} \\ \mathrm{Var}(y_{gj}|n_g) &= n_g \mathrm{E}(p_{gj})(1-\mathrm{E}(p_{gj}))\left(\frac{n_g+\alpha_g}{1+\alpha_g}\right), \end{aligned} \tag{4.4}$$

respectively, where $E(p_{gj}) = \frac{\alpha_{gj}}{\sum_{j=1}^{J}\alpha_{gj}}$. The last term in (4.4) inflates the variance in order to accommodate overdispersion.[1] It is this additional term which distinguishes the DM distribution from the multinomial distribution.

4.2.2 Derivation of DM regression model for grouped count data

DM regression for grouped data assumes that individuals within a group share common characteristics and are faced with the same choice set; hence, the true level of variation is at the group level (Guimarães and Lindrooth, 2007). Let X be a $G \times J \times K$ design array containing covariate information, with entries x_{gjk}. We use $\boldsymbol{x}_{gj}$ to represent the K-length vector of covariates associated with group g and choice j. The multinomial probabilities $\boldsymbol{p}_g$ are modelled as a function of covariates through a logit formulation,[2]

$$\log\left(\frac{p_{gj}}{1-p_{gj}}\right) = \boldsymbol{\beta}'\boldsymbol{x}_{gj} + \eta_{gj}, \tag{4.5}$$

for $g = 1, \ldots, G$ and $j = 1, \ldots, J$, where $\boldsymbol{\beta}'$ is a K-length vector of unknown regression coefficients associated with the covariates in $\boldsymbol{x}_{gj}$ and η_{gj} is a random group effect that accounts for unobservable heterogeneity among individuals within a group.[3] By rearranging (4.5), we get an expression for the probability that an individual in group g selects choice j:

$$p_{gj} = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{gj}+\eta_{gj})}{\sum_{j=1}^{J}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{gj}+\eta_{gj})} = \frac{\lambda_{gj}\exp(\eta_{gj})}{\sum_{j=1}^{J}\lambda_{gj}\exp(\eta_{gj})}. \tag{4.6}$$

We further make the assumption that the $\exp(\eta_{gj})$'s follow independent gamma distributions with both shape and scale parameters $\delta_g\lambda_{gj}$, $\delta_g > 0$. Under these assumptions, it can

[1] Note that in our implementation of DM regression, $\alpha_g = \delta_g^{-1}\lambda_g$, for some $\delta_g > 0$, hence the last term in (4.4) can be rewritten as $\frac{n_g\delta_g+\lambda_g}{\delta_g+\lambda_g}$.

[2] Under the RUM framework with utility function given by the RHS of (4.5) plus independent errors that follow an extreme value distribution, the logit formulation follows directly. See Appendix G.

be shown that the probabilities $\boldsymbol{p_g}$ follow a Dirichlet distribution with parameters $\boldsymbol{\alpha_g} = (\delta_g^{-1}\lambda_{g1}, \ldots, \delta_g^{-1}\lambda_{gJ})$, where δ_g is used to quantify the overdispersion. Hence the log-likelihood associated with the DM distribution in (4.2) becomes:

$$\begin{aligned} l_{DMd} &= \sum_{g=1}^{G} \Big\{ \log(n_g!) + \log\Gamma(\delta_g^{-1}\lambda_g) - \log\Gamma(\delta_g^{-1}\lambda_g + n_g) \\ &\quad + \sum_{j=1}^{J} \log\Gamma(\delta_g^{-1}\lambda_{gj} + y_{gj}) - \log\Gamma(\delta_g^{-1}\lambda_{gj}) - \log(y_{gj}!) \Big\}. \end{aligned} \tag{4.7}$$

where, $\lambda_g = \sum_{j=1}^{J} \lambda_{gj}$. Finally, the DM model can be re-parameterized in terms of the intragroup correlation coefficient, ρ_g, since $\rho_g = \frac{\delta_g}{\lambda_g + \delta_g}$, where $0 < \rho_g < 1$. This alternate parameterization results in the following DM log-likelihood function:

$$\begin{aligned} l_{DMr} &= \sum_{g=1}^{G} \Big\{ \log(n_g!) + \log\Gamma(\rho_g^{-1} - 1) - \log\Gamma([\rho_g^{-1} - 1] + n_g) \\ &\quad + \sum_{j=1}^{J} \log\Gamma([\rho_g^{-1} - 1]p_{gj} + y_{gj}) - \log\Gamma([\rho_g^{-1} - 1]p_{gj}) - \log(y_{gj}!) \Big\}. \end{aligned} \tag{4.8}$$

Maximization of (4.7) or (4.8) provides estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}_g$ (or $\boldsymbol{\rho}_g$). In the special case that $\eta_{gj} = 0$, there is no random group effect and the DM model reduces to a group conditional logit model. If, further, there are only group-specific covariates, then the model reduces to a standard multinomial logit model. Under an alternative parameterization of the DM model, the cell frequencies y_{gj} can be shown to follow a negative multinomial distribution with fixed effects (Guimarães and Lindrooth, 2007). Conditional on the row sums, the cell frequencies form overdispersed count variables following a Poisson distribution with rate $\lambda_{gj} = \exp(\boldsymbol{\beta' x_{gj}})$, where λ_{gj} itself is a random Gamma variate with parameters $(\delta_g^{-1}\lambda_{gj}, \delta_g^{-1})$.

Modelling Overdispersion

The overdispersion associated with group g can be represented in several ways. The simplest case is constant overdispersion across all groups: $\delta_g = \delta$. However, sometimes it may be reasonable to assume that overdispersion is a function of group-specific traits: $\delta_g = \exp(\gamma_0 + \gamma_1 z_g)$, where $\boldsymbol{z}_g$ is an L-length vector of group specific characteristics and $\gamma = (\gamma_0, \gamma_1)$ is a

[3] In general, the addition of η_{gj} may induce correlation among the choices. Refer to Appendix H for more details and the relation of this model to the mixed logit model.

coefficient vector to be estimated. Similarly, when overdispersion is modelled as a function of the intragroup correlation, ρ_g can be modelled as a constant, $\rho_g = \rho$, or as a function of group level covariates, $\text{logit}(\rho_g) = \gamma_0 + \gamma_1 z_g$.

4.2.3 Optimization Methods

Methods for unconstrained nonlinear optimization problems can be broadly classified based on the amount of derivative information used: (i) higher-order methods that require both first and second derivatives, e.g., Newton-Raphson; (ii) gradient methods that require only first derivatives, e.g., BFGS; and (iii) direct search methods that require no derivatives, e.g., Nelder-Mead. The former two approaches are based on line searches, while the latter is based on a pattern search. We implemented DM regression for grouped data in `R` (Crea, 2014), using the `optimx` function to perform all three optimization methods [4].

The Newton-Raphson algorithm (Ypma, 1995) is a commonly used method for maximizing the likelihood (or minimizing the negative log-likelihood) function to obtain $\hat{\boldsymbol{\theta}}$, the vector of maximum likelihood estimates of the unknown model parameters $\boldsymbol{\theta}$. The algorithm essentially iterates between calculating a direction vector $\boldsymbol{d}_i$ that is based on the score vector and Hessian matrix $H(\boldsymbol{\theta})$, and updating parameter estimates by moving them in that direction until the likelihood converges. Unfortunately, derivatives are not always analytically available, the Newton-Raphson algorithm is computationally expensive and it has poor global convergence properties. In practice, a modified Newton-Raphson (MNR) in which the direction vector is iteratively scaled by a step size α_i is used to speed up convergence. Here, we consider a MNR with $\alpha_i = \underset{\alpha_i > 0}{\text{argmin}}\, f(\boldsymbol{\theta}_i + \alpha_i \boldsymbol{d}_i)$.

The BFGS (suggested independently by Broyden, Fletcher, Goldfarb, and Shanno in 1970) approximates the Newton-Raphson by iteratively calculating $B(\boldsymbol{\theta})$ to approximate $H(\boldsymbol{\theta})$ based on changes in the gradient vector. Unlike the MNR method, the BFGS does not directly evaluate the Hessian, and explicit conditions are enforced to ensure $B(\boldsymbol{\theta})$ is a symmetric positive definite matrix. Local convergence of the BFGS has been established (Dennis and Moré, 1974; Dennis and Moré, 1977); however, the BFGS algorithm converges only superlinearly, compared to the quadratic convergence of the MNR method. On the other hand,

[4] In optimx, we provided the DM likelihood function and the analytical gradient function (Appendix I), but second derivatives for the MNR method were calculated numerically using a finite-difference approximation. Most default parameters were used, except for the following: maxit = 10,000, kkt=F and starttests=F.

unlike MNR, BFGS is globally convergent for convex problems when an exact or some special inexact line search is used (Byrd and Nocedal, 1989; Byrd et al., 1987; Dixon, 1972; Griewank, 1991; Powell, 1971; Powell, 1976; Toint, 1986) and for non-convex problems when the function has Lipschitz continuous gradients (Li and Fukushima, 2001).

The NM method (Nelder and Mead, 1965) is a commonly used heuristic search procedure for finding the minimum (or maximum) of an objective function in n-dimensional space. Unlike the aformentioned line search methods, this direct search approach does not require any derivative information. In brief, the method takes in $n+1$ test points, defines a simplex as the convex hull of the $n+1$ points and replaces the point associated with the highest objective function value, among all test points, by a new point that has a lower function value. While NM can handle complicated functions for which first and second derivatives are tedious to derive analytically, or are analytically intractable, it may take an unnecessarily large number of function evaluations before reaching convergence (Nash, 1990). In practice, Newton or quasi-Newton methods should be used over NM when higher order derivative information is available. However, direct search methods are still an effective option and sometimes the only option for a variety of difficult optimization problems, e.g., when a function is non-smooth or for noisy data (Kolda et al., 2003).

4.3 Simulation Study

4.3.1 Simulation Design

The simulation study was motivated by the problem of modelling (counts of) interactions made between plant and pollinator species in a given ecosystem over a given sampling period. Several studies suggest that these interactions are driven, or facilitated, by linkage rules defined by plant and pollinator species traits (Santamaría and Rodríguez-Gironés, 2007; Allesina et al., 2008; Stang et al., 2009; Vázquez et al., 2009b; Blüthgen, 2010; Rohr et al., 2010; Olesen et al., 2011; Bartomeus, 2013; Junker et al., 2013; Eklöf et al., 2013; Gravel et al., 2013). Networks were generated in `R`, per Crea et al. (2016), based on the gamma-Poisson parameterization of the DM model. Each network was based on a $G \times J \times 4$ design array X, where x_k is a covariate-specific $G \times J$ design matrix with entries x_{jgk}, $k = 1, \ldots, 4$. Binary covariates x_1, x_2, and x_3 represented linkage rules. Continuous covariates x_4 and z_1

were sampled from the relative species abundance distribution Ravasz et al. (2005) for plant and pollinator species, respectively, using the inverse cdf method (Devroye, 1986). We considered five dispersion structures: none, constant (dconst: $\delta_g = \delta$), δ as a function of group covariates (dfunc: $\delta_g = \exp(\gamma_0 + \gamma_1 z_g)$), constant intragroup correlation (rconst: $\rho_g = \rho$), and ρ as a function of group covariates (rfunc: $\text{logit}(\rho_g) = \gamma_0 + \gamma_1 z_g$).

To evaluate model fit, we studied two network sizes: 53×20 and 78×40. Parameter values were independently sampled from uniform distributions over equal-sized and non-overlapping low, medium, and high intervals that covered the following respective parameter ranges: $\beta_k \in (-1, 5)$, $\delta^{-1} \in (0, 12)$, $\rho \in (0, 1)$, and $\gamma_\ell \in (-4, 2)$. These parameter settings corresponded to networks that were neither too sparse nor too highly populated. Networks were generated assuming all combinations of parameter values (4 coefficients and 1 or 2 dispersion parameters) according to a factorial design with 10 replicates. Accordingly, for each network size, there were $3^4 \times 10 \times 2 = 810$ networks with no dispersion, $3^5 \times 10 = 2,430$ networks with constant dispersion, and $3^6 \times 10 = 7,290$ networks with dispersion as a function of covariates.

Slices of the negative log-likelihood were examined to study the behaviour of the DM model under the various parameterizations. For each slice, all elements of the $\boldsymbol{\beta}$ vector were equal and fixed at either 0.5, 2, or 4; e.g., $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)$. The value of the dispersion parameter was set at a low, medium, or high value ($\delta^{-1} \in (2, 6, 10)$, $\rho \in (0.15, 0.50, 0.85)$, and $\gamma_1 \in (-3, -1, 1)$[5]). Further, all networks were sized 53×20.

4.3.2 Simulation Study Results

DM Model Fits

Table 4.1 shows the results for correctly specified models, i.e., when true dispersion matched the modelled dispersion. Since similar trends were observed across network sizes, only results associated with the 53×20 networks are discussed (see Appendix J for the 78×40 network results). Further, since the trends tend to persist regardless of the magnitude of the coefficients, the following results are collapsed across all values of $\boldsymbol{\beta}$.

In general, all three optimization methods were competitive and produced similar results for

[5]For dispersion as a function of group covariates, dfunc: $\delta_g = \exp(\gamma_1 z_g)$, and rfunc: $\text{logit}(\rho_g) = \gamma_1 z_g$.

the no dispersion parameterization. Although results were almost identical across optimization methods for the rconst parameterization, the percent relative bias tended to be a bit lower with the BFGS method. For the dconst, dfunc and rfunc parameterizations, the NM method performed best in that estimates converged, with invertible Hessians, more often compared to the BFGS and MNR methods, respectively. We also found that the highest family-wise coverage probability of 85% was associated with NM, based on Wald type 95% confidence intervals. In general, goodness-of-fit (GOF) statistics were similar across optimization methods except for the rfunc parameterization, where BFGS showed considerably poorer GOF and corresponding high bias. Across all methods and parameterizations, higher percent relative bias was observed for $\hat{\beta}_4$ (the continuous covariate).

Tables 4.2 and 4.3 show the performance of the optimization methods when the dispersion structure is misspecified. In general, NM produced model fits that converged to estimates with invertible Hessians almost 100% of the time, except when true dispersion was none, where convergence was less than 75%. NM was also associated with better GOF for the δ parameterizations (i.e. high percentage of models having χ^2 p-values greater than 0.05) and competitive GOF for the rconst parameterization.

Similar to the correctly specified model fits, the median and IQR percent relative bias of β was low across all methods and parameterizations, but notably higher for parameterizations in terms of ρ. Further, percent relative bias was typically higher for $\hat{\beta}_4$. The individual and family-wise coverages of β were lower when true dispersion was rconst or rfunc compared to when true dispersion was dconst or dfunc for all methods. However, it was the BFGS (not NM) that showed the highest family-wise coverages when modelled dispersion was rconst. Upon closer inspection, it appears that higher bias and lower coverage of β was associated with higher levels of dispersion in the rconst and rfunc settings. These trends did not persist for the dconst and dfunc parameterizations.

Table 4.4 succinctly summarizes the results of the simulation study based on the performance metrics discussed above. In cases where the likelihood was convex, i.e., modelled dispersion was none (standard discrete choice logit model) or rconst, then higher order methods performed just as well or better than NM with respect to the performance metrics. For dconst, dfunc, and rfunc, NM generally performed best, but MNR performed better than NM with respect to the execution time. Overall, all three methods reached convergence relatively

quickly with times ranging from 0.01 seconds to 19.61 seconds (median time of 0.31 seconds). However, BFGS had the lowest median time of 0.28 seconds, while NM had the highest median time of 0.55 seconds.

Behaviour of the Negative Log-likelihood

Figure 4.1 shows slices of the negative log-likelihood for each of the four dispersion structures, evaluated at the correct (true) value of the coefficients. In general, the negative log-likelihood curves appear to become less convex as dispersion increases, regardless of the value of $\boldsymbol{\beta}$. For the dconst and dfunc parameterizations, there is a distinct minimum when data are generated with low dispersion values (dashed curves on first two rows of plots, where $\delta = 2$ and $\gamma_1 = -3$, respectively). However, there is somewhat of a flattening of the negative log-likelihood when the dispersion parameter is set to higher values. The same trends are observed for the rconst and rfunc parameterizations (last two rows of plots). Note that the y-axis range is very large so the curves are not as flat as they appear. Regardless, this flattening suggests that the estimates of the dispersion parameters become more imprecise as dispersion increases. These same trends persist even if the y-axis range was adjusted to show a few negative log-likelihood units above the minimum.

Table 4.1: Results of DM model fits when true dispersion matches modelled dispersion for network size 53×20.

True		Converge	Good Fit	Percent Relative Bias of $\hat{\beta}$		Coverage of 95% Wald CI's	
Disp.[1]	Method	(%)[2]	(%)[3]	Median	IQR[4]	Individual	Family
none	BFGS	100	100	(0.64, 0.40, 0.45, 1.50)	(2.55, 1.68, 1.61, 4.23)	(93, 95, 94, 95)	80
	NM	100	100	(0.64, 0.40, 0.46, 1.48)	(2.48, 1.66, 1.62, 4.19)	(93, 96, 94, 95)	81
	MNR	100	100	(0.65, 0.40, 0.46, 1.51)	(2.55, 1.68, 1.60, 4.18)	(93, 96, 94, 95)	81
dconst	BFGS	83	84	(1.50, 1.20, 1.18, 3.36)	(5.02, 4.43, 4.26, 11.27)	(95, 96, 95, 95)	68
	NM	100	83	(1.43, 1.04, 1.11, 3.29)	(4.53, 3.83, 3.84, 10.68)	(95, 96, 95, 95)	82
	MNR	98	83	(1.44, 1.06, 1.12, 3.28)	(4.49, 3.85, 3.9, 10.66)	(95, 96, 95, 95)	80
rconst	BFGS	98	75	(9.13, 7.24, 7.40, 24.62)	(21.47, 18.21, 15.69, 72.22)	(95, 95, 95, 95)	80
	NM	98	75	(9.18, 7.37, 7.39, 24.84)	(21.76, 18.73, 15.64, 73.66)	(95, 95, 95, 95)	80
	MNR	98	75	(9.17, 7.26, 7.41, 24.80)	(21.61, 18.37, 15.77, 73.06)	(95, 95, 95, 95)	81
dfunc	BFGS	62	86	(1.19, 0.80, 0.86, 2.90)	(4.22, 3.68, 3.08, 9.56)	(94, 95, 94, 95)	50
	NM	97	88	(1.17, 0.77, 0.82, 2.72)	(3.76, 3.24, 2.66, 8.54)	(94, 95, 95, 95)	78
	MNR	91	87	(1.12, 0.73, 0.80, 2.62)	(3.41, 2.90, 2.49, 8.31)	(94, 95, 95, 94)	73
rfunc	BFGS	60	44	(20.56, 17.21, 20.42, 51.25)	(38.11, 32.20, 38.55, 149.99)	(62, 57, 69, 70)	14
	NM	99.5	74	(4.90, 4.09, 5.79, 17.34)	(11.91, 10.54, 14.24, 49.91)	(95, 94, 95, 96)	85
	MNR	67	72	(5.88, 5.11, 6.83, 20.95)	(13.57, 12.74, 16.48, 57.44)	(93, 92, 93, 94)	54

[1] Based on total of 810 (none), 2430 (dconst/rconst), and 7290 (dfunc/rfunc) networks, respectively.
[2] Percentage of total networks that converged with invertible Hessian matrices.
[3] Percentage of total networks that converged with invertible Hessian matrices that had χ^2 p-values greater than 0.05.
[4] IQR - Interquartile range.

Table 4.2: Results of misspecified DM model fits parameterized in terms of δ (Network Size 53×20)

True		Converge	Good Fit	Percent Relative Bias of β		Coverage of 95% Wald CI's	
Disp.[1]	Method	(%)[2]	(%)[3]	Median	IQR[4]	Individual	Family
Modelled Dispersion: dconst							
none	BFGS	24	97	(0.22, 0.17, 0.18, 0.83)	(0.59, 0.46, 0.34, 2.49)	(89, 93, 93, 92)	19
	NM	71	100	(0.75, 0.45, 0.53, 1.56)	(2.92, 1.84, 2.04, 4.24)	(93, 97, 94, 95)	63
	MNR	64	100	(0.52, 0.30, 0.38, 1.32)	(1.44, 0.98, 1.18, 3.45)	(93, 96, 95, 94)	52
rconst	BFGS	33	87	(31.92, 15.39, 15.52, 28.88)	(41.13, 27.60, 24.24, 99.44)	(59, 73, 77, 91)	13
	NM	100	92	(32.34, 17.82, 16.55, 29.32)	(25.96, 19.78, 19.74, 83.10)	(43, 61, 65, 89)	26
	MNR	66	92	(31.86, 17.80, 16.06, 26.61)	(24.56, 19.44, 18.39, 73.47)	(39, 57, 61, 88)	15
dfunc	BFGS	62	86	(1.21, 0.82, 0.85, 2.82)	(4.43, 3.81, 3.09, 9.15)	(94, 95, 94, 95)	50
	NM	100	88	(1.18, 0.78, 0.83, 2.69)	(3.79, 3.28, 2.68, 8.49)	(94, 95, 95, 95)	80
	MNR	93	87	(1.11, 0.73, 0.78, 2.58)	(3.41, 2.94, 2.45, 8.10)	(94, 95, 95, 94)	75
rfunc	BFGS	35	82	(16.73, 14.29, 29.91, 24.51)	(25.48, 21.88, 36.71, 63.97)	(64, 61, 50, 89)	9
	NM	100	90	(18.29, 16.63, 31.41, 26.90)	(20.99, 21.22, 28.54, 64.37)	(49, 49, 35, 84)	16
	MNR	74	89	(17.68, 15.71, 30.23, 24.20)	(20.59, 21.01, 29.09, 60.66)	(45, 48, 33, 84)	10
Modelled Dispersion: dfunc							
rfunc	BFGS	24	84	(13.66, 10.42, 24.11, 24.49)	(28.54, 20.30, 46.10, 76.90)	(79, 75, 61, 88)	10
	NM	99.8	91	(12.35, 10.98, 24.35, 23.06)	(18.14, 16.88, 27.12, 66.73)	(62, 63, 42, 85)	24
	MNR	74	92	(11.80, 10.04, 23.21, 20.09)	(17.44, 16.51, 27.13, 54.71)	(59, 62, 41, 88)	17

[1] Total Networks: none=810; dconst/rconst=2430; and dfunc/rfunc=7290.
[2] Percentage of total networks that converged with invertible Hessian matrices.
[3] Percentage of total networks that converged with invertible Hessian matrices that had χ^2 p-values greater than 0.05.
[4] IQR - Interquartile range.

Table 4.3: Results of misspecified DM model fits parameterized in terms of ρ (Network Size 53×20)

True		Converge	Good Fit	Percent Relative Bias of β		Coverage of 95% Wald CI's	
Disp.[1]	Method	(%)[2]	(%)[3]	Median	IQR[4]	Individual	Family
Modelled Dispersion: rconst							
none	BFGS	73	99	(0.62, 0.41, 0.51, 1.43)	(2.51, 1.45, 1.87, 4.21)	(94, 96, 95, 94)	59
	NM	72	87	(2.35, 0.75, 1.33, 4.49)	(5.72, 2.61, 3.32, 13.08)	(60, 77, 69, 57)	28
	MNR	37	99	(1.06, 0.54, 0.90, 2.02)	(6.40, 2.76, 3.39, 8.16)	(90, 93, 93, 92)	46
dconst	BFGS	100	22	(1.97, 1.56, 1.61, 4.12)	(6.08, 4.46, 4.76, 11.52)	(86, 90, 88, 93)	65
	NM	100	22	(2.36, 1.65, 1.78, 5.24)	(6.26, 4.65, 4.93, 14.19)	(81, 87, 85, 84)	57
	MNR	95	23	(2.19, 1.78, 1.81, 4.39)	(6.34, 4.72, 5.11, 12.14)	(86, 89, 88, 93)	65
dfunc	BFGS	99	40	(1.53, 1.20, 1.12, 3.31)	(5.15, 4.78, 3.77, 10.20)	(88, 87, 88, 93)	67
	NM	98	39	(2.21, 1.44, 1.44, 4.99)	(5.72, 5.09, 3.98, 13.85)	(78, 81, 81, 79)	52
	MNR	80	40	(2.38, 1.89, 1.77, 4.57)	(6.78, 6.19, 4.77, 12.81)	(87, 85, 87, 93)	64
rfunc	BFGS	100	74	(8.41, 7.22, 8.88, 21.36)	(15.80, 14.36, 17.82, 55.65)	(78, 80, 83, 91)	58
	NM	100	74	(8.40, 7.24, 8.85, 21.38)	(15.84, 14.39, 17.90, 55.77)	(78, 80, 83, 91)	58
	MNR	100	74	(8.41, 7.22, 8.88, 21.36)	(15.80, 14.35, 17.84, 55.65)	(78, 80, 83, 91)	58
Modelled Dispersion: rfunc							
dfunc	BFGS	20	77	(48.51, 47.06, 46.37, 40.41)	(33.06, 48.12, 32.70, 56.89)	(28, 31, 24, 54)	1
	NM	99	39	(1.51, 1.16, 1.10, 3.28)	(4.98, 4.62, 3.69, 10.15)	(99, 98, 99, 99)	95
	MNR	20	41	(6.54, 7.38, 4.99, 9.85)	(14.49, 25.94, 12.48, 24.16)	(97, 95, 97, 95)	18

[1] Total Networks: none=810; dconst/rconst=2430; and dfunc/rfunc=7290.
[2] Percentage of total networks that converged with invertible Hessian matrices.
[3] Percentage of total networks that converged with invertible Hessian matrices that had χ^2 p-values greater than 0.05.
[4] IQR - Interquartile range.

Table 4.4: Overall Simulation Study Results

	Modelled Dispersion				
True Dispersion	**none**	**dconst**	**dfunc**	**rconst**	**rfunc**
none	equivalent	NM[a,c]/MNR[b]	–	BFGS[a,b,c,d]	–
dconst	equivalent	NM[a]/MNR[d]	–	BFGS[a,b,c,d]	–
rconst	equivalent	NM[a,b,c,d]	–	BFGS[b,d]	–
dfunc	equivalent	NM[a,b,c,d]	NM[a]/MNR[d]	BFGS[a,b,c,d]	NM[a,b,c]
rfunc	equivalent	NM[a,b,c,d]	NM[a,d]	BFGS[d]	NM[a,b,c]

BFGS - Broyden-Fletcher-Goldfarb-Shanno; NM - Nelder-Mead; MNR - modified Newton-Raphson

[a] method converged at a point where Hessian was negative definite

[b] median or IQR of percent relative bias of regression coefficients

[c] individual or family-wise coverage of regression coefficients

[d] IQR - reported execution time for the method

4.4 Empirical Networks

We used DM regression to analyze 12 networks from the Natural Park of Cap de Creus (Catalonia, northeastern Spain), along the coastal Mediterranean shrublands (Bartomeus et al., 2008). The networks were surveyed in the spring and summer of 2005 at six sites: three sites for each of two invasive plant species of interest (*Carpobrotus*, and *Opuntia*). Each site was further divided into paired 50m $\times$ 50m plots, one invaded by the plant species of interest and the other uninvaded. Half of the plots were sampled six times over March-April and the remaining half were sampled four times over June-July. An interaction was recorded if an insect touched the reproductive organs of the flower.

Only plant traits gathered from the literature, following Bosch et al. (1997) classification, were considered for covariates in the analysis. Plant traits included: (1) corolla colour: yellow, white, purple, green or blue; (2) floral symmetry: actinomorphic and zygomorphic; (3) corolla shape: disc-bowl shaped, restrictive tubular, or very restrictive papilionaceous; and (4) flower type: solitary, in raceme, or composite flowers. Mean plant flower abundance for each species at plot level was used as an additional covariate in the variable selection process. Data and traits are available at `http://figshare.com/articles/Plant_Pollinator_Network_Data/154863`. The final model was chosen through a stepwise selection method based on Pearson χ^2 GOF statistics and the Akaike Information Criterion (AIC).

dconst:
$\boldsymbol{\delta_g = \delta}$

dfunc:
$\boldsymbol{\delta_g} =$
$\exp(\boldsymbol{\gamma_1 z_g})$

rconst:
$\boldsymbol{\rho_g = \rho}$

rfunc:
$\text{logit}(\boldsymbol{\rho_g}) =$
$\boldsymbol{\gamma_1 z_g}$

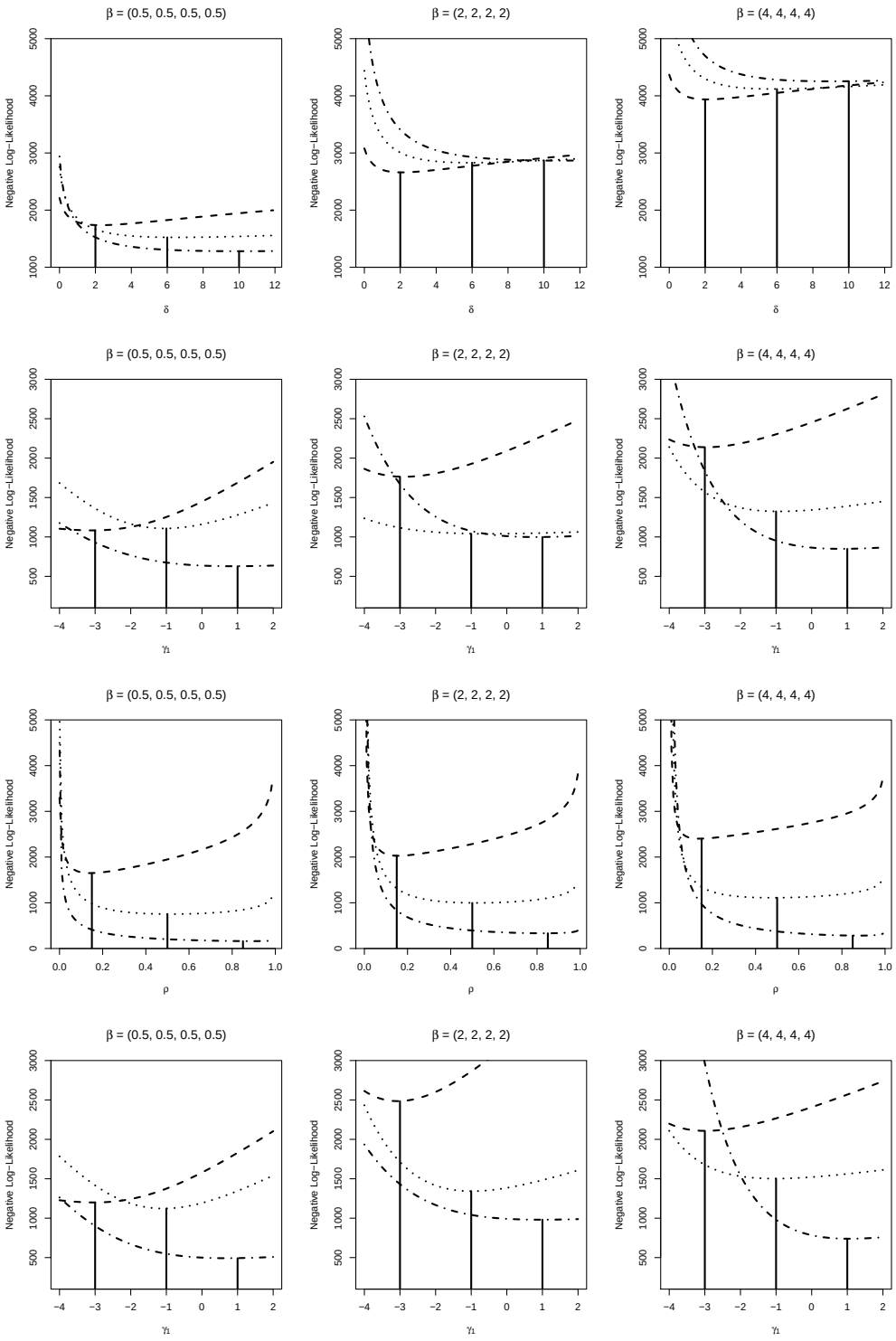

Figure 4.1: Dashed line ≡ low dispersion; dotted line ≡ medium dispersion; dot-dashed line ≡ high dispersion. Solid line marks the true value of the dispersion parameter for each profile.

4.4.1 Empirical Network Results

Network size ranged from 21 to 47 pollinator species by 8 to 14 plant species. On average, networks had 281 potential links, and 102 observed interactions. We found that floral symmetry (zygomorphic vs. actinomorphic) was not statistically significant for any of the models fits. Additionally, covariate estimates tended to be more statistically significant for the no dispersion model compared to the DM models. However, the DM models provided a better fit to the data than the no dispersion models and the dispersion parameter was consistently significant, suggesting that these data do exhibit overdispersion. Among the DM models, the dconst model did not converge for several fits that included the continuous covariate mean flower abundance, unlike the analogous rconst model. When estimates were available, the dconst and rconst models produced similar results, but the AIC was lowest for the rconst models. Looking across optimization methods for the rconst model, estimates were almost identical. When coupled with the results of the simulation study, these trends suggest that the underlying dispersion structure is likely to be rconst.

Table 4.5 presents parameter estimates under the BFGS method for the rconst models since this method performed best when data were modelled by the rconst dispersion structure in our simulation study. On average, two covariates were sufficient in explaining the distribution of counts across plant and pollinator species and most models included mean flower abundance, corolla shape, and type of flower as significant predictors. In particular, mean floral abundance was found to be significant in 9 out of the 12 models and type of flower was significant for 8 of the models. Corolla shape was statistically significant for only 2 invaded plots when invasive species was *Carpobrotus* (sites 1-2), but appeared in models for both invaded and uninvaded when invasive species was *Opuntia* (sites 4-6). It is interesting to note that when using the χ^2 p-value to assess fit, among sites 1-3 (associated with invasive species *Carpobrotus*), the uninvaded plots consistently had better model fits, whereas among sites 4-6 (associated with invasive species *Opuntia*), it was the invaded plots that consistently had better model fits. However, when using the AIC value to assess fit, the uninvaded plots consistently had lower AICs (and hence better model fits) for all sites 1-6.

4.5 Discussion

In this paper we have studied the behaviour of the negative log-likelihood associated with different parameterizations of grouped DM regression, evaluated the performance of three optimization methods for these models, and provided a stable implementation to obtain associated MLEs. We found that the pattern search method NM performed best for the δ (dconst and dfunc) and rfunc (ρ as a function of covariates) parameterizations, while the line search method BFGS performed best for the rconst (ρ constant) parameterization. Although the MNR was competitive for some parameterizations, it never out-performed either of the other two methods.

When the negative log-likelihood function is globally convex, as for discrete choice logit models with linear-in-parameters utility, then there is only one minimum; however, the DM model (along with many other discrete choice models) is not globally convex. For globally convex functions, like the no dispersion parameterization, the local minimum is the global minimum, and theoretically, BFGS and MNR do work better and have a superior rate of convergence, as compared to NM. This is in line with the findings of the simulation study for the no dispersion case, but in terms of bias and coverage, all optimizations methods performed equally. Interestingly, the BFGS did perform better than the NM for the rconst parameterization. We used a starting value of $\rho = 0.1$, which seemed to be an adequate starting point, and may explain why BFGS was the best option for the rconst parameterization, even though the DM likelihood is, in general, non-convex.

For dconst, dfunc and rfunc parameterizations, the NM converged at a point where the Hessian was negative definite more often than BFGS and MNR. Generally, these latter two methods diverged and/or did not settle at an optimal point. Possible reasons for these results include: (i) multiple local minima, (ii) a saddlepoint, (iii) poor starting values for the dispersion parameters, and (iv) gradients that are not well-behaved or well-defined at the minimum (e.g., non-smooth regions in the likelihood). Regardless, NM seems to be more robust to (i) to (iv) than the other two methods for these parameterizations of the DM regression model. Future work in deriving better starting values for the dispersion parameters could circumvent some of the poor global convergence issues associated with the BFGS and MNR methods.

The log-likelihood plots showed that the DM likelihood flattened with increasing dispersion. This is a simple way of taking a slice of the likelihood for one parameter of interest (the dispersion parameter), but it can be statistically misleading because the other parameters are fixed (Bolker, 2008). Consequently, this slice does not show the behaviour of the negative log-likelihood near the minimum when the other parameters (β) are varied, which would require a profile of the log-likelihood. Further, one cannot access convexity of the likelihood surface at the minimum. An examination of the eigenvalues of the Hessian evaluated near the minimum is needed to assess convexity. If the Hessian contains both negative and positive values, then the optimization algorithm settled at a saddle point, which could explain why BFGS and MNR performed poorly compared to NM for some of the DM parameterizations.

The percentage of GOF χ^2 p-values > 0.05 rarely reached the nominal value of 95% for misspecified models across all parameterizations. It may be that the quality of the asymptotic approximation of the χ^2-statistic to follow a χ^2 distribution is poor for the DM model and which can lead to inaccurate p-values and inflated Type I error rates. A viable alternative is to use measures such as the AIC or BIC for model selection.

It is also worth noting that although the percent relative bias of β was typically low for all covariates, the bias of β_4 was consistently highest. Recall that β_4 was associated with relative species abundance, a continuous covariate. Further, by the way the covariate was defined, it is likely that there were not many species that shared the same relative abundance. As such, it might be expected that there was not enough information at the group level to estimate this parameter. In practice, this limitation may not pose a problem if covariates are binary or represent linkage rules.

Bartomeus (2013) analyzed the same empirical networks as we did but using a separate Poisson regression for each pollinator species. Similar to our analyses, on average two covariates were retained in his final models, with flower abundance and type of flower appearing in most models. Flower abundance may be a marker for generalist (or social) species. Generalists tend to be the most abundant species and vice-versa (Chittka and Thomson, 2005), and often more interactions are observed among abundant species (Vázquez and Aizen, 2004). However, the corresponding parameter estimates were not very large, ranging from 1.002 to 1.07, it may not be a significant ecological predictor of plant-pollinator interaction.

Table 4.5: Odds ratios for the 12 Mediterranean scrubland plant-pollinator networks under rconst model with BFGS optimization. Sites 1-3 are associated with invasive plant species *Carpobrotus*, while sites 4-6 are associated with *Opuntia*.

Site	**1**		**2**		**3**		**4**		**5**		**6**	
Invaded Plot	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Network Size	26×9	47×10	25×10	43×14	30×10	27×11	24×7	25×8	27×8	24×8	21×9	28×9
Total Counts	70	185	88	177	93	114	63	97	77	112	74	77
Variable:[1]												
Colour												
White	–	–	1.00	1.00	–	–	–	–	1.00	–	–	1.00
Coloured	–	–	0.41^{*}	0.59	–	–	–	–	0.32^{**}	–	–	0.13^{***}
Type of Flower[2]												
Solitary	1.00	1.00	1.00	–	1.00	1.00	–	1.00	–	–	1.00	1.00
Raceme	0.55	0.09^{***}	0.26^{*}	–	0.14^{***}	0.17^{***}	–	0.17^{***}	–	–	–	–
Comp.	4.42^{*}	3.69^{*}	0.83	–	0.36^{**}	0.31^{***}	–	0.94	–	–	–	–
Raceme/Comp.	–	–	–	–	–	–	–	–	–	–	7.43^{*}	0.28^{*}
Corolla Shape[3]												
Disc-bowl	–	1.00	–	1.00	–	–	1.00	1.00	–	1.00	1.00	–
Tubular	–	0.11^{***}	–	0.42^{**}	–	–	–	–	–	–	0.18^{**}	–
Pap.	–	0.03^{***}	–	0.10^{**}	–	–	–	–	–	–	0.27	–
Tubular/Pap.	–	–	–	–	–	–	0.32^{*}	0.17^{**}	–	0.24^{**}	–	–
Mean Abund.	1.003^{**}	1.002^{*}	1.02^{***}	1.01^{**}	–	–	1.03^{***}	1.03^{**}	1.06^{***}	1.06^{8*}	–	1.07^{***}
Dispersion												
Constant	0.30^{**}	0.28^{***}	0.34^{***}	0.25^{***}	0.37^{***}	0.15^{**}	0.45^{***}	0.56^{***}	0.31^{***}	0.55^{***}	0.38^{***}	0.35^{***}
AIC	180.05	401.24	206.16	459.60	240.97	296.97	133.11	160.69	175.06	181.94	153.82	194.23
χ^2 p-value	>0.999	0.966	0.927	0.141	0.745	0.087	0.096	0.572	0.241	0.944	0.386	0.991

[1] Comp.=composite flower; Pap.= papilionaceous; and Abund.=abundance.

[2] Raceme and Comp. were collapsed into one category when there was an insufficient number of plant species to estimate the effects separately.

[3] Tubular and Pap. were collapsed into one category when there was an insufficient number of plant species to estimate the effects separately.

* $p < 0.05$; ** $p < 0.01$; and *** $p < 0.001$.

It is known that abundance will vary throughout a season, depending on the phenology of the floral species. As such, the mean abundance of the plot over the study period may not be the best representation of abundance since it ignores the abundance of other floral species at any given time. Perhaps average relative abundance would give a better summary of this phenomenon. Contrary to our findings, floral symmetry appeared in some of his final models. This discrepancy could be attributed to the fact that he did not account for overdispersion in his analysis.

4.6 Relation to Other Work and Conclusions

The most current literature on the implementation of the DM regression model focuses on ungrouped applications in which counts are at the individual level and the log odds are modelled as a function of the individuals' characteristics. These models parallel the structure of the standard multinomial logistic model, where one of the J response categories is set as the baseline and the log odds for all other categories are calculated relative to the baseline. Effectively, the model estimates $J-1$ regression coefficients for each covariate, and each β_j can be interpreted as the effect of the covariates on the odds of making a given choice over another. Thus, this model aims to explain how an individual's characteristics affect the likelihood of falling in a particular response category, as often arises in bioinformatics where interest lies in the association between microbiome composition and environmental factors. For example, Chen and Li (2013) use the DM regression model to link nutrient intake to the human gut microbiome.

In topic modelling, DM regression is used within a mixture model (Mimno and McCallum, 2012). There, the DM model is derived from a probability framework in which the multivariate counts are assumed to follow a mixture of multinomial distributions with a Dirichlet prior over the multinomial parameters and the Dirichlet prior is modelled as a function of document features. Zhang et al. (2017) give a comprehensive evaluation of DM regression models arising under varying correlation structures, such as the generalized DM, but motivated from high-throughput data in genomics. However, it should be noted that our implementation of DM regression is for grouped data and does not represent a mixture model.

There exist several packages in `R` that provide fitting of multinomial logistic models and their DM extensions; of note are `nnet` (Venables and Ripley, 2002), `MGLM` (Zhang et al.,

2017), `MCMCglmm` (Hadfield, 2010), and `VGAM` (Yee, 2010). Similarly, numerous packages are available to fit discrete choice models derived from random utility theory, such as the conditional and mixed logit models; of note being the `mlogit` (Croissant, 2013), `mclogit` (Elff, 2014), and `mnlogit` (Hasan et al., 2015) packages. However, none of these routines can accommodate the grouped DM model framework, which explicitly incorporates overdispersion.

Currently, there are two implementations of DM regression for grouped data: (i) the `dirmul` function (Guimarães and Lindrooth, 2007) in Stata (StataCorp, 2011), which can incorporate both group-level and choice-specific covariates; and (ii) our `dirmultreg` function in `R` (R Core Team, 2015), available on GitHub (Crea, 2014). The former uses a modified NR method in which numerical derivatives via a finite-difference approximation are used for first derivatives in the δ parameterizations and for both first and second derivatives in the ρ parameterizations[6]. However, our simulations suggest that an MNR algorithm can be problematic for complex parameterizations, particularly when dispersion is modelled as a function of covariates. The latter implementation uses the `optimx` package to do the optimization for all three methods[7], thereby providing a more stable implementation of grouped DM regression.

In practice, one typically does not know the true underlying dispersion structure that generated the data. We applied DM regression to model several real world plant-pollinator networks and demonstrated how the results of our simulation study can guide one to choose an appropriate dispersion structure. However, when there are several available covariates, either group-specific or choice-specific, it is often difficult to know which, if any, of the covariates are important predictors of interaction. Consequently, model selection becomes a practical problem. Some form of regularization that can automatically identify the correct predictors and dispersion structure would be ideal. Model selection may be particularly challenging because the likelihood surface is prone to flat or near flat regions under the Dirichlet-multinomial distribution. The authors are currently working on this problem.

[6] In Stata, the `dirmul` function uses $\rho_g = \gamma_0 + \gamma_1 z_g$.

[7] For MNR, only second derivatives are numerically approximated for all DM parameterizations

Acknowledgements

The authors gratefully acknowledge *NSERC* for funding this work and T. Desmond for discussions on likelihood curves.

Chapter 5

Regularization for the grouped Dirichlet-multinomial regression model

Variable selection for grouped Dirichlet-multinomial regression

Catherine Crea and R. Ayesha Ali

Abstract

We present regularized regression for various parameterizations of the grouped Dirichlet-multinomial (DM) model using standard and adaptive lasso methods. The grouped DM regression model is often used in econometrics and ecology to model group choice behaviour, though here we focus on ecological applications within a plant-pollinator context. We implement the fast iterative shrinkage-thresholding algorithm (FISTA) for estimating the model parameters and conducting variable selection simultaneously. Our implementation of FISTA adapts the learning rate, or stepsize, both globally (using the Armijo rule) or on a dimension-specific basis (using the ADADELTA rule) and information criterion (AIC and BIC) are used to determine the optimal tuning parameter. We evaluate the performance of the proposed lasso approaches through extensive simulation studies and show that the adaptive lasso using BIC is more consistent with respect to parameter estimation and variable selection. We provide our implementation of lasso for grouped DM regression as a publicly available routine in `R` and demonstrate its performance in the analysis of an empirical plant-pollinator network.

Keywords: *Dirichlet-multinomial regression, lasso, adaptive lasso, variable selection, information criteria, proximal gradient methods, plant-pollinator networks.*

5.1 Introduction

The ultimate goal of any model selection task is two-fold: selecting the appropriate mathematical (or statistical) representation of the data and building a model that is interpretable and usable. The nature of the data will drive the former while variable selection methods can facilitate the latter. In this paper, we are concerned with the modelling of multivariate categorical count data. The ubiquitous mutinomial logit (MNL) model (McCullagh and Nelder, 1983) is frequently used to model these data, but in practice observed counts exhibit more variation than assumed by the multinomial distribution, a phenomenon referred to as overdispersion.

The Dirichlet-multinomial (DM) regression model (Guimarães and Lindrooth, 2007) has become a popular alternative to the MNL because its mean-variance structure can accommodate overdispersion. Because of its flexibility, DM regression has applications in several fields, including econometrics, genomics, social sciences and ecology. In these applications (and many others), there can be a large number of features (or covariates) that may or may not be relevant in the modelling of the counts. To reduce the size and complexity of the model, we can use variable selection techniques to pare down the features to just the relevant ones while avoiding an overfit model and increasing interpretability. One such method is regularization, where the likelihood function is constrained by some function of the model parameters, which encourages sparsity and parsimony. The goal of this paper is to provide an implementation of regularized regression for the grouped DM model for the purposes of variable selection.

Our application of DM regression is motivated by the analysis of ecological data, in particular, the interaction frequencies between plant species and pollinator species in a mutualistic network. These networks can be represented by a matrix of counts, where the pollinator species represent rows and the plant species represent the columns (or categories). The pollinator species counts across the plant species can, therefore, be assumed to follow a multinomial distribution. These networks are known to be heterogeneous (Bascompte and Jordano, 2007) and so the DM model is more appropriate than the standard MNL model. In order to better understand the mechanisms driving these interactions, we model the counts as a function of plant and pollinator traits or derived interactions between the traits, known as trait matching or linkage rules in the ecological literature (Blüthgen, 2006; Vázquez et

al., 2009a; Santamaría and Rodríguez-Gironés, 2007; Allesina et al., 2008; Stang et al., 2009; Blüthgen, 2010; Olesen et al., 2011; Bartomeus, 2013; Junker et al., 2013; Eklöf et al., 2013; Rohr et al., 2010; Gravel et al., 2013).

It is only within the past few years that field ecologists have collected detailed species data to better understand the mechanisms that drive pollination. Consequently, there is a growing need to develop variable selection methods for the grouped DM model to be able to discern which of the covariates, if any, are important indicators of interaction. Here the objective of the model fitting exercise is variable selection rather than prediction accuracy because the ultimate goal is to identify the "true" underlying mechanisms driving the interactions between plant and pollinators. If prediction is the focus, then the selected model could contain some irrelevant variables, as long as the coefficients of those variables are small (Hesterberg, 2008). This distinction is important when building models for a particular application because it facilitates the selection of the appropriate statistical methods.

Standard methods for variable selection use forward/backward strategies; however, these methods can yield models that are too small, that have MLEs with high sampling variability, and that are computationally expensive, particularly in high dimensional settings (Hastie et al., 2009). Even in lower dimensions, where the number of predictors is less than the number of observations, $p < n$, best subset selection methods can be unstable due to the inherent discreteness in the process (Breiman, 1996; Hastie et al., 2009).

Regularized regression is a popular alternative that can conduct variable selection and parameter estimation simultaneously. Regularization algorithms can also deal with multicollinearity and redundant predictors. These methods typically apply a penalty term for complexity to the log-likelihood based on the model parameters and then maximization is done on the penalized log-likelihood. Of the most prominent examples is the least absolute shrinkage and soft-thresholding operator (lasso) of Tibshirani (1996), which induces sparsity by using an $l1$ norm penalty term. Other sparsity-inducing penalties include the adaptive lasso (Zou, 2006), the fused lasso (Tibshirani et al., 2005), the group lasso (Yuan and Lin, 2006; Meier et al., 2008), and the sparse group lasso (Yuan and Lin, 2007). Alternative regularized estimators include ridge regression (Tikhonov, 1977), the elastic net (Zou and Hastie, 2005), and the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001). Although these methods have been well developed for many linear and generalized linear (GLM) regression

models, particularly those with a univariate response, regularized regression has not been developed for this implementation of the grouped DM model.

In this paper, we focus on implementing lasso-type methods to various parameterizations of the grouped DM regression model and use accelerated proximal gradient methods (Beck and Teboulle, 2009; Nesterov, 2007) for optimizing the penalized log-likelihood. More specifically, the objectives of this work include: (i) an evaluation of the performance of the standard and adaptive lasso with respect to variable selection and parameter estimation in the fixed dimension and diverging dimension scenarios, (ii) a comparison of the AIC (Akaike, 1973) and the BIC (Schwarz, 1978) for tuning parameter selection, and (iii) implementation of the fast iterative shrinkage thresholding algorithm (FISTA) which adapts the learning rate, or stepsize, both globally (using the Armijo rule (Armijo, 1966)) and on a dimension-specific basis (using the ADADELTA rule (Zeiler, 2013)). The remainder of this paper is organized as follows. Section 2 provides a brief description of the DM parameterizations for both the unpenalized and penalized grouped DM regression model and technical details of our implementation of FISTA. Section 3 investigates the performance of lasso-type methods for the regularized DM regression model via simulation and Section 4 shows the results of the analysis of an empirical plant-pollinator network using our proposed lasso approach. Discussion and conclusions are provided in Section 5.

5.1.1 Relation to Other Work

Note that there are some distinct differences between our implementation of DM regression and those in other fields. First, these counts are aggregated over all individuals within a pollinator species and so we are modelling count data at the group level, not the individual level, as is often done in most other applications, e.g., bioinformatics (Chen and Li, 2013; Zhang et al., 2016; Wang and Zhao, 2017). In light of this difference, we refer to our model as the grouped DM regression model. Secondly, since we have adapted this implementation from an econometric discrete choice model, our model is a form of the grouped conditional logit (GCL) (McFadden, 1974) for overdispersed count data (Guimarães and Lindrooth, 2007), where the logit probabilities are modelled as a function of the category-specific traits (i.e., plant species) and not the individuals' characteristics (i.e., pollinator species). Therefore, there is only one regression coefficient to estimate for each covariate introduced into the model and it is through the DM dispersion parameter that we can

choose to include pollinator- or group-specific covariates (although this parameterization is outside of the scope of this paper). Finally, within the GCL framework, we condition on the sum of counts for each pollinator species and so the group-level intercepts are dropped from the likelihood, but for the grouped DM model these intercepts do not condition out of the model and are absorbed into the dispersion parameter (Guimarães and Lindrooth, 2007).

There exist several `R` packages that provide procedures for fitting various lasso-type methods for the MNL models referenced throughout this paper, most of which use a variant of the coordinate descent algorithm. Of note is the `glmnet` package (Freidman et al., 2010) which fits lasso or elastic-net for the grouped MNL model, among many other GLMs and the `msgl` package (Vincent, 2014) which provides model fits for a sparse group lasso. Friedman et al. (2014) implement the lasso and elastic net for the conditional logistic regression model used in case-control studies (Friedman, 2014), which is available in the `clogitL1` package (Reid and Tibshirani, 2014). Adhikari et al. (2015) implement a high dimensional fused lasso using proximal gradient methods to the MNL model with time-varying covariates and demonstrate its performance on a longitudinal data set on Alzheimer's disease (available in the `glmgen` package). However, these implementations fit MNL models with only individual-specific covariates. In terms of MNL models that incorporate choice-specific covariates, Tutz et al. (2015) and Maurerer et al. (2015) develop lasso-type methods in the context of modelling electoral choices in multiparty systems. The `MRSP` (Pößnecker, 2014) package fits the group lasso, sparse group lasso and the categorically structured lasso of Tutz et al. (2015) using accelerated proximal gradient methods, while Maurerer et al. (2015) implement the lasso and adaptive lasso using proximal gradients (currently there is no official `R` package for the latter). Although these MNL models can accommodate choice-specific covariates, they are used for individual counts, not grouped counts, and they cannot accommodate overdispersion.

Currently, there is only one `R` package available for the regularized DM model. The `MGML` package (Zhang and Zou, 2017) provides an implementation of the lasso and group lasso using accelerated proximal gradients for a variety of multivariate count models applied to high-throughput data in genomics. There are two other implementations of regularized DM regression motivated from genomics, which use DM regression link nutrient intake to the human gut microbiome. The first is the sparse group lasso of Chen and Li (2013) and the second is an extension by Wang and Zhao (2017) to incorporate the phylogenetic information

using a penalized version of the Dirichlet-Tree multinomial regression model. The former uses a coordinate descent algorithm, while the latter uses accelerated proximal gradient methods and, to date, neither are official `R` packages, but are available from the authors' websites. Although these implementations can accommodate overdispersion, they do not accommodate grouped data or choice-specific covariates. We provide our implementation of regularized regression for the grouped DM model as an `R` program on GitHub (Crea, 2017).

Implementing regularization for the DM model, for both the grouped and ungrouped setting, is not a trivial task. In most applications, the model belongs to the exponential family of distributions and so the log-likelihood is convex. The MNL model is convex and so model consistency and oracle properties of the resulting method can be easily derived assuming some regularity conditions. Since the grouped DM likelihood is generally not convex and has non-smooth regions, it is difficult to prove these properties. None of the implementations for regularized DM regression have attempted to do so. Additionally, when minimizing the penalized log-likelihood, convergence is not guaranteed, but the algorithm will generally converge at least to a local minimum (Zhang et al., 2016; Wang and Zhao, 2017).

5.2 Methods

5.2.1 DM regression model for grouped data

DM regression for grouped data assumes that individuals within a group share common characteristics and are faced with the same choice set; hence, the true level of variation is at the group level (Guimarães and Lindrooth, 2007). Let X be a $G \times J \times K$ design array containing covariate information, with entries x_{gjk}. We use $\boldsymbol{x}_{gj}$ to represent the K-length vector of covariates associated with group g and choice j. The multinomial probabilities $\boldsymbol{p}_g$ are modelled as a function of covariates through a logit formulation,[1]

$$\log\left(\frac{p_{gj}}{1-p_{gj}}\right) = \boldsymbol{\beta}'\boldsymbol{x}_{gj} + \eta_{gj}, \tag{5.1}$$

for $g = 1, \ldots, G$ and $j = 1, \ldots, J$, where $\boldsymbol{\beta}'$ is a K-length vector of unknown regression coefficients associated with the covariates in $\boldsymbol{x}_{gj}$ and η_{gj} is a random group effect that

[1]Under the random utility framework (McFadden, 1974) with utility function given by the RHS of (5.1) plus independent errors that follow an extreme value distribution, the logit formulation follows directly.

accounts for unobservable heterogeneity among individuals within a group. By rearranging (5.1), we get an expression for the probability that an individual in group g selects choice j:

$$p_{gj} = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{gj} + \eta_{gj})}{\sum_{j=1}^{J}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{gj} + \eta_{gj})} = \frac{\lambda_{gj}\exp(\eta_{gj})}{\sum_{j=1}^{J}\lambda_{gj}\exp(\eta_{gj})}. \tag{5.2}$$

We further make the assumption that the $\exp(\eta_{gj})$'s follow independent gamma distributions with both shape and scale parameters $\delta_g\lambda_{gj}$, $\delta_g > 0$. Under these assumptions, it can be shown that the probabilities $\boldsymbol{p_g}$ follow a Dirichlet distribution with parameters $\boldsymbol{\alpha_g} = (\delta_g^{-1}\lambda_{g1}, \ldots, \delta_g^{-1}\lambda_{gJ})$, where δ_g is used to quantify the overdispersion. The log-likelihood for the DM distribution is:

$$\begin{aligned} l_{DMd} &= \sum_{g=1}^{G}\Big\{\log(n_g!) + \log\Gamma(\delta_g^{-1}\lambda_g) - \log\Gamma(\delta_g^{-1}\lambda_g + n_g) \\ &\quad + \sum_{j=1}^{J}\log\Gamma(\delta_g^{-1}\lambda_{gj} + y_{gj}) - \log\Gamma(\delta_g^{-1}\lambda_{gj}) - \log(y_{gj}!)\Big\}. \end{aligned} \tag{5.3}$$

where, $\lambda_g = \sum_{j=1}^{J}\lambda_{gj}$. Finally, the DM model can be re-parameterized in terms of the intragroup correlation coefficient, ρ_g, since $\rho_g = \frac{\delta_g}{\lambda_g + \delta_g}$, where $0 < \rho_g < 1$. This alternate parameterization results in the following DM log-likelihood function:

$$\begin{aligned} l_{DMr} &= \sum_{g=1}^{G}\Big\{\log(n_g!) + \log\Gamma(\rho_g^{-1} - 1) - \log\Gamma([\rho_g^{-1} - 1] + n_g) \\ &\quad + \sum_{j=1}^{J}\log\Gamma([\rho_g^{-1} - 1]p_{gj} + y_{gj}) - \log\Gamma([\rho_g^{-1} - 1]p_{gj}) - \log(y_{gj}!)\Big\}. \end{aligned} \tag{5.4}$$

Maximization of (5.3) or (5.4) provides estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}_g$ (or $\boldsymbol{\rho}_g$). In the special case that $\eta_{gj} = 0$, there is no random group effect and the DM model reduces to a group conditional logit model[2]. If, further, there are only group-specific covariates, then the model reduces to a standard multinomial logit model. Under an alternative parameterization of the DM model, the cell frequencies y_{gj} can be shown to follow a negative multinomial distribution with fixed effects (Guimarães and Lindrooth, 2007). Conditional on the row sums, the cell frequencies form overdispersed count variables following a Poisson distribution with rate $\lambda_{gj} = \exp(\boldsymbol{\beta}'\boldsymbol{x_{gj}})$, where λ_{gj} itself is a random Gamma variate with parameters $(\delta_g^{-1}\lambda_{gj}, \delta_g^{-1})$.

5.2.2 Regularized DM regression

To perform variable selection, we add a penalty term to the log-likelihood models in (5.3) and (5.4) and proceed by minimizing the penalized (negative) log-likelihood function:

$$l_p(\beta^*, \tilde{\lambda}) \quad = \quad -l(\beta^*) + \tilde{\lambda} J(\beta^*), \tag{5.5}$$

where β^* is the M-length parameter vector for which the first $K = M - 1$ elements contain the regression coefficients and the last element is the dispersion parameter (δ and ρ), $-l(\beta^*)$ is the negative DM likelihood, $J(\beta^*)$ is the penalty term, and $\tilde{\lambda} > 0$ is the tuning parameter. The penalty term, $J(\beta^*)$, determines the properties of the penalized estimator, such as inducing shrinkage, sparsity, or group structure. while the tuning parameter, $\tilde{\lambda}$, controls the balance between model fit and complexity.

Since we are interested in a sparse solution where some elements of β^* are shrunk exactly to zero, we let $J(\beta^*)$ be the l_1-norm, which is the standard lasso (Tibshirani, 1996). Within a lasso framework, the tuning parameter $\tilde{\lambda}$ determines the strength of the regularization such that smaller values of $\tilde{\lambda}$ correspond to less shrinkage and larger values of $\tilde{\lambda}$ lead to sparser solutions, i.e., $\tilde{\lambda} = 0 \Rightarrow \beta^* = \text{MLEs}$ and $\tilde{\lambda} = \infty \Rightarrow \beta^* = 0$. The penalized likelihood for the DM lasso becomes:

$$l_p(\beta^*, \tilde{\lambda}) \quad = \quad -l(\beta^*) + \tilde{\lambda} \sum_{m=1}^{M-1} |\beta^*_M|, \tag{5.6}$$

where $\sum_{m=1}^{M-1} |\beta^*_m|$ is the l_1-norm. Note we do not penalize the M^{th} element, i.e., dispersion parameter δ or ρ, but we do estimate it.

One main issue with the lasso is that it must satisfy a non-trivial necessary condition, known as the *Irrepresentable Condition*, to select the true model (see Zhao and Yu (2006) for more details). Fan and Li (2001) and Meinshausen and Bühlmann (2004) also showed that the oracle properties (consistency in variable selection and parameter estimation) do not hold for the lasso. To mitigate these issues, we also consider the adaptive lasso (Zou, 2006),

[2] $l_{GCL} = \sum_{g=1}^{G} \sum_{j=1}^{J} n_{gj} \log(p_{gj})$

which scales the l_1-norm term by an adaptive data-driven weight vector, $\hat{w}_m$. The penalized log-likelihood for the adaptive DM lasso is:

$$l_p(\beta^*, \tilde{\lambda}) \;\; = \;\; -l(\beta^*) + \tilde{\lambda} \sum_{m=1}^{M-1} \hat{w}_m |\beta^*_m|, \tag{5.7}$$

where the $\hat{w}_m$ are the adaptive weights. The weights can be any $\sqrt{n}$-consistent estimator, but here we use the MLEs, i.e., $\hat{w}_m = |\hat{\beta}^{ML}|^{-\tilde{\gamma}}$, where $\tilde{\gamma} > 0$ is a positive constant and an additional tuning parameter that adjusts the weights. The idea is to penalize the irrelevant predictors more than the relevant predictors leading to consistent model selection and optimal prediction (Zou, 2006). Minimization of the function $l_p(\beta^*, \tilde{\lambda})$ in (5.6) and (5.7) provides the penalized MLEs with certain parameters shrunk to exactly zero thus achieving variable selection and parameter estimation simultaneously.

5.2.3 Proximal gradient method

We implement FISTA (Beck and Teboulle, 2009) to minimize the penalized negative log-likelihood. FISTA is an accelerated proximal gradient method which is an extension of the classical gradient algorithm that can handle constrained non-smooth minimization problems. It is particularly useful for composite functions with non-smooth, but inexpensive penalties, such as those that often arise in regularization problems (e.g., lasso). It is attractive due to its simplicity: (i) take an (accelerated) gradient step to create a search point and (ii) project the search point onto the constrained set via the proximal operator. FISTA only uses gradient information for the updates so there is no need to compute and store Hessians, and it can easily be applied to large-scale problems (Beck and Teboulle 2009). Additionally, the acceleration remedies the slow convergence of first-order methods while the use of proximal gradients remedies the non-smooth and constrained term in the penalized log-likelihood.

More formally, let us consider an objective function $l_p(\theta)$ of the form (5.5), with non-smooth penalty term $J(\theta)$. The proximal map for $J(\theta)$ is a type of generalized projection operator:

$$\text{prox}_J(\theta) \;\; = \;\; \underset{\theta \in \boldsymbol{R}^n}{\text{argmin}} \frac{1}{2} ||\theta - z||_2^2 + \tilde{\lambda} J(\theta), \tag{5.8}$$

where the search point z is obtained by taking a gradient step with respect to the differen-

tiable term l_p. In other words, we have:

$$z \;=\; \hat{\theta}^{t+1} = \hat{\theta}^t + s\nabla l(\hat{\theta}^t), \tag{5.9}$$

where s is a stepsize that can be fixed, or chosen adaptively (see below for more details). If $J(\theta)$ is the $l1$ norm of the model parameters, e.g., $||\beta^*||_1$, then $\text{prox}_{slp}(\theta)$ is the soft-thresholding operator:

$$\text{prox}_{slp} \;=\; \text{sign}(\theta) \cdot \max(|\theta| - s, 0). \tag{5.10}$$

Using the update from (5.9), the iterations of the proximal gradient method become:

$$\hat{\theta}^{t+1} \;=\; \text{prox}_{slp}\big(\hat{\theta}^t + s\nabla l(\hat{\theta}^t)\big). \tag{5.11}$$

For FISTA, a momentum term, known as Nesterov acceleration (Nesterov, 1983; Beck and Teboulle, 2009), is added to the gradient step to reach a faster convergence rate. The momentum term, $\hat{\nu}^t$, extrapolates $\hat{\theta}^t$ by forming a linear combination of the previous two points using acceleration factors a_t[3]:

$$\hat{\nu}^t \;=\; \hat{\theta}^t + \frac{a_{t-1} - 1}{a_t}(\hat{\theta}^t - \hat{\theta}^{t-1}). \tag{5.12}$$

This final extrapolated point, $\hat{\nu}^t$, is used in place of $\hat{\theta}^t$ for the proximal updates in (5.11).

Two points are noteworthy. First, the complexity of the algorithm scales well with both the number of covariates and the network size. Second, the DM likelihood is generally non-convex; therefore, convergence is not guaranteed. Generally, the algorithm converges to a local minimizer, since the objective value decreases over iterations.

Adaptive learning rate

In practice, the choice of the stepsize significantly affects the rate of convergence, stability and computational efficiency of iterative direction methods. If it is too small, convergence may be very slow. If it is too large, convergence may not be smooth (divergence). Exact

[3]For the first iteration of the proximal gradient method, we initialize $a_0 = 0$, $a_1 = 1$, and for subsequent updates $a_{t+1} = \frac{1+\sqrt{1+4a_t^2}}{2}$, as per Beck and Teboulle (2009).

computation can be expensive and common methods include a constant or global stepsize using the Armijo rule for a backtracking line search or a dimension-specific stepsize, such as the ADADELTA rule (see Zeiler, 2013 for specific algorithmic details).

Briefly, ADADELTA is used for gradient descent methods. Unlike the backtracking line search, it requires no manual tuning of the stepsize. This is because each parameter has its own stepsize that is inversely proportional to the exponential moving average of historical gradients. The idea behind ADADELTA is that parameters whose gradients were high in previous iterations may be close to the optimal value and so a slower stepsize may be more appropriate, and vice versa. We use this stepsize rule when running FISTA for the constant intracorrelation parameterization of the DM model.

5.2.4 Tuning parameter selection

Selecting the optimal value of $\tilde{\lambda}$ is essential to the performance of the lasso because it controls the trade-off between model fit and model size/complexity. We use the AIC and the BIC to select the optimal $\tilde{\lambda}$, defined as

$$AIC(\tilde{\lambda}) \quad = \quad -2l_p(\hat{\beta}, \tilde{\lambda}) + 2\text{df}(\hat{\beta}, \tilde{\lambda}) \tag{5.13}$$

and

$$BIC(\tilde{\lambda}) \quad = \quad -2l_p(\hat{\beta}, \tilde{\lambda}) + \log(N)\text{df}(\hat{\beta}, \tilde{\lambda}), \text{ repectively,} \tag{5.14}$$

where $\text{df}(\hat{\beta}, \tilde{\lambda})$ is the effective degrees of freedom (or nonzero coefficients) implied by $\hat{\beta}(\tilde{\lambda})$ (Zou et al., 2007). We then select $\tilde{\lambda}$ by minimizing AIC or BIC over an equally-spaced log grid of 100 $\tilde{\lambda}$ values, starting from $\tilde{\lambda}_{\max}$ to $\tilde{\lambda}_{\min}$. To determine $\tilde{\lambda}_{\max}$, we begin with the smallest $\tilde{\lambda}$ such that all regression coefficients are 0, except for the dispersion parameter, δ or ρ. This can be achieved by setting $\tilde{\lambda}_{\max} = \max_M \frac{|S_M|}{\tilde{w}_M}$, where S_M is the gradient vector (Appendix I) evaluated at $\beta^*_m = 0, \ (m \neq M)$[4]. Note that we set $\tilde{\lambda}_{\min} = 0.01$ to avoid numerically unstable solutions at $\tilde{\lambda} = 0$.

[4]For the δ and ρ parameterizations, the last element of β^* contains the (unpenalized) MLE of the dispersion parameter.

For the adaptive lasso, there is an additional tuning parameter, $\tilde{\gamma}$, that needs to be optimized. We consider this additional tuning parameter from the set $\{\tilde{\gamma} = 1, 2, 3\}$[5]. For each $\tilde{\gamma}$, we run FISTA over a log grid of $\tilde{\lambda}$, to find the optimal $\tilde{\lambda}$; however, an alternative is to simultaneously optimize the tuning parameters using a two-dimensional lattice grid of values for both $\tilde{\lambda}$ and $\tilde{\gamma}$, following the suggestion of Zou (2006).

For linear regression models with a fixed predictor dimension, it has been shown that using BIC-type criteria can identify the true model consistently for the adaptive lasso (Wang and Leng, 2007) and for the SCAD (Wang et al., 2007). Wang et al. (2009) show the same model consistency property for a diverging number of predictors. Model consistency using BIC-type criteria has not been proven for implementations of the lasso for the DM regression model, but BIC is often the preferred selection criterion (see for example Wang and Zhao, 2017; Chen and Li, 2013; and Zhang et al., 2016). Maurerer et al. (2015) are in favour of using AIC, as opposed to BIC or cross-validation (CV), to select $\tilde{\lambda}$ when the adaptive lasso is applied to the MNL model that incorporates both individual- and category-specific variables. In their experience, BIC results in the most amount of sparsity and CV results in the least amount of sparsity. Here we investigate both AIC and BIC. Although AIC has been shown to overfit, it is asymptotically loss efficient. Zhang et al. (2010) define an efficient criterion as one that "selects the model so that its average squared error is asymptotically equivalent to the minimum offered by the candidate models when the true model is approximated by a family of candidate models". We expect these theoretical findings to hold empirically for our application as well.

Note that cross-validation is an alternative and widely used procedure for selecting $\tilde{\lambda}$ in the lasso; however, it is computationally expensive, particularly in high dimensions, and it is better suited when the model fitting goal is prediction rather than selection, since it is asymptotically loss efficient and selection inconsistent (Wang et al. 2007; Zhang et al., 2010). Also, standard techniques for estimating the prediction error do not apply to the DM models. More specifically, in group conditional logit models, the nuisance parameters (or group specific intercepts) are conditioned out of the likelihood so care must be taken when data are being split into training and testing sets. Reid and Tibshirani (2014) address this issue when implementing the lasso for the conditional logistic regression model used in case-

[5]Zou (2006) suggests varying $\tilde{\gamma}$ between 0.5 and 2. Preliminary simulation results suggested that values of $\tilde{\gamma} \geq 1$ were needed to ensure selection consistency; therefore, we chose to vary $\tilde{\gamma}$ between 1 and 3.

control studies. They adopt an approach by van Houwelingen et al. (2006), where complete strata are left out at a time and the CV error is the sum over the left out strata of each stratum's log conditional likelihood contribution (Reid and Tibshirani, 2014). However, in their framework, strata are of the same size and the number of cases and controls in each strata are fixed. This is not the case for the DM model because the group sizes are different for each pollinator species and in the DM setting the group level intercepts are not conditioned out of the model, but are absorbed into the dispersion parameter, δ (Guimarães and Lindrooth, 2007). Nonetheless, preliminary analyses (not shown) included CV as a tuning parameter selection method, where both the approach by Reid and Tibshirani (2014) and the DM deviance was used to select the optimal $\tilde{\lambda}$. In general, the lasso solutions tended to overfit the model by including many of the irrelevant predictors (a weakness of both CV and AIC methods), which is not surprising because CV is loss efficient.

5.3 Simulation Study

The simulation study was designed to evaluate the performance of lasso-type methods for three parameterizations of the DM regression model: no dispersion (none)[6], constant (dconst: $\delta_g = \delta$), and constant intragroup correlation (rconst: $\rho_g = \rho$). Networks were generated in `R`, per Crea et al. (2016), based on the gamma-Poisson parameterization of the DM model. We considered three network sizes: small (25×10), medium (50×20) and large (100×30), i.e., $N = G \times J = 250, 1000, 3000$, respectively. For the dconst and rconst parameterizations, we considered low and medium levels of dispersion, specifically $\delta = 2, 6$ and $\rho = 0.2, 0.5$. Entries of the covariate array X were generated based on complementarity linkage rules following Santamaría and Rodríguez-Gironés (2007). Refer to Crea et al. (2016) for specific details on the construction of these binary covariates. For each parameterization, network size and dispersion level, 100 replicates are simulated for each scenario defined below.

5.3.1 Simulation Scenarios

Scenario 1: Fixed K

For this scenario, we assess the performance of DM lasso in the fixed parameter dimension

[6]The no dispersion case corresponds to assuming that the variance of the random group effects is zero, $\text{Var}(\eta_{gj}) = 0$, which corresponds to the GCL model.

setting. In this setting, we based our cases on personal experience with analyzing empirical networks and those typically found in the ecological literature. We consider three cases, where the predictor dimension increases with increasing K, the proportion of relevant to irrelevant predictors is fixed at 0.25 for all K, and we increase the network size:

Case 1a. $K = 10$, $\alpha_0 = 2$, $\beta^* = (-0.5, 2, 0, \ldots, 0)$, $N = 250, 1000, 3000$.
Case 1b. $K = 20$, $\alpha_0 = 4$, $\beta^* = (-0.5, 1, -1, 2, 0, \ldots, 0)$, $N = 250, 1000, 3000$.
Case 1c. $K = 30$, $\alpha_0 = 6$, $\beta^* = (-0.5, 0.5, -1, 1, -2, 2, 0, \ldots, 0)$, $N = 250, 1000, 3000$.

Scenario 2: Diverging K_N
For this scenario, we investigate the effects of the dimensionality on the performance of DM lasso. In most model selection problems, it is valid to assume that the number of parameters should be large and grow with network size. We consider three cases, where the predictor dimension K is diverging at a rate of $\sqrt{N}$, which corresponds to a decreasing proportion of relevant to irrelevant predictors, and the dimension of the true model is fixed:

Case 2a. $K_{250} = 16$, $\alpha_0 = 10$, $\beta^* = (-0.5, 0.5, 1, -1, 1.5, -1.5, 2, -2, 2.5, -2.5, \ldots, 0)$.
Case 2b. $K_{1000} = 32$, $\alpha_0 = 10$, $\beta^* = (-0.5, 0.5, 1, -1, 1.5, -1.5, 2, -2, 2.5, -2.5, \ldots, 0)$.
Case 2b. $K_{3000} = 55$, $\alpha_0 = 10$, $\beta^* = (-0.5, 0.5, 1, -1, 1.5, -1.5, 2, -2, 2.5, -2.5, \ldots, 0)$.

Note that for the δ and ρ parameterizations, β^* also includes the dispersion parameter as the last element of this parameter vector. For example, for Case 1a, when the rconst DM model is being fit, then $\beta^* = (-0.5, 2, 0, \ldots, 0, 0.2)$ and $\beta^* = (-0.5, 2, 0, \ldots, 0, 0.5)$ for low and medium dispersion levels, respectively. For both scenarios we fit the simulated networks using standard lasso (lasso) and adaptive lasso (alasso). All simulations and model fits were performed in `R` (R Core Team, 2017) using a self-written implementation available online (Crea, 2017).

5.3.2 Performance Metrics

We compare methods with respect to variable selection and parameter estimation. To evaluate variable selection, we calculate the percent of models among the 100 replicates that selected the true model, an underfit model and an overfit model. We also calculate the average number of zero coefficients correctly estimated to be nonzero (i.e., true positives) and the

average number of nonzero coefficients incorrectly estimated to be zero (i.e., false negatives) over the 100 replicates. To evaluate parameter estimation, we calculate the mean squared error of $\hat{\beta}^*$ by MSE $= \sum_M (\hat{\beta}^* - \beta^*)^2/M$ and then take the mean over the 100 replicates.

5.3.3 Simulation results

Scenario 1: Fixed K

Tables 5.1 through 5.3 present the results of the simulation study for Scenario 1: the fixed dimension case. In general, the adaptive lasso using BIC outperformed the standard lasso for all parameterizations of the DM model and for all cases, i.e., fixed values of K. For brevity, we show the results for Case 1b for each of the three parameterizations of the DM models. Results for Cases 1a and 1c are in Appendix K.

Table 5.1 shows the results for Case 1b, i.e., $K = 20$, for the no dispersion parameterization of the DM model. None of the methods returned underfitted models; they either provided the correct fit or overfitted ones. None of the methods resulted in false negatives and all mean MSEs were relatively low and comparable. For a given network size, the adaptive lasso with $\tilde{\gamma} = 3$ had the highest percentage of correct model fits, the highest true positives, and the lowest mean MSE for both BIC and AIC. Further, the standard lasso was never able to recover the true model using either AIC or BIC, whereas for the adaptive lasso, we see that the performance with respect to all metrics increased with increasing $\tilde{\gamma}$. For these adaptive lasso models, BIC consistently outperformed AIC, particularly for model selection. For example, for the adaptive lasso with $\tilde{\gamma} = 3$, the percent of correct fits ranged from 75% to 96% using BIC and ranged from 19% to 78% using AIC. Note that the performance of all methods increased with increasing network size and these same trends persisted for Cases 1a and 1b; however, the overall performance across methods slightly decreased with increasing K.

Table 5.2 shows the results for Case 1b, i.e., $K = 20$, for the constant dispersion, or dconst, parameterization of the DM model. None of the methods returned underfitted models, except for four instances when $N = 250$ for the adaptive lasso. Also, none of the methods showed any false negatives, except for eight instances when $N = 250$ for the adaptive lasso, although they were all very low. The same trends observed for the no dispersion parameterization persisted for the constant dispersion parameterization: standard lasso was never

able to recover the true model; the adaptive lasso with $\tilde{\gamma} = 3$ had the highest percentage of correct model fits, the highest number of true positives, and the lowest mean MSE; the performance of all methods increased with increasing network size; and, BIC outperformed AIC. In fact, the improvement gained by using BIC over AIC for this parameterization was more prominent for all performance metrics, especially for variable selection. This result is not surprising since BIC has been shown to be variable consistent for other implementations of the lasso-type methods (Wang and Leng, 2007; Wang et al., 2009). Note that the performance of all methods slightly decreased with increasing dispersion and increasing K (see Appendix K).

Table 5.3 shows the results for Case 1b, i.e., $K = 20$, for the constant intra-correlation, or rconst, parameterization of the DM model. Some of the same trends observed for the other two parameterizations persisted for this parameterization: standard lasso was never able to recover the true model when using AIC, the adaptive lasso with $\tilde{\gamma} = 3$ had the highest percentage of correct model fits, the highest true positives, and the lowest mean MSE, the performance of all methods increased with increasing network size, BIC outperformed AIC significantly with respect to model selection, and performance decreased as dispersion increased. It is interesting to note that, in general, model selection was lowest and the MMSE was highest for this parameterization. For example, when $N = 250$ and $\rho = 0.5$, BIC tended to underfit a large percentage of models, particularly for the standard lasso, and the mean MSE was as high as 13.75 when the model was selected using AIC.

Table 5.1: Simulation results for Scenario 1 (Fixed K), Case 1b ($K = 20$), No Dispersion Model*

		AIC						BIC					
N	Method	Under	Correct	Over	TP	FN	MMSE	Under	Correct	Over	TP	FN	MMSE
250	lasso	0	0	100	0.18	0	0.82	0	0	100	1.52	0	0.78
	alasso ($\gamma = 1$)	0	0	100	8.96	0	0.56	0	14	86	13.54	0	0.33
	alasso ($\gamma = 2$)	0	9	91	12.72	0	0.44	0	57	43	15.27	0	0.26
	alasso ($\gamma = 3$)	0	19	81	13.78	0	0.38	0	75	25	15.62	0	0.24
1000	lasso	0	0	100	0.13	0	0.15	0	0	100	0.47	0	0.15
	alasso ($\gamma = 1$)	0	1	99	8.68	0	0.11	0	17	83	13.81	0	0.06
	alasso ($\gamma = 2$)	0	13	87	12.64	0	0.08	0	84	16	15.78	0	0.04
	alasso ($\gamma = 3$)	0	37	63	14.81	0	0.05	0	93	7	15.92	0	0.04
3000	lasso	0	0	100	0.31	0	0.04	0	0	100	1.66	0	0.05
	alasso ($\gamma = 1$)	0	0	100	9.43	0	0.03	0	22	78	14.44	0	0.02
	alasso ($\gamma = 2$)	0	14	86	13.45	0	0.02	0	87	13	15.86	0	0.01
	alasso ($\gamma = 3$)	0	78	22	15.76	0	0.01	0	96	4	15.96	0	0.01

* alasso - adaptive lasso; Under - percentage of under-fit models; Correct - percentage of correct fit models; Over - percentage of over-fit models; TP - mean number of true positives; FN - mean number of false negatives; MMSE - mean of mean squared errors ($\times 100$).

Table 5.2: Simulation results for Scenario 1 (Fixed K), Case 1b ($K = 20$), Constant Dispersion Model*

		AIC						BIC					
N	Method	Under	Correct	Over	TP	FN	MMSE	Under	Correct	Over	TP	FN	MMSE
$\boldsymbol{\delta = 2}$													
250	lasso	0	0	100	0.27	0	2.58	0	0	100	2.91	0	2.31
	alasso ($\gamma = 1$)	0	0	100	8.29	0	1.87	0	9	91	12.91	0	1.22
	alasso ($\gamma = 2$)	0	1	99	11.67	0	1.57	1	36	63	14.74	0.01	1.04
	alasso ($\gamma = 3$)	0	8	92	12.87	0	1.47	2	53	45	15.26	0.02	0.99
1000	lasso	0	0	100	0.1	0	0.53	0	0	100	0.9	0	0.52
	alasso ($\gamma = 1$)	0	0	100	8.42	0	0.4	0	16	84	13.8	0	0.23
	alasso ($\gamma = 2$)	0	9	91	12.74	0	0.31	0	64	36	15.54	0	0.19
	alasso ($\gamma = 3$)	0	21	79	14.17	0	0.26	0	85	15	15.83	0	0.18
3000	lasso	0	0	100	0.14	0	0.2	0	0	100	0.49	0	0.2
	alasso ($\gamma = 1$)	0	1	99	7.72	0	0.12	0	21	79	13.92	0	0.07
	alasso ($\gamma = 2$)	0	10	90	12.79	0	0.09	0	79	21	15.74	0	0.06
	alasso ($\gamma = 3$)	0	41	59	14.96	0	0.07	0	92	8	15.91	0	0.05
$\boldsymbol{\delta = 6}$													
250	lasso	0	0	100	0.36	0	4.88	0	0	100	3.39	0	4.35
	alasso ($\gamma = 1$)	0	0	100	6.77	0.01	4.01	0	4	96	11.83	0.04	2.44
	alasso ($\gamma = 2$)	0	2	98	10.07	0.03	3.42	3	21	76	13.87	0.13	2.08
	alasso ($\gamma = 3$)	0	7	93	11.72	0.04	3.07	5	34	61	14.7	0.16	1.84
1000	lasso	0	0	100	0.15	0	0.78	0	0	100	1.24	0	0.76
	alasso ($\gamma = 1$)	0	0	100	7.76	0	0.6	0	7	93	13.19	0	0.35
	alasso ($\gamma = 2$)	0	8	92	12.55	0	0.45	0	49	51	15.17	0	0.27
	alasso ($\gamma = 3$)	0	12	88	13.59	0	0.4	0	76	24	15.72	0	0.23
3000	lasso	0	0	100	0.15	0	0.24	0	0	100	0.47	0	0.24
	alasso ($\gamma = 1$)	0	0	100	6.55	0	0.19	0	11	89	12.81	0	0.11
	alasso ($\gamma = 2$)	0	8	92	12.42	0	0.14	0	69	31	15.66	0	0.08
	alasso ($\gamma = 3$)	0	31	69	14.6	0	0.1	0	90	10	15.9	0	0.07

* alasso - adaptive lasso; Under - percentage of under-fit models; Correct - percentage of correct fit models; Over - percentage of over-fit models; TP - mean number of true positives; FN - mean number of false negatives; MMSE - mean of mean squared errors ($\times 100$).

Table 5.3: Simulation results for Scenario 1 (Fixed K), Case 1b ($K = 20$), Constant Intra-correlation Model*

		AIC						BIC					
N	Method	Under	Correct	Over	TP	FN	MMSE	Under	Correct	Over	TP	FN	MMSE
$\rho = 0.2$													
250	lasso	0	0	100	0.61	0	5.07	4	7	89	9.6	0.15	3.52
	alasso ($\gamma = 1$)	0	0	100	7.88	0.04	3.77	4	14	82	13.34	0.15	2.18
	alasso ($\gamma = 2$)	0	1	99	10.87	0.08	3.24	10	28	62	14.43	0.23	2.08
	alasso ($\gamma = 3$)	0	5	95	12.21	0.1	3.06	12	37	51	14.86	0.26	2.03
1000	lasso	0	0	100	0.64	0	1.26	0	1	99	5.49	0	1.09
	alasso ($\gamma = 1$)	0	1	99	8.77	0	0.92	0	23	77	14.08	0	0.51
	alasso ($\gamma = 2$)	0	9	91	12.14	0	0.76	0	56	44	15.37	0	0.42
	alasso ($\gamma = 3$)	0	14	86	13.24	0	0.69	0	69	31	15.58	0	0.41
3000	lasso	0	0	100	0.5	0	0.49	0	0	100	2.07	0	0.47
	alasso ($\gamma = 1$)	0	1	99	8.9	0	0.35	0	22	78	14.17	0	0.2
	alasso ($\gamma = 2$)	0	10	90	12.55	0	0.28	0	71	29	15.65	0	0.15
	alasso ($\gamma = 3$)	0	17	83	13.85	0	0.24	0	85	15	15.84	0	0.14
$\rho = 0.5$													
250	lasso	0	0	100	1.27	0.04	13.75	66	10	24	14.99	1.84	12.38
	alasso ($\gamma = 1$)	0	1	99	7.26	0.19	11.08	28	12	60	13.62	0.75	6.43
	alasso ($\gamma = 2$)	1	3	96	9.68	0.24	10.17	23	13	64	13.67	0.66	6.68
	alasso ($\gamma = 3$)	2	3	95	10.83	0.33	9.59	26	13	61	14.12	0.72	6.69
1000	lasso	0	0	100	0.77	0	3.11	3	9	88	11.28	0.09	2.28
	alasso ($\gamma = 1$)	0	1	99	8.28	0.03	2.29	5	20	75	14.2	0.11	1.27
	alasso ($\gamma = 2$)	0	2	98	11.33	0.04	1.96	7	46	47	15.18	0.13	1.17
	alasso ($\gamma = 3$)	0	5	95	12.4	0.04	1.88	12	58	30	15.49	0.19	1.17
3000	lasso	0	0	100	0.45	0	1.23	0	3	97	7.02	0	1.04
	alasso ($\gamma = 1$)	0	1	99	8.59	0	0.89	0	24	76	14.45	0	0.5
	alasso ($\gamma = 2$)	0	7	93	12.45	0	0.72	3	62	35	15.52	0.03	0.47
	alasso ($\gamma = 3$)	0	16	84	13.62	0	0.67	4	76	20	15.77	0.05	0.44

* alasso - adaptive lasso; Under - percentage of under-fit models; Correct - percentage of correct fit models; Over - percentage of over-fit models; TP - mean number of true positives; FN - mean number of false negatives; MMSE - mean of mean squared errors ($\times 100$).

Scenario 2: Diverging K_N

Tables 5.4 through 5.6 present the results of the simulation study for Scenario 2: the diverging dimension setting. Note that for $N = 3000$ and $K_N = 55$, the number of replicates were 45 for the no dispersion and dconst model fits and 10 for the rconst model fits. Nonetheless, in general, the results parallel those from Scenario 1 for the fixed K: the adaptive lasso with $\tilde{\gamma} = 3$ had the highest percentage of correct model fits, the highest true positives, and the lowest mean MSE, though the adaptive lasso with $\tilde{\gamma} = 2$ was competitive for the latter two metrics. Further, BIC outperformed AIC with respect to model selection, and performance decreased as dispersion increased. Interestingly, for the no dispersion and dconst parameterizations (Tables 5.4 and 5.5), performance improved with increasing K_N when BIC was used, but worsened when AIC was used. On the other hand, for rconst (Table 5.6), the performance of all methods decreased with increasing K_N using AIC, and most methods when using BIC, with the exception of the adaptive lasso fits where $\gamma = 2$ or $\gamma = 3$. For these latter two methods, the highest percent of correct models was 60% when $\rho = 0.2$ and 17% when $\rho = 0.5$, compared to the dconst models which were able to achieve 84% and 89% when $\delta = 2$ and $\delta = 6$, respectively. Finally, similar to the fixed K case, the mean MSE was highest for the rconst parameterization.

Table 5.4: Simulation results for Scenario 2 (Diverging K_N), Case 2a (No Dispersion Model)*

		AIC						BIC					
N/K	Method	Under	Correct	Over	TP	FN	MMSE	Under	Correct	Over	TP	FN	MMSE
250/16	lasso	0	0	100	0.02	0	0.59	0	0	100	0.1	0	0.59
	alasso ($\gamma = 1$)	0	0	100	1.75	0	0.57	0	3	97	2.82	0	0.53
	alasso ($\gamma = 2$)	0	19	81	4.33	0	0.47	0	46	54	5.19	0	0.42
	alasso ($\gamma = 3$)	0	56	44	5.39	0	0.4	0	84	16	5.84	0	0.37
1000/32	lasso	0	0	100	0.08	0	0.07	0	0	100	0.29	0	0.07
	alasso ($\gamma = 1$)	0	0	100	8.58	0	0.06	0	0	100	14.89	0	0.05
	alasso ($\gamma = 2$)	0	5	95	17.27	0	0.04	0	50	50	21.29	0	0.03
	alasso ($\gamma = 3$)	0	57	43	21.46	0	0.03	0	89	11	21.88	0	0.03
3000/55	lasso	0	0	100	0.05	0	0.02	0	0	100	0.21	0	0.02
	alasso ($\gamma = 1$)	0	0	100	20.84	0	0.01	0	2	98	38.41	0	0.01
	alasso ($\gamma = 2$)	0	7	93	37.85	0	0.01	0	74	26	44.65	0	< 0.01
	alasso ($\gamma = 3$)	0	91	9	44.89	0	0	0	96	4	44.95	0	< 0.01

* alasso - adaptive lasso; Under - percentage of under-fit models; Correct - percentage of correct fit models; Over - percentage of over-fit models; TP - mean number of true positives; FN - mean number of false negatives; MMSE - mean of mean squared errors ($\times 100$).

Table 5.5: Simulation results for Scenario 2 (Diverging K_N), Case 2b (Constant Dispersion Model)*

		AIC						BIC					
N/K	Method	Under	Correct	Over	TP	FN	MMSE	Under	Correct	Over	TP	FN	MMSE
$\boldsymbol{\delta = 2}$													
250/16	lasso	0	0	100	0.10	0	2.43	0	0	100	0.15	0	2.43
	alasso ($\gamma = 1$)	0	0	100	1.72	0	2.32	0	2	98	2.75	0	2.21
	alasso ($\gamma = 2$)	0	9	91	3.83	0	2.09	0	31	69	4.92	0	1.89
	alasso ($\gamma = 3$)	0	36	64	4.95	0	1.87	0	68	32	5.62	0	1.75
1000/32	lasso	0	0	100	0.11	0	0.28	0	0	100	0.34	0	0.28
	alasso ($\gamma = 1$)	0	0	100	7.67	0	0.22	0	2	98	15.96	0	0.15
	alasso ($\gamma = 2$)	0	4	96	16.76	0	0.16	0	48	52	21.2	0	0.12
	alasso ($\gamma = 3$)	0	32	68	20.62	0	0.12	0	84	16	21.83	0	0.11
3000/55	lasso	0	0	100	0.07	0	0.07	0	0	100	0.32	0	0.07
	alasso ($\gamma = 1$)	0	0	100	20.98	0	0.05	0	0	100	37.81	0	0.03
	alasso ($\gamma = 2$)	0	4	96	37.85	0	0.03	0	57	43	44.32	0	0.02
	alasso ($\gamma = 3$)	0	51	49	44.14	0	0.02	0	86	14	44.86	0	0.02
$\boldsymbol{\delta = 6}$													
250/16	lasso	0	0	100	0.08	0	3.93	0	0	100	0.23	0	3.92
	alasso ($\gamma = 1$)	0	0	100	1.45	0	3.82	0	1	99	2.52	0.02	3.66
	alasso ($\gamma = 2$)	0	3	97	3.42	0.03	3.57	3	25	72	4.54	0.08	3.32
	alasso ($\gamma = 3$)	3	21	76	4.47	0.07	3.34	7	54	39	5.32	0.11	3.12
1000/32	lasso	0	0	100	0.12	0	0.56	0	0	100	0.36	0	0.56
	alasso ($\gamma = 1$)	0	0	100	7.32	0	0.47	0	1	99	15.87	0	0.31
	alasso ($\gamma = 2$)	0	0	100	16.26	0	0.34	0	26	74	20.53	0	0.23
	alasso ($\gamma = 3$)	0	15	85	19.39	0	0.28	0	76	24	21.72	0	0.20
3000/55	lasso	0	0	100	0.04	0	0.13	0	0	100	0.18	0	0.13
	alasso ($\gamma = 1$)	0	0	100	15.9	0	0.09	0	0	100	33.94	0	0.06
	alasso ($\gamma = 2$)	0	0	100	36.79	0	0.06	0	42	58	44.00	0	0.04
	alasso ($\gamma = 3$)	0	26	74	42.64	0	0.04	0	89	11	44.88	0	0.03

* alasso - adaptive lasso; Under - percentage of under-fit models; Correct - percentage of correct fit models; Over - percentage of over-fit models; TP - mean number of true positives; FN - mean number of false negatives; MMSE - mean of mean squared errors ($\times 100$).

Table 5.6: Simulation results for Scenario 2 (Diverging K_N), Case 2c (Constant Intra-correlation Model)*

		AIC						BIC					
N	Method	Under	Correct	Over	TP	FN	MMSE	Under	Correct	Over	TP	FN	MMSE
$\rho = 0.2$													
250/16	lasso	0	0	100	0.19	0.01	9.29	0	0	100	0.30	0.04	9.27
	alasso ($\gamma = 1$)	0	0	100	1.33	0.08	9.15	2	1	97	2.27	0.17	8.75
	alasso ($\gamma = 2$)	1	5	94	2.81	0.19	8.85	5	14	81	3.88	0.34	8.16
	alasso ($\gamma = 3$)	3	10	87	3.60	0.31	8.57	14	32	54	4.69	0.44	8.02
1000/32	lasso	0	0	100	0.47	0	2.16	0	0	100	1.25	0	2.11
	alasso ($\gamma = 1$)	0	0	100	7.25	0	1.91	0	0	100	15.13	0.02	1.21
	alasso ($\gamma = 2$)	0	0	100	13.81	0.02	1.56	1	9	90	19.46	0.04	1.02
	alasso ($\gamma = 3$)	0	2	98	16.70	0.03	1.36	7	30	63	20.89	0.08	0.92
3000/55	lasso	0	0	100	0.37	0	0.77	0	0	100	3.10	0	0.74
	alasso ($\gamma = 1$)	0	0	100	19.76	0	0.57	0	0	100	38.31	0	0.25
	alasso ($\gamma = 2$)	0	0	100	32.80	0	0.41	0	31	69	43.23	0	0.20
	alasso ($\gamma = 3$)	0	1	99	36.79	0	0.36	0	58	42	44.40	0	0.18
$\rho = 0.5$													
250/16	lasso	0	0	100	0.19	0.02	31.50	0	0	100	0.37	0.07	30.88
	alasso ($\gamma = 1$)	0	1	99	1.33	0.19	30.96	0	5	95	2.33	0.41	28.75
	alasso ($\gamma = 2$)	0	3	97	2.41	0.40	30.95	5	11	84	3.36	0.61	29.24
	alasso ($\gamma = 3$)	2	11	87	3.20	0.59	30.40	20	17	63	4.09	0.83	28.10
1000/32	lasso	0	0	100	0.44	0	5.74	0	0	100	2.32	0	5.35
	alasso ($\gamma = 1$)	0	0	100	6.91	0.03	5.12	0	1	99	15.84	0.14	2.94
	alasso ($\gamma = 2$)	0	0	100	13.09	0.11	4.29	0	2	98	18.36	0.25	2.89
	alasso ($\gamma = 3$)	0	0	100	15.84	0.15	3.88	5	12	83	19.69	0.34	2.82
3000/55	lasso	0	0	100	0.70	0	2.14	0	0	100	13.25	0	1.81
	alasso ($\gamma = 1$)	0	0	100	19.29	0	1.59	0	1	99	37.64	0.02	0.71
	alasso ($\gamma = 2$)	0	0	100	30.45	0	1.27	0	14	86	41.81	0.03	0.60
	alasso ($\gamma = 3$)	0	0	100	34.53	0.01	1.16	4	26	70	43.45	0.07	0.57

* alasso - adaptive lasso; Under - percentage of under-fit models; Correct - percentage of correct fit models; Over - percentage of over-fit models; TP - mean number of true positives; FN - mean number of false negatives; MMSE - mean of mean squared errors ($\times 100$).

5.4 Regularized analysis of ecological networks

We analyzed an empirical plant-pollinator network located in Terceira Island (Azores) in the North Atlantic Ocean region, and compared the penalized and unpenalized approaches. The final models were compared using deviance, AIC, BIC and Pearson χ^2 goodness-of-fit statistics. For lasso-type estimators, we also adopted the hybrid approach of Efron et al. (2004) and refit the model using standard DM regression, but only for those predictors estimated to be nonzero, to obtain the unpenalized MLEs and standard errors for the model parameters. Following the results of the simulation study, we only consider the adaptive lasso with $\tilde{\gamma} = 3$. Since we do not know the true underlying dispersion structure for these data, we decided to fit both the no dispersion and dconst models, but not the rconst model since the lasso performance was suboptimal for this parameterization. Finally, we also fit the DM model where δ is modelled as a function of pollinator specific covariates (dfunc) because pollinator-specific covariates were available. Note that this parameterization was not studied via simulation.

5.4.1 Description of Terceira Island Network

Insect sampling was carried out from June to September in 2013 and 2014 from 50 transects, ten (10m $\times$ 1m) transects for each of five sites (or habitat types): natural forest, naturalized vegetation areas, exotic forest, semi-natural pasture, and intensively managed pasture. Transects were selected to include areas of dense flowering and each flower along a transect was surveyed for four minutes. An interaction was recorded if an insect was observed on an flower exhibiting pollinator-like behaviour, such as, probing for nectar or eating/collecting pollen. These flower-visiting insects were observed and collected with a pooter when it was not possible to identify them in the field. The specimens collected were sorted first into morphospecies and later identified following the taxonomic nomenclature in Borges et al. (2010). Data were pooled across habitat type and time because the primary interest was in the range of plant-insect interactions across the region. The pooled network consisted of G=54 insect species, J=48 plant species (i.e., N=2592), and a total of 2134 observed interactions (flower visits). There were 9 unidentified insect species that were removed from the network because no trait information was available; therefore, the final network consisted of G=45, J=48 (i.e., N=2160), and a total of 2018 observed flower visits.

Plant and insect traits were compiled using existing published and unpublished datasets

from the region: all plant species traits were classified following the information on Schaefer (2003, 2005), Sjgren (2001), Streeter (1983) and Dias (1996); and the insect species traits were classified following the information on Barnard (2011), Base de Dados da Biodiversidade dos Aores (2016), Bee, Wasp and Ant Recording Society (2016), Evans (2008), Jentzsch (2014), Marsh (1994), Pendleton and Pendleton (2016), Speight and Sarthou (2012), Rojo et al. (1997), Royal Entomological Society of London (1950, 1972), and Lafontaine and Troubridge (1998). Plant traits included: (1) life span: short (annual/biennial) or long (perennial); (2) flower type: single flower or inflorescence; (3) corolla colour: red/blue, white, or yellow; (4) flower size: small, medium, or large; (5) floral symmetry: zygomorphic or actinomorphic; (6) plant origin: native/endemic or introduced; (7) plant morphology: monoecious or dioecious; (8) corolla shape: regular or irregular; and (9) type of plant: herbaceous or woody. Pollinator traits included: (1) minimum and maximum body length (mm); (2) insect behaviour: social or solitary; (3) insect trophic: herbivore or non-herbivore; and (4) insect origin: native/endemic or introduced.

5.4.2 Results and Interpretation of Terceira Island Network

Table 5.7 summarizes the results of the penalized and unpenalized model fits to the Terceira Island network data. Since the underlying dispersion structure is unknown for these data, we fit the data using the none, dconst and dfunc parameterizations. For each parameterization we provide the unpenalized MLEs, the refit of the final models chosen by adaptive lasso with $\tilde{\gamma} = 3$ using AIC and BIC. Note that none of the model fits provided a good fit to the data according to the χ^2 goodness-of-fit test (p-values < 0.05). Possible reasons include: (i) the quality of the asymptotic approximation of the χ^2-statistic to follow a χ^2 distribution is poor for the DM model, (ii) the large amount of zeroes in the data, or (iii) the analysis did not consider interactions or any higher order effects. So instead we report the χ^2 statistic and use it as a reference measurement of model assessment.

For the dconst parameterization, the final model selected using BIC was sparser (6/11 non-zero coefficients) than the final model using AIC (8/11 non-zero coefficients), which is in line with the results of the simulation study. When using BIC to select the tuning parameter, the predictors that were shrunk to zero were the same ones with p-values less than 0.05 in the unpenalized model, with the exception of floral symmetry. The parameter estimates for this same set of predictors were generally close to the unpenalized MLEs, but the discrepancy

between the unpenalized and penalized refit using AIC was higher than between the unpenalized and penalized refit using BIC. In terms of goodness of fit, the unpenalized model has the lowest χ^2 statistic; however, the refit adaptive lasso models had the lowest AIC and BIC values. For the no dispersion parameterization the results were rather inconclusive since the adaptive lasso only shrunk one predictor to zero, and, in general, there was little difference between the unpenalized and penalized estimates and goodness-of-fit measurements.

Most trends from the dconst model fits persisted for the dfunc ones. For the dfunc parameterization, the final model selected using BIC was sparser (10/15 non-zero coefficients) than the final model using AIC (12/15 non-zero coefficients). When using BIC to select the tuning parameter, the predictors that were shrunk to zero were the same ones with p-values less than 0.05 in the unpenalized model, with the exception of Corolla Shape Rotate and Actinomorphic. The parameter estimates for these same set of predictors were generally close to the unpenalized MLEs, but a bit lower for both the penalized refits using AIC and BIC. In terms of goodness of fit, the unpenalized model has the lowest log-likelihood, which is expected since it contains all the predictors. However, the refit adaptive lasso models had the lowest AIC, BIC and χ^2 statistic.

A comparison of the model fit statistics across parameterizations suggests that the dfunc and dconst models provide a better fit than the none model. For example, the AIC, BIC, and the χ^2 statistic was always lowest for the dispersion models, with the lowest being achieved with the dfunc model. These discrepancies suggest that the data do exhibit overdispersion and that the dfunc model may be a more appropriate choice.

The plant traits that were estimated to have non-zero coefficients using the adaptive lasso (dfunc parameterization) can be partially explained by pollination syndromes. Pollination syndromes are defined as evolved suites of floral traits (e.g., colour, shape, size, etc.) among flowers pollinated by a particular functional group (e.g., bees, beetles, flies, etc.). For example, moth pollinated flowers tend to be large and white with long tubular corollas. Accordingly, a changing insect species composition may be expected to lead to different suites of floral traits, depending on the dominant flower-visiting guilds and the floral characteristics that appeal to them. For the Terceira Island data set, flies are the dominant guild (Picanço et al., unpublished) and the floral traits associated with fly pollination, namely, white corollas (Arnold et al., 2009), symmetric flowers (actinomorphic), and regular (wheel-shaped)

corollas, had positive non-zero coefficients. Hence, there is a higher log-odds of a pollinator species interacting with a plant species that possesses these traits.

Additionally, plant morphology (plants with inflorescences) and flower type (dioecious plants) also had non-zero coefficients. This result is not surprising since pollinators are attracted to inflorescences, which tend to have a greater nectar/pollen reward, and dioecious plants are generally preferred over monoecious plants (Zito et al., 2016). On the other hand, introduced plant species was estimated to have a negative non-zero coefficient. Hence there is a lower log-odds of selecting an introduced plant over an native/endemic plant. While the geographical origin of the plant species (i.e., native/endemic or introduced) is not a species trait per se, from an ecological adaptation perspective insects may be driven to native/endemic plants because they are familiar or well known. In fact, although there were more than two times the number of introduced (exotic) species than native/endemic species, the proportions of total number of visits to these latter two groups of plants were 54% and 46%, respectively. This result provides some evidence in favour of the notion of ecological adaption.

Finally, two out of the four available insect species traits used to model dispersion (δ_g) had non-zero coefficients. Insect behaviour (social) had a positive coefficient, while insect trophic (herbivorous) had a negative coefficient. These dispersion traits provide information on the impact that these insect traits have on the number of times each plant species is visited. Since social flies and non-herbivorous dominate this island network (Picanço et al., 2017), insect species with these traits have a higher impact on the number of visits made to each of the plant species.

Table 5.7: Unpenalized and Penalized Model Fits for the Terceira Island Network*

	dfunc			dconst			none	
Predictor	MLE	BIC refit	AIC refit	MLE	BIC refit	AIC refit	MLE	AIC/BIC refit
Monoecious	-0.584	-0.63	-0.633	-0.585	-0.605	-0.52	-1.155	-1.156
Corolla White	0.646	0.551	0.44	0.504	0.387	0.54	0.544	0.551
Corolla Yellow	0.152	0	0.199	0.15	0	0.28	0.678	0.687
Flower Size Medium	0.485	0.146	0.09	0.127	0	0	-0.452	-0.461
Flower Size Large	0.758	0.315	0.217	0.236	0	0.17	0.233	0.23
Corolla Shape Regular	0.312	0.458	0.316	0.389	0.53	0.4	0.543	0.545
Actinomorphic	0.332	0.17	0.425	0.433	0.282	0.54	1.235	1.239
Inflorescence	0.722	0.646	0.464	0.581	0.518	0.57	0.818	0.813
Plant Type Woody	0.118	0	0	0.049	0	0	0.02	0.168
Perreniel	0.083	0	0	0.038	0	0	0.167	0
Introduced Plant Species	-0.311	-0.361	-0.398	-0.333	-0.416	-0.27	-0.452	-0.45
Dispersion:								
Introduced Insect Species	0.244	0	0.092	-	-	-	-	-
Social	1.76	0.885	1.926	-	-	-	-	-
Herbivorous	-0.923	-0.753	-0.898	-	-	-	-	-
Median Body Length	-0.023	0	0	-	-	-	-	-
Intercept	-4.551	-3.353	-4.284	-2.996	-2.574	-3.086	-	-
Log-Likelihood	-1605.993	-1601.696	-1600.21	-1617.225	-1619.182	-1618.238	-7172.473	-7172.505
BIC	3334.832	3287.849	3300.232	3326.584	3292.109	3305.577	14429.403	14421.7886
AIC	3243.986	3225.392	3226.42	3258.45	3252.364	3254.48	14366.946	14365.01
Non-zero Coefficients	15	10	12	11	6	8	11	10
χ^2 Statistic	2397.258	2266.115	2313.309	2344.507	2388.464	2364.08	14320.51	14317.41

* dfunc - δ as a function of pollinator-specific covariates; dconst - δ as a constant; none - no dispersion; MLE - unpenalized maximum likelihood estimator; BIC refit - unpenalized refit estimators for adaptive lasso ($\tilde{\gamma} = 3$) using BIC; AIC refit - unpenalized refit estimators for adaptive lasso ($\tilde{\gamma} = 3$) using AIC; AIC/BIC refit - unpenalized refit estimators for adaptive lasso ($\tilde{\gamma} = 3$) using AIC or BIC (variable selection resulted in the same selected predictors).

5.5 Discussion and Conclusions

In this paper we developed regularized regression for three parameterizations of grouped DM model using lasso-type methods and considered two information criteria, AIC and BIC, for tuning parameter selection. We provide a stable implementation of the accelerated proximal gradient method FISTA, where the stepsize was adapted globally and on a dimension-specific basis, to obtain the penalized MLEs. We compared the performance of the lasso and adaptive lasso for both the fixed and diverging dimension scenarios and found that the adaptive lasso using BIC performed best with respect to model selection and parameter estimation. We also applied our lasso approach to perform variable selection on an empirical plant-pollinator network and showed that it is a viable method in practice.

Although AIC and BIC were developed primarily for unpenalized likelihoods, in most cases, their use within a penalized framework has successfully facilitated the selection of the tuning parameter needed to achieve simultaneous variable selection and parameter estimation. More specifically, BIC is known to be asymptotically selection consistent, and we found that this also seems to be the case, at least empirically, for the penalized grouped DM model. This finding is not surprising since BIC does account for model fit and model complexity, which makes it a reasonable choice for either variable selection or prediction (Hastie et al., 2009). On the other hand, AIC is not asymptotically consistent and tends to choose models that are more complex, which was also demonstrated empirically for our penalized DM models. However, the results of the simulation study did show that AIC was competitive for parameter estimation which can be attributed to the fact that it is loss efficient. Note also that AIC is sometimes referred to as a predictive method in a penalized framework and is used when prediction accuracy is of main interest.

In general, the lasso-type methods for the none and dconst parameterizations performed better than the rconst parameterization. In fact, in the diverging K_N scenario, the percent of correct models was extremely low for the rconst models, even for the largest network sizes with low dispersion. Also, some of the highest mean MSEs were observed for the rconst models. This might suggest that AIC and BIC are inconsistent for both model selection and parameter estimation for this parameterization of the grouped DM model. However, these results are in line with some of our previous studies which showed none and dconst to be the more stable parameterizations (Crea et al., 2016). Alternatively, potential improvements

could be attained with respect to model selection consistency by using an adjusted BIC. For example, Hui et al. (2015) suggest the extended regularization information criteria (ERIC), a modification of the BIC, where the model complexity term is also a function of the tuning parameter $\tilde{\lambda}$.

Although accelerated proximal gradient methods were developed for composite functions where the first term is convex and smooth (i.e., continuously differentiable with Lipschitz continuous gradient) and the second term is non-smooth, it seemed to work well for the penalized DM likelihood. In general, optimizing the unpenalized DM likelihood is nontrivial because it is non-convex and can have non-smooth regions. Consequently, establishing global convergence properties using FISTA for the penalized DM likelihood is difficult since the convergence rates for FISTA rely on convexity. However, we did not encounter any convergence issues and the algorithm did converge to a local minimum of the objective function, which was also noted by others who used FISTA for the penalized DM regression model (Zhang et al., 2016 and Wang and Zhao, 2017). Scientists are beginning to tackle non-convex and non-smooth problems, particularly for these large scale problems, but more work is needed to establish global convergence rates.

Our implementation of the grouped DM lasso was motivated from an ecological context. Ecologists and evolutionary biologists seek to understand how species traits and linkage rules influence the interactions in mutualistic networks. Recent ecological studies involve collecting interaction frequency and detailed trait data and it is for this reason that model selection is necessary for analyzing ecological networks. Burnham and Anderson (2009) advocate the use of AIC over the BIC because they feel that all covariates included in an ecological regression model have non-zero effects, albeit some may be very small. As such, identifying the covariates that improve prediction, rather than identifying the "true" dimension of a model, should be the objective of model selection. However, it is unlikely that every trait or combination of traits is an important predictor for a given network, particularly since we are analyzing a sampled network.

The data contained in a sampled network has a fixed amount of information and so the gain in accuracy from having a more complex model can lead to a loss in parameter precision (Bolker, 2008). In the empirical analysis, the adaptive lasso models using BIC performed best and the dconst and dfunc models seemed to provide the best fit to the data, even though

some important predictors may have been missing from the set of covariates. Nonetheless, the relevant predictors chosen by the adaptive lasso for the dfunc parameterization could be explained from an ecological perspective.

We also refit the final models to obtain the unpenalized MLEs and associated standard errors as suggested by Efron et al. (2004). Alternatively, methods have been proposed to calculate standard errors directly from the penalized estimates. The bootstrap method can be used to calculate standard errors, but is computationally expensive. Additionally, lasso estimates are known to be biased, hence the standard errors can only account for the variance of the estimates and not the bias. For the adaptive lasso, Zou (2006) provides a sandwich estimator for the standard errors using a locally quadratic approximation within the linear and (univariate) nonlinear regression context. Zeng et al. (2014) adopt this approach for the adaptive lasso of the zero-inflated Poisson and negative binomial models. Nonetheless, there is no real consensus in the literature about the validity of these standard errors and theory is still being developed.

In future research, an evaluation of the robustness to misspecification of dispersion structure for the penalized δ and ρ parameterizations should be considered since, in practice, we do not know the true underlying dispersion structure of the data (as demonstrated in the empirical analysis). Also, we used our lasso approach to fit a DM model where the dispersion parameter is modelled as a function of group-specific covariates, but a more comprehensive study of this parameterization is needed to truly understand the performance of the proposed lasso approaches. There are two approaches that can be considered: estimate all effects simultaneously or use a two-stage process. We chose to estimate both types of covariates simultaneously, similar to Zhao et al. (2014) who consider the lasso, SCAD, and MCP for the varying dispersion beta regression model and estimate both the fixed and random effects simultaneously. On the other hand, Pan and Shang (2018) implement an adaptive lasso for the linear mixed model using a two-stage approach, where the random effects are selected via a penalized restricted profile log-likelihood and then the fixed effects are selected using a penalized log-likelihood.

Finally, although the scope of this paper is within the moderate parameter dimension setting, we have set up the framework to be easily extended to the high dimensional setting, i.e., $K >> N$. A great deal of work has been devoted to the high dimensional case, particularly

within the linear and nonlinear regression frameworks (e.g., Wang, Li, and Leng 2009; Wang and Zhu 2011; Fan and Tang 2012; Hui et al., 2015). This will require further study, but since the BIC performed best for this implementation of the DM model, continuing to use BIC (or a modified version tailored specifically for the high dimensional case) to select the tuning parameter seems to be the most logical path forward.

Acknowledgements

The authors gratefully acknowledge Ana Picanço for providing the Terceira Island data set and for her insightful and constructive comments on the interpretation of the analysis results.

Chapter 6

Conclusions And Future Work

In this thesis, we studied the optimization and regularization of the grouped DM regression model with applications to ecological networks. Though this model has several applications in various fields, our application of grouped DM regression to plant-pollinator networks was motivated from an econometric consumer-choice framework, and is a novel approach to analyzing plant-pollinator networks.

In Chapter 3, we tested whether this econometric approach is applicable within an ecological setting. We showed that the DM model was able to generate networks that exhibited a diversity of topological features that resembled those in observed networks and that analyzing these networks using the DM model provided useful insights into the mechanisms driving network structure. The DM model not only identified the most relevant predictors of interaction, it was also able to quantify their relative contributions. Further, since the nature of ecological data sets differs from that of econometric data sets, particularly with respect to the magnitude/sparsity of the observed counts and the types of covariates, an exploration of the parameter space was imperative in evaluating applicability of this model to ecological data. By determining the most likely ranges of the model parameters, we were able to study and evaluate the performance of the DM model within these parameter boundaries.

In Chapter 4 we studied the behaviour of the grouped DM likelihood under various parameterizations and further evaluated the robustness to misspecification of dispersion structure. Guimarães and Lindrooth (2007) introduced this (econometric) model, but did not investigate these topics, which are essential to the use of the DM model by ecologists in practice. The DM distribution is not part of the natural exponential family and, in general is non-

convex. The optimization of this likelihood is nontrivial. Standard methods for optimizing convex functions, such as MNR and BFGS, can become unstable, particularly when the algorithms have poor starting values, which is often the case when fitting a regression model.

Since our model has an analogous Poisson-Gamma representation, we could use the Poisson estimates as starting values for the regression coefficients, which proved to be a good starting point for these parameters. However, best "guesses" for the dispersion parameters was not as straightforward. NM was more robust to these poor starting values for the dconst, dfunc and rfunc parameterizations. Further, NM performed best with respect to convergence including Hessian stability, parameter estimation, and coverage. However, for the rconst parameterization, our starting value of 0.1 seemed to be an adequate starting value and BFGS performed best for this parameterization. Note that all optimizations methods performed equally for the no dispersion case, which is not surprising since the multinomial likelihood is convex.

We also provide a stable implementation in `R` for fitting the grouped DM regression model, where the user can choose from the abovementioned optimization methods. Our `R` program, Stata and LIMDEP are the only programs that can fit this implementation of the grouped DM regression model; however, the latter ones use a MNR for the optimization, and based on our experience, is prone to convergence issues.

Our evaluation of the robustness to dispersion structure revealed that, in general, the no dispersion and δ parameterizations were more robust and stable compared to the ρ parameterization. When the dispersion structure was misspecified and a ρ parameterization was used to fit the data, we observed more convergence and Hessian instability issues, higher bias and variability in regression coefficients, and poorer model fits. However, we also noted that the χ^2 approximation can be poor for the DM model and can lead to inflated Type I errors, so we suggest that information criteria, such as AIC or BIC, be used instead for model comparisons.

In Chapter 5 we developed a regularization procedure for DM regression models. Through the analysis of several plant-pollinator networks, it became clear that the lack of an automatic or systematic variable selection method for the DM model was inconvenient, impractical, and computationally expensive. However, field ecologists are continually collecting more detailed

species data with a goal of better understanding the mechanisms that drive pollination, and discerning which covariates, if any, are important indicators of interaction. Consequently, our regularized regression for the DM model focused on variable selection rather than prediction accuracy.

Using lasso-type methods for variable selection, we implemented FISTA for optimizing the penalized log-likelihood, and studied both AIC and BIC for selecting the optimal tuning parameter. For a moderate parameter dimension setting, we evaluated the performance of our proposed lasso approach for both the fixed K and diverging K_N scenarios. The results of the simulation study suggest that the adaptive lasso ($\tilde{\gamma} = 3$) performed the best for the grouped DM model, that BIC outperformed AIC with respect to model selection and parameter estimation, and that these methods worked best for the none and δ parameterizations. We applied our lasso approach to an empirical plant-pollinator network and show that it is a viable method for variable selection in practice. Although, the DM likelihood is generally non-convex and can be non-smooth, FISTA at least converged to a local minimum of the objective values. A stable implementation of the regularized grouped DM regression model in `R`, though its performance is yet to be well-studied.

In closing, this thesis provides ecologists with robust statistical methods to conduct community-level analyses for studying the mechanisms that drive mutualistic networks. It also provides useful insights into the behaviour of the DM likelihood and the model's robustness to misspecification of dispersion structure. Again, from a practical perspective, such an evaluation facilitates the selection of the appropriate combination of parameterization and optimization method when analyzing empirical data sets. Finally, we have set up the framework and foundation for the use of regularization for the grouped DM regression model, which simultaneously performs variable selection and parameter estimation.

6.1 Future Work

This thesis has explored several aspects of the grouped DM regression model and its application to ecological problems. However, there are several additional areas of future work that will continue to advance this model both from an ecological and a statistical perspective. The following are some potential directions for future work from an ecological perspective.

The presence of zero counts in a network can be attributed to sampling effects, but also other ecological and evolutionary factors, such as, forbidden links. Currently, the DM model can accommodate these structural zeroes, but because these networks are based on a sample, the presence of a zero count in the interaction matrix does not necessarily imply an interaction cannot occur. It simply means that an interaction was not observed, i.e., missing interaction vs. forbidden interaction. An extension to DM regression that accommodates zero-inflated adjustments may lead to better parameter estimates of the regression coefficients and overall model fits. These extensions, which involve creating a new category for representing structural zeroes, exist for the unpenalized Poisson and negative binomial models (Moghimbeigi et al., 2008; Lee et al., 2006) and the penalized ones (Zeng et al., 2014).

Ecological networks are known to be built using visitation counts sampled from finite populations. Such sampling effects (e.g. observation error, sampling effort) typically create a discrepancy between the true and observed network structure thereby making these networks appear heterogeneous and difficult to study. Collecting interaction network data is both time-consuming and expensive; therefore, networks are often under-sampled (Vázquez et al. 2009a). Although there is a surge in the study of mutualistic network analysis, the effects of sampling completeness and sampling effects are not well understood. It is clear that the observed networks are typically under-sampled, but the amount of sampling effort needed to adjust for under-coverage is costly. Consequently, network analyses can often result in biased estimates of network statistics which may lead to skewed or misleading network patterns. Also, these sampling biases can also affect the parameter estimation used in other quantitative network analyses, such as DM regression, which greatly affect the inferences made about the mechanisms driving network structure. One approach to account for sampling effort is to adjust the actual observed counts for the sampling weights and have these adjusted counts propagate through the likelihood or weight the multinomial probabilities directly using the weighted sampling distribution and re-write the DM likelihood in terms of the weighted probabilities for parameter estimation.

From a statistical perspective, there are many potential directions for future work with respect to optimization and regularization. With respect to the optimization of the unpenalized DM regression model, it is clear that the Poisson estimates are not good starting points for the dispersion parameters. Since the DM likelihood is not part of the exponential family, there are potentially many local optima and no way of knowing which is the global.

One way to deal with this is by taking a multi-start approach, which involves running the optimization algorithm from randomly sampled starting values, then selecting the best as the proposed globally optimal solution. Alternatively, using accelerated gradients may be a viable approach since the accelerated proximal gradients performed well for the penalized DM likelihood. Finally, consideration of other methods specifically tailored to non-convex and possibly non-smooth optimization problems could be explored, for example, the adaptive scaled BFGS or the double parameter scaled BFGS algorithm, which is globally convergent in very general conditions without the convexity assumption (Andrei, 2018a; Andrei, 2018b). An evaluation of these new methods can also accommodate larger scaled problems since the NM can degenerate in medium to high dimensional settings.

Although we have now established and tested a regularized framework using lasso-type penalties for the grouped DM model, there are several other penalties that can be easily implemented depending on the model fitting goals. Examples include ridge (if only shrinkage, not sparsity is the goal), elastic net (if one suspects multicollinearity among the variables), group or sparse group lasso (if variables are multilevel categories), or non-convex penalties, such as SCAD and MCP. Finally, other tuning parameter selection criteria, such as CV can be evaluated, particularly if prediction accuracy becomes the main interest, or a modified BIC to improve the model selection for the ρ parameterization.

Bibliography

Adhikari, S., Lecci, F., Becker, J. T., Junker, B. W., Kuller, L. H., Lopez, O. L. and Tibshirani, R. J. (2015). High-Dimensional Longitudinal Classification with the Multinomial Fused Lasso, *arXiv preprint arXiv:1501.07518* .

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *in* B. N. Petrov and F. Csaki (eds), *Second International Symposium on Information Theory*, Akadémiai Kiado, Budapest, pp. 267–281.

Alarcón, R. (2010). Congruence between visitation and pollen-transport networks in a california plant–pollinator community, *Oikos* **119**(1): 35–44.

Allesina, S., Alonso, D. and Pascual, M. (2008). A general model for food web structure, *Science* **320**(5876): 658–661.

Andrei, N. (2018a). A double parameter scaled BFGS method for unconstrained optimization, *Journal of Computational and Applied Mathematics* **332**: 26–44.

Andrei, N. (2018b). An adaptive scaled BFGS method for unconstrained optimization, *Numerical Algorithms* **77**(2): 413–432.

Armijo, L. (1966). Minimization of functions having Lipschitz continuous first partial derivatives, *Pacific Journal of Mathematics* **16**(1): 1–3.

Arnold, S. E., Savolainen, V. and Chittka, L. (2009). Flower colours along an alpine altitude gradient, seen through the eyes of fly and bee pollinators, *Arthropod-Plant Interactions* **3**(1): 27–43.

Azores Biodiversity Database (2016). Base de dados da biodiversidade dos açores. **URL:** *http://www.atlantis.angra.uac.pt/*

Barnard, P. C. (2011). *The Royal Entomological Society Book of British Insects*, John Wiley & Sons.

Bartomeus, I. (2013). Understanding linkage rules in plant-pollinator networks by using hierarchical models that incorporate pollinator detectability and plant traits, *PloS One* **8**(7): e69200.

Bartomeus, I., Vilà, M. and Santamaría, L. (2008). Contrasting effects of invasive plants in plant–pollinator networks, *Oecologia* **155**(4): 761–770.

Bascompte, J. and Jordano, P. (2007). Plant-Animal Mutualistic Networks: The Architecture of Biodiversity, *Annual Review of Ecology Evolution and Systematics* **38**(1): 567–593.

Bawa, K. S. (1980). Evolution of dioecy in flowering plants, *Annual Review of Ecology and Systematics* **11**: 15–39.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences* **2**(1): 183–202.

Bee, Wasp and Ants Recording Society (2016).
URL: *http://www.bwars.com/*

Bersier, L.-F., Banašek-Richter, C. and Cattin, M.-F. (2002). Quantitative descriptors of food-web matrices, *Ecology* **83**(9): 2394–2407.

Blüthgen, N. (2010). Why network analysis is often disconnected from community ecology: a critique and an ecologist's guide, *Basic and Applied Ecology* **11**(3): 185–195.

Blüthgen, N., Menzel, F. and Blüthgen, N. (2006). Measuring specialization in species interaction networks, *BMC Ecology* **6**(1): 9.

Bolker, B. M. (2008). *Ecological Models and Data in R*, Princeton University Press, Princeton and Oxford.

Borges, P., Costa, A., Cunha, R., Gabriel, R., Gonalves, V., Martins, A., Melo, I., Parente, M., Raposeiro, P., Rodrigues, P., Santos, R., Silva, L., Vieira, P. and Vieira, V. (2010). *A List of the Terrestrial and Marine Biota from the Azores*, Princípia, Oeiras.

Bosch, J., Retana, J. and Cerdá, X. (1997). Flowering phenology, floral traits and pollinator composition in a herbaceous Mediterranean plant community, *Oecologia* **109**(4): 583–591.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *The Annals of Statistics* **24**(6): 2350–2383.

Brownstone, D. and Train, K. (1998). Forecasting new product penetration with flexible substitution patterns, *Journal of Econometrics* **89**(1): 109–129.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms: 2. The new algorithm, *IMA Journal of Applied Mathematics* **6**(3): 222–231.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection, *Sociological Methods & Research* **33**(2): 261–304.

Byrd, R. H. and Nocedal, J. (1989). A Tool for the Analysis of Quasi-Newton Methods with Application to Unconstrained Minimization, *SIAM Journal on Numerical Analysis* **26**(3): 727–739.

Byrd, R. H., Nocedal, J. and Yuan, Y.-X. (1987). Global Convergence of a Class of Quasi-Newton Methods on Convex Problems, *SIAM Journal on Numerical Analysis* **24**(5): 1171–1190.

Chacoff, N. P., Vazquez, D. P., Lomascolo, S. B., Stevani, E. L., Dorado, J. and Padron, B. (2012). Evaluating sampling completeness in a desert plant–pollinator network, *Journal of Animal Ecology* **81**(1): 190–200.

Chen, J. and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis, *The Annals of Applied Statistics* **7**(1).

Chittka, L. and Thomson, J. D. (2005). *Cognitive Ecology of Pollination: Animal Behaviour and Floral Evolution*, Cambridge University Press.

Crea, C. (2014). GCL and DM regression, `https://github.-com/ccrea/GCL_DM_Regression`.

Crea, C. (2017). Varibale Selection for Grouped DM regression, `https://github.-com/ccrea/VarSelect_DM_Regression`.

Crea, C., Ali, R. A. and Rader, R. (2016). A new model for ecological networks using species-level traits, *Methods in Ecology and Evolution* **7**(2): 232–241.

Croissant, Y. (2013). *mlogit: multinomial logit model*. R package version 0.2-4.
URL: *https://CRAN.R-project.org/package=mlogit*

Davidson, W. C. (1991). Variable Metric Method for Minimisation, *SIAM Journal on Optimization* **1**(1): 1–17.

de Valpine, P. and Harmon-Threatt, A. N. (2013). General models for resource use or other compositional count data using the Dirichlet-multinomial distribution, *Ecology* **94**(12): 2678–2687.

Dennis, J. E. and Moré, J. J. (1974). A Characterization of Superlinear Convergence and Its Application to Quasi-Newton Methods, *Mathematics of Computation* **28**(126): 549–560.

Dennis, Jr, J. E. and Moré, J. J. (1977). Quasi-Newton Methods, Motivation and Theory, *SIAM Review* **19**(1): 46–89.

Devroye, L. (1986). *Non-uniform random variate generation*, Springer-Varlag, New York.

Dias, E. (1996). Vegetação natural dos açores. ecologia e sintaxonomia das florestas naturais, *Departamento de Ciências Agrárias Universidade dos Açores Angra do Heroísmo, Spain Ph. D thesis* .

Díaz, S., Quétier, F., Cáceres, D. M., Trainor, S. F., Pérez-Harguindeguy, N., Bret-Harte, M. S., Finegan, B., Peña-Claros, M. and Poorter, L. (2011). Linking functional diversity and social actor strategies in a framework for interdisciplinary analysis of nature's benefits to society, *Proceedings of the National Academy of Sciences* **108**(3): 895–902.

Dixon, L. (1972). Variable metric algorithms: Necessary and sufficient conditions for identical behavior of nonquadratic functions, *Journal of Optimization Theory and Applications* **10**(1): 34–40.

Dormann, C. F., Fründ, J., Blüthgen, N. and Gruber, B. (2009). Indices, graphs and null models: analyzing bipartite ecological networks.

Dupont, Y. L. and Olesen, J. M. (2009). Ecological modules and roles of species in heathland plant–insect flower visitor networks, *Journal of Animal Ecology* **78**(2): 346–353.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004). Least angle regression, *The Annals of Statistics* **32**(2): 407–499.

Eklöf, A., Jacob, U., Kopp, J., Bosch, J., Castro-Urgal, R., Chacoff, N. P., Dalsgaard, B., Sassi, C., Galetti, M., Guimarães, P. R. et al. (2013). The dimensionality of ecological networks, *Ecology Letters* **16**(5): 577–583.

Elff, M. (2014). *mclogit: Mixed Conditional Logit*. R package version 0.3-1.
URL: *https://CRAN.R-project.org/package=mclogit*

Evans, A. V. (2008). *Field Guide to Insects and Spiders of North America*, Sterling, New York.

Faegri, K. and Van Der Pijl, L. (2013). *Principles of Pollination Ecology*, Elsevier.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**(456): 1348–1360.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(3): 531–552.

Fletcher, R. (1970). A new approach to variable metric algorithms, *The Computer Journal* **13**(3): 317–322.

Fletcher, R. and Powell, M. J. (1963). A rapidly convergent descent method for minimization, *The Computer Journal* **6**(2): 163–168.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software* **33**(1): 1–22.
URL: *http://www.jstatsoft.org/v33/i01/*

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means, *Mathematics of Computation* **24**(109): 23–26.

González-Castro, A., Yang, S., Nogales, M. and Carlo, T. A. (2015). Relative importance of phenotypic trait matching and species' abundances in determining plant–avian seed dispersal interactions in a small insular community, *AoB Plants* **7**: plv017.

Gravel, D., Poisot, T., Albouy, C., Velez, L. and Mouillot, D. (2013). Inferring food web structure from predator–prey body size relationships, *Methods in Ecology and Evolution* **4**(11): 1083–1090.

Greenleaf, S. S., Williams, N. M., Winfree, R. and Kremen, C. (2007). Bee foraging ranges and their relationship to body size, *Oecologia* **153**(3): 589–596.

Griewank, A. (1991). The global convergence of partitioned BFGS on problems with convex decompositions and Lipschitzian gradients, *Mathematical Programming* **50**(1): 141–175.

Guimarães, P., Galdini Raimundo, R. and Cagnolo, L. (2011). *Interaction Web Database*, National Center for Ecological Analysis and Synthesis, University of Califoria, Santa Barbara, USA.
URL: *http://www.nceas.ucsb.edu/interactionweb/index.html*

Guimarães, P. and Lindrooth, R. C. (2007). Controlling for overdispersion in grouped conditional logit models: A computationally simple application of Dirichlet-multinomial regression, *The Econometrics Journal* **10**(2): 439–452.

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package, *Journal of Statistical Software* **33**(2): 1–22.
URL: *http://www.jstatsoft.org/v33/i02/*

Hasan, A., Zhiyu, W. and Mahani, A. S. (2015). *mnlogit: Multinomial Logit Model*. R package version 1.2.4.
URL: *https://CRAN.R-project.org/package=mnlogit*

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Hausman, J., Hall, B. H. and Griliches, Z. (1984). Econometric Models for Count Data with an Application to the Patents-R & D Relationship, *Econometrica* **52**(4): 909–938.

Heinrich, B. (1979). Majoring and minoring by foraging bumblebees, Bombus vagans: an experimental analysis, *Ecology* **60**(2): 245–255.

Hesterberg, T., Choi, N. H., Meier, L., Fraley, C. et al. (2008). Least angle and $l1$ penalized regression: A review, *Statistics Surveys* **2**: 61–93.

Hudson, L. N., Newbold, T., Contu, S., Hill, S. L., Lysenko, I., De Palma, A., Phillips, H. R., Senior, R. A., Bennett, D. J., Booth, H. et al. (2014). The predicts database: a global database of how local terrestrial biodiversity responds to human impacts, *Ecology and Evolution* **4**(24): 4701–4735.

Hui, F. K., Warton, D. I. and Foster, S. D. (2015). Tuning parameter selection for the adaptive lasso using ERIC, *Journal of the American Statistical Association* **110**(509): 262–269.

James, J. (2016). An MM Algorithm for General Mixed Multinomial Logit Models, *Journal of Applied Econometrics* .

Jentzsch, M. (2014). New data on hoverflies (Diptera: Syrphidae) from the Azorean island Pico (Portugal), *Arquipélago - Life and Marine Science* **31**.

Jordán, K. (1965). *Calculus of finite differences*, Vol. 33, American Mathematical Society.

Jordano, P. (1987). Patterns of mutualistic interactions in pollination and seed dispersal: connectance, dependence asymmetries, and coevolution, *The American Naturalist* **129**(5): 657–677.

Jordano, P. and Bascompte, J. and Olesen, J. M. (2006). The ecological consequences of complex topology and nested structure in pollination webs, *in* N. M. Waser and J. Ollerton (eds), *Plant-Pollinator Interactions: From Specialization to Generalization*, University Of Chicago Press, EEUU, chapter 8, pp. 173–199.

Jordano, P., Bascompte, J. and Olesen, J. M. (2003). Invariant properties in coevolutionary networks of plant–animal interactions, *Ecology Letters* **6**(1): 69–81.

Junker, R. R., Blüthgen, N., Brehm, T., Binkenstein, J., Paulus, J., Martin Schaefer, H. and Stang, M. (2013). Specialization on traits as basis for the niche-breadth of flower visitors and as structuring mechanism of ecological networks, *Functional Ecology* **27**(2): 329–341.

Klami, A., Tripathi, A., Sirola, J., Väre, L. and Roulland, F. (2015). Latent feature regression for multivariate count data, *Artificial Intelligence and Statistics*, pp. 462–470.

Kolda, T. G., Lewis, R. M. and Torczon, V. (2003). Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods, *SIAM Review* **45**(3): 385–482.

Kremen, C. (2005). Managing ecosystem services: what do we need to know about their ecology?, *Ecology Letters* **8**(5): 468–479.

Lafontaine, J. and Troubridge, J. (1998). Moths and butterflies (lepidoptera), *Assessment of species diversity in the Montane Cordillera Ecozone. Eco-logical Monitoring and Assessment Network, Burlington, ON. http://www. naturewatch. ca/eman/reports/publications/99_montane/lepidopt/intro. html [accessed 22 July 2009]* .

Larsen, T. H., Williams, N. M. and Kremen, C. (2005). Extinction order and altered community structure rapidly disrupt ecosystem functioning, *Ecology Letters* **8**(5): 538–547.

Lavorel, S. and Grigulis, K. (2012). How fundamental plant functional trait relationships scale-up to trade-offs and synergies in ecosystem services, *Journal of Ecology* **100**(1): 128–140.

Lee, A. H., Wang, K., Scott, J. A., Yau, K. K. and McLachlan, G. J. (2006). Multi-level zero-inflated poisson regression modelling of correlated count data with excess zeros, *Statistical Methods in Medical Research* **15**(1): 47–61.

Li, D.-H. and Fukushima, M. (2001). On the Global Convergence of the BFGS Method for Nonconvex Unconstrained Optimization Problems, *SIAM Journal on Optimization* **11**(4): 1054–1064.

Maddala, G. S. (1986). *Limited-dependent and qualitative variables in econometrics*, number 3, Cambridge university press.

Marsh, P. M. (1994). Hymenoptera of the World: An Identification Guide to Families, *American Entomologist* **40**(2): 115–116.

Mauerer, I., Pößnecker, W., Thurner, P. W. and Tutz, G. (2015). Modeling electoral choices in multiparty systems with high-dimensional data: A regularized selection of parameters using the lasso approach, *Journal of Choice Modelling* **16**: 23–42.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, Vol. 37, CRC press.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior, *in* P. Zarembka (ed.), *Frontiers in econometrics*, Academic Press, New York, pp. 105–142.

McFadden, D. and Train, K. (2000). Mixed MNL models for discrete response, *Journal of Applied Econometrics* **15**: 447–470.

Meier, L., Van De Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1): 53–71.

Mello, M. A. R., Marquitti, F. M. D., Guimarães, P. R., Kalko, E. K. V., Jordano, P. and de Aguiar, M. A. M. (2011). The modularity of seed dispersal: differences in structure and robustness between bat–and bird–fruit networks, *Oecologia* **167**(1): 131.

Milne-Thomson, L. M. (2000). *The Calculus of Finite Differences*, American Mathematical Society, Providence, Rhode Island.

Mimno, D. and McCallum, A. (2012). Topic models conditioned on arbitrary features with Dirichlet-multinomial regression, *arXiv preprint arXiv:1206.3278* .

Moghimbeigi, A., Eshraghian, M. R., Mohammad, K. and Mcardle, B. (2008). Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros, *Journal of Applied Statistics* **35**(10): 1193–1202.

Montoya, J. M., Pimm, S. L. and Solé, R. V. (2006). Ecological networks and their fragility, *Nature* **442**(7100): 259–264.

Montoya, J. M. and Solé, R. V. (2002). Small world patterns in food webs, *Journal of Theoretical Biology* **214**(3): 405–412.

Mosimann, J. E. (1962). On the Compound Multinomial Distribution, the Multivariate β-Distribution, and Correlations Among Proportions, *Biometrika* **49**(1/2): 65–82.

Nash, J. C. (1990). *Compact numerical methods for computers: linear algebra and function minimisation*, CRC Press.

Nash, J. C. (2014). On best practice optimization methods in R, *Journal of Statistical Software* **60**(2): 1–14.

Nash, J. C. and Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: optimx for R, *Journal of Statistical Software* **43**(9): 1–14.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization, *The computer journal* **7**(4): 308–313.

Nesterov, Y. (2007). Smoothing technique and its applications in semidefinite optimization, *Mathematical Programming* **110**(2): 245–259.

Nielsen, A. and Bascompte, J. (2007). Ecological networks, nestedness and sampling effort, *Journal of Ecology* **95**(5): 1134–1141.

of London, R. E. S., Coe, R., Freeman, P. and Mattingly, P. (1950). Handbooks for the Identification of British Insects, Royal Entomological Society of London.

of London, R. E. S. and Spencer, K. (1972). Handbooks for the Identification of British Insects, Royal Entomological Society of London.

Olesen, J. M., Bascompte, J., Dupont, Y. L., Elberling, H., Rasmussen, C. and Jordano, P. (2011). Missing and forbidden links in mutualistic networks, *Proceedings of the Royal Society of London B: Biological Sciences* **278**(1706): 725–732.

Olito, C. and Fox, J. W. (2015). Species traits and abundances predict metrics of plant–pollinator network structure, but not pairwise interactions, *Oikos* **124**(4): 428–436.

Pan, J. and Huang, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function, *Statistics and Computing* **24**(5): 725–738.

Pan, J. and Shang, J. (2018). Adaptive LASSO for linear mixed model selection via profile log-likelihood, *Communications in Statistics-Theory and Methods* **47**(8): 1882–1900.

Pendleton, T. and Pendleton, D. (1997). The website dedicated to Nottinghamshire's invertebrate fauna.
URL: *http://www.eakringbirds.com/*

Picanço, A., Rigal, F., Matthews, T., Cardoso, P. and Borges, P. (2017). Island flower-visiting insect communities are dominated by widespread native species in oceanic islands: a case study in the Azores, *Insect Conservation and Diversity* **10**(3): 211–223.

Picanço, A., Rigal, F., Matthews, T. J., Cardoso, P. and Borges, P. A. (2017). Impact of land-use change on flower-visiting insect communities on an oceanic island, *Insect Conservation and Diversity* **10**(3): 211–223.

Poisot, T., Canard, E., Mouquet, N. and Hochberg, M. E. (2012). A comparative study of ecological specialization estimators, *Methods in Ecology and Evolution* **3**(3): 537–544.

Poisot, T., Stouffer, D. B. and Gravel, D. (2015). Beyond species: why ecological interaction networks vary through space and time, *Oikos* **124**(3): 243–251.

Polson, N. G., Scott, J. G. and Willard, B. T. (2015). Proximal Algorithms in Statistics and Machine Learning, *Statistical Science* **30**(4): 559–581.
URL: *https://doi.org/10.1214/15-STS530*

Popic, T. J., Wardle, G. M. and Davila, Y. C. (2013). Flower-visitor networks only partially predict the function of pollen transport by bees, *Austral Ecology* **38**(1): 76–86.

Pößnecker, W. (2014). *MRSP: Multinomial Response Models with Structured Penalties*. R package version 0.4.3.
URL: *http://CRAN.R-project.org/package=MRSP*

Powell, M. (1971). On the Convergence of the Variable Metric Algorithm, *IMA Journal of Applied Mathematics* **7**(1): 21–36.

Powell, M. J. (1976). Some global convergence properties of a variable metric algorithm for minimization without exact line searches, *Nonlinear programming* **9**(1): 53–72.

R Core Team (2011). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

R Core Team (2015). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

R Core Team (2017). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Ravasz, M., Balog, A., Markó, V. and Néda, Z. (2005). The species abundances distribution in a new perspective, *arXiv preprint q-bio/0502029* .

Reid, S. and Tibshirani, R. (2014). Regularization Paths for Conditional Logistic Regression: The clogitL1 Package, *Journal of Statistical Software* **58**(12): 1–23.
URL: *http://www.jstatsoft.org/v58/i12/*

Revelt, D. and Train, K. (1998). Mixed logit with repeated choices: households' choices of appliance efficiency level, *Review of Economics and Statistics* **80**(4): 647–657.

Rezende, E. L., Jordano, P. and Bascompte, J. (2007). Effects of phenotypic complementarity and phylogeny on the nested structure of mutualistic networks, *Oikos* **116**(11): 1919–1929.

Robert, C. and Casella, G. (2010). *Introducing Monte Carlo Methods with R*, Springer-Verlag, New York.

Rohr, R. P., Scherer, H., Kehrli, P., Mazza, C. and Bersier, L.-F. (2010). Modeling food webs: exploring unexplained structure using latent traits, *The American Naturalist* **176**(2): 170–177.

Rojo, S., Isidro, P., Perez-Bañón, M. and Marcos-García, M. (1997). Revision of the hoverflies (Diptera: Syrphidae) from the Azores archipelago with notes on Macaronesian Syrphid fauna, *Arquipélago. Life and Marine Sciences* **15A**: 65–82.

Santamaría, L. and Rodríguez-Gironés, M. A. (2007). Linkage rules for plant–pollinator networks: trait complementarity or exploitation barriers?, *PLoS Biology* **5**(2): e31.

Schäfer, H. (2002). *Flora of the Azores: a Field Guide*, Margraf Verlag, Weikersheim.

Schäfer, H. (2003). Chorology and diversity of the Azorean flora, *Willdenowia* **33**: 481–482.

Schwarz, G. (1978). Estimating the Dimension of a Model, *The Annals of Statistics* **6**(2): 461–464.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization, *Mathematics of Computation* **24**(111): 647–656.

Shonkwiler, J. and Hanley, N. (2003). A new approach to random utility modeling using the Dirichlet multinomial distribution, *Environmental and Resource Economics* **26**(3): 401–416.

Simon, N., Friedman, J. and Hastie, T. (2013). A blockwise descent algorithm for group-penalized multiresponse and multinomial regression, *arXiv preprint arXiv:1311.6529* .

Sjögren, E. (2001). *Plants and Flowers of the Azores*, Eget förlag.

Sorensen, P. B., Damgaard, C. F., Strandberg, B., Dupont, Y. L., Pedersen, M. B., Carvalheiro, L. G., Biesmeijer, J. C., Olsen, J. M., Hagen, M. and Potts, S. G. (2012). A method for under-sampled ecological network data analysis: plant-pollination as case study, *Journal of Pollination Ecology* **6**.

Speight, M. and Sarthou, J. (2012). STN keys for the identification of adult European Syrphidae (Diptera) 2012, *Syrph the Net, the database of European Syrphidae, Syrph the Net publications, Dublin* **70**: 1–130.

Stang, M., Klinkhamer, P. G., Waser, N. M., Stang, I. and van der Meijden, E. (2009). Size-specific interaction patterns and size matching in a plant–pollinator interaction web, *Annals of Botany* **103**(9): 1459–1469.

StataCorp (2011). *Stata Statistical Software: Release 11*, Stata Corp LP, College Station, TX.

Stone, G. N. (1994). Activity patterns of females of the solitary bee *Anthophora plumipes* in relation to temperature, nectar supplies and body size, *Ecological Entomology* **19**(2): 177–189.

Stone, G. and Willmer, P. (1989). Warm-up rates and body temperatures in bees: the importance of body size, thermal regime and phylogeny, *Journal of Experimental Biology* **147**(1): 303–328.

Streeter, D. (1983). *The Wild Flowers of the British Isles*, Vol. 1, Macmillan, London.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1): 91–108.

Tikhonov, A. (1977). *Solutions of Ill-Posed Problems*, Vol. 14, Winston, Washington, DC.

Toint, P. L. (1986). Global convergence of the partitioned BFGS algorithm for convex partially separable optimization, *Mathematical Programming* **36**(3): 290–306.

Torres, C. and Galetto, L. (1998). Patterns and implications of floral nectar secretion, chemical composition, removal effects and standing crop in *Mandevilla pentlandiana (Apocynaceae)*, *Botanical Journal of the Linnean Society* **127**(3): 207–223.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*, Cambridge University Press.

Tutz, G., Pößnecker, W. and Uhlmann, L. (2015). Variable selection in general multinomial logit models, *Computational Statistics & Data Analysis* **82**: 207–222.

Tylianakis, J. M., Tscharntke, T. and Lewis, O. T. (2007). Habitat modification alters the structure of tropical host–parasitoid food webs, *Nature* **445**(7124): 202–205.

van Houwelingen, H. C., Bruinsma, T., Hart, A. A., van't Veer, L. J. and Wessels, L. F. (2006). Cross-validated Cox regression on microarray gene expression data, *Statistics in Medicine* **25**(18): 3201–3216.

Vázquez, D. P. and Aizen, M. A. (2004). Asymmetric specialization: a pervasive feature of plant–pollinator interactions, *Ecology* **85**(5): 1251–1257.

Vázquez, D. P., Blüthgen, N., Cagnolo, L. and Chacoff, N. P. (2009a). Uniting pattern and process in plant–animal mutualistic networks: a review, *Annals of Botany* **103**(9): 1445–1457.

Vázquez, D. P., Chacoff, N. P. and Cagnolo, L. (2009b). Evaluating multiple determinants of the structure of plant–animal mutualistic networks, *Ecology* **90**(8): 2039–2046.

Vázquez, D. P., Morris, W. F. and Jordano, P. (2005). Interaction frequency as a surrogate for the total effect of animal mutualists on plants, *Ecology Letters* **8**(10): 1088–1094.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, fourth edn, Springer, New York. ISBN 0-387-95457-0.
URL: *http://www.stats.ox.ac.uk/pub/MASS4*

Vincent, M. and Hansen, N. R. (2014). Sparse group lasso and high dimensional multinomial classification, *Computational Statistics & Data Analysis* **71**: 771–786.

Wang, H. and Leng, C. (2007). Unified LASSO Estimation by Least Squares Approximation, *Journal of the American Statistical Association* **102**(479): 1039–1048.

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(3): 671–683.

Wang, H., Li, R. and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika* **94**(3): 553–568.

Wang, T. and Zhao, H. (2017). A Dirichlet-Tree Multinomial Regression Model for Associating Dietary Nutrients with Gut Microorganisms, *Biometrics* **73**: 792–801.

Wang, T. and Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression, *Journal of Multivariate Analysis* **102**(7): 1141–1151.

Wells, K. and O'Hara, R. B. (2013). Species interactions: estimating per-individual interaction strength and covariates before simplifying data into per-species ecological networks, *Methods in Ecology and Evolution* **4**(1): 1–8.

Williams, N. M., Crone, E. E., Tai, H. R., Minckley, R. L., Packer, L. and Potts, S. G. (2010). Ecological and life-history traits predict bee species responses to environmental disturbances, *Biological Conservation* **143**(10): 2280–2291.

Willmer, P. (2011). *Pollination and floral ecology*, Princeton University Press.

Wright, S. and Nocedal, J. (1999). Numerical optimization, *Springer Science* **35**: 67–68.

Yee, T. W. (2010). The VGAM package for categorical data analysis, *Journal of Statistical Software* **32**(10): 1–34.

Ypma, T. J. (1995). Historical development of the Newton-Raphson method, *SIAM Review* **37**(4): 531–551.

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.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method, *arXiv preprint arXiv:1212.5701* .

Zeng, P., Wei, Y., Zhao, Y., Liu, J., Liu, L., Zhang, R., Gou, J., Huang, S. and Chen, F. (2014). Variable selection approach for zero-inflated count data via adaptive lasso, *Journal of Applied Statistics* **41**(4): 879–894.

Zhang, Y., Li, R. and Tsai, C.-L. (2010). Regularization Parameter Selections via Generalized Information Criterion, *Journal of the American Statistical Association* **105**(489): 312–323.

Zhang, Y. and Zhou, H. (2016). *MGLM: Multivariate Response Generalized Linear Models*. R package version 0.0.7.
URL: *https://CRAN.R-project.org/package=MGLM*

Zhang, Y., Zhou, H., Zhou, J. and Sun, W. (2017). Regression Models for Multivariate Count Data, *Journal of Computational and Graphical Statistics* **26**(1): 1–37.

Zhao, W., Zhang, R., Lv, Y. and Liu, J. (2014). Variable selection for varying dispersion beta regression model, *Journal of Applied Statistics* **41**(1): 95–108.

Zhou, H. and Lange, K. (2010). MM Algorithms for Some Discrete Multivariate Distributions, *Journal of Computational and Graphical Statistics* **19**(3): 645–665.

Zhou, H. and Zhang, Y. (2012). EM vs MM: A case study, *Computational Statistics & Data Analysis* **56**(12): 3909–3920.

Zito, P., Scrima, A., Sajeva, M., Carimi, F. and Dötterl, S. (2016). Dimorphism in inflorescence scent of dioecious wild grapevine, *Biochemical Systematics and Ecology* **66**: 58–62.

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association* **101**(476): 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2): 301–320.

Zou, H., Hastie, T., Tibshirani, R. et al. (2007). On the Degrees of Freedom of the Lasso, *The Annals of Statistics* **35**(5): 2173–2192.

Appendix A

DM regression and relationship between δ and ρ

We assume that pollinators choose plant species according to some multinomial distribution with probabilities that are random themselves. In particular, the multinomial probabilities are the result of: (i) a structural model defined by relevant linkage rules and plant species traits; and (ii) a measurement model for overdisperion, possibly defined by pollinator specific traits, which leads to a Dirichlet prior distribution for the multinomial probabilities. Note that the Dirichlet distribution (multivariate version of a beta distribution) is the conjugate prior for the multinomial distribution.

We assume that individual pollinators assign a utility to plant species according to,

$$U_{igj} = \beta' x_{gj} + \eta_{gj} + \epsilon_{igj},$$

where β' is a K-length vector of unknown regression coefficients associated with the covariates in x_{gj} and η_{gj} is a scalar random group effect for pollinator species g and plant species j. On a technical note, since pollinators select the plant species with the maximum utility, we take the random error term ϵ_{igj} to follow a generalized extreme value (type I) distribution. The probability that pollinator g selects (visits) plant j is given by:

$$p_{gj} = \frac{\exp(\beta' x_{gj} + \eta_{gj})}{\sum_{j=1}^{J} \exp(\beta' x_{gj} + \eta_{gj})}, \text{ for } g = 1, \ldots, G; \text{ and } j = 1, \ldots, J,$$

where η_{gj} is a random pollinator effect, and β contains K coefficients. This model is known

as the Dirichlet-multinomial regression model (Guimarães and Lindrooth, 2007). The goal of fitting a DM regression is to estimate the regression coefficients, which summarize the contribution of each linkage rule or covariate to the plant-pollinator interaction probabilities. Briefly, p_{gj} provides information on the strength of the links in the network and parameter vector β summarizes the covariates contributions to those probabilities.

By letting $\lambda_{gj} = \exp(\beta' x_{gj})$ for $g = 1, \ldots, G$; $j = 1, \ldots, J$, we can re-express the interaction probabilities by,

$$p_{gj} = \frac{\lambda_{gj}\exp(\eta_{gj})}{\sum_{j=1}^{J} \lambda_{gj}\exp(\eta_{gj})}, \text{ for } g = 1, \ldots, G; \text{ and } j = 1, \ldots, J.$$

We assume that the λ_{gj} are independent and identically gamma distributed with both shape and scale (rate) parameters $\delta_g \lambda_{gj}$, where $\delta_g > 0$. Consequently, the vector of probabilities for pollinator species g, $(p_{g1}, \ldots, p_{gJ})$, follows a Dirichlet distribution (Mosimann, 1962) and the group random effect η_{gj} translates into the pollinator-specific over-dispersion parameter δ_g (Guimarães and Lindrooth, 2007).

The interaction counts, covariates, and model parameters are related to each other as shown in Figure A.1. The observed counts in the rows of the $G \times J$ matrix Y follow a multinomial distribution with parameters given by rows of the $G \times J$ interaction matrix $P = (p_{11}, \ldots, p_{GJ})$. Further, rows of P follow Dirichlet distributions with parameters depending on the product of the observed covariates in the $G \times J \times K$ array X and the associated K-vector parameter β (through λ_{gj}), and on the over-dispersion parameter δ_g (through $\exp(\eta_{gj})$). Hence DM regression is an example of a hierarchical model in which the probabilities associated with a multinomial random variable are themselves random variables following a Dirichlet distribution.

Plant-pollinator network data can be modelled assuming five different dispersion structures: no dispersion, constant dispersion ($\delta_g = \delta$), dispersion as a function of pollinator specific covariates ($\delta_g = \mathrm{f}(z_g)$), constant intra-correlation ($\rho_g = \rho$), and intra-correlation as a function of pollinator specific covariates ($\rho_g = \mathrm{f}(z_g)$)[1]. Whenever we assume δ_g or ρ_g is a function

[1] Although modeling the intra-class correlation as a function of pollinator specific covariates is theoretically feasible, we found that, in practice, this parameterization tends to be unstable and prone to non-convergence. As such, we did not include this in our simulation study or `R` program for DM regression.

of covariates, we also assume that the function is non-linear, e.g., $\mathrm{f}(z_g) = \exp\{-(\gamma_0 + \gamma_1 z_1 + \gamma_2 z_2)\}$ for pollinator species covariates z_1 and z_2. Although the pollinator-specific parameters δ_g or ρ_g account for extra-multinomial variability without affecting the choice probabilities per se, these models can provide additional insight into the nature of the heterogeneity in plant-pollinator networks (Guimarães and Lindrooth, 2007).

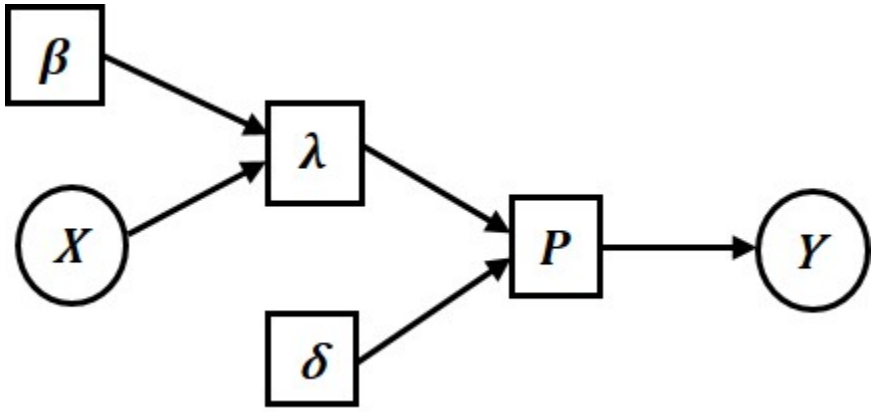

Figure A.1: Graphical model of random variables and parameters in a Dirichlet-multinomial regression model. Boxes represent unobserved quantities to be estimated from data. Circles represent observed variables.

Note that there is a relationship between the dispersion structure in terms of the intra-group correlation coefficient ρ_g and the group over-dispersion parameter δ_g. In particular, whenever $\rho_g = \rho$ we have,

$$\delta_g = \frac{\rho}{1-\rho}\lambda_g, \text{ for } g = 1, \ldots, G,$$

where $\lambda_g = \sum_{j=1}^{J} \lambda_{gj}$. This relation can be exploited whenever over-dispersion is parameterized in terms of ρ so that the model is parameterized in terms of δ_g. Further, under an alternative parameterization of the DM model, the cell frequencies y_{gj} can be shown to follow a negative multinomial distribution with fixed effects (Guimarães and Lindrooth, 2007). Consequently, conditional on the row sums (total number observed interactions, by pollinator species), the cell frequencies follow a Poisson distribution. These parameterizations and relationships are exploited in the simulation study. Finally, maximum likelihood estimates of the model parameters can be obtained using an iterative estimation method and is implemented in our `R` program.

Appendix B

GCL and DM regression R program

Computer code in R (R Core Team, 2014) is available on GitHub (https://github.com/) under repository "GCL_DM_Regression" (Crea, 2015) to estimate the parameters of the GCL and DM regression models. This program consists of a main function named dirmultreg which calls to a number of utility functions written by the authors and R's optim function (R Core Team, 2014).

Here we provide an example of how to use the R program to analyze the artificial data sets discussed in this paper.

Copy and paste the R program into a .R script file and load the file into R so that the dirmultreg function is available for use:

1. Launch the R console
2. File -> New Script. An untitled R Editor window will open.
3. Paste R program into R Editor.
4. File -> Save. A "Save script as" window will open. Navigate to the folder you would like to save this script file; give script a name. Note under "Save as type", it will default to "R files (*.R)". Do not change this file format.
5. Load and execute the script file with the R commands using the source function:
   source(myfile.R)
   where the myfile.R argument is a character string giving the path name and file name.

The data needs to be an R `data.frame` object that is in the following format, for example.

> mydata

plant	poll	Y	x1	x2	x3	x4	z1
1	1	0	1	1	0	0.068115	0.000987
2	1	0	1	0	0	0.062685	0.177443
3	1	1	1	0	1	0.129319	0.001481
4	1	2	1	0	0	0.003455	0.023445
5	1	1	1	0	1	0.017029	0.047631
6	1	2	1	0	0	0.007404	0.08539
7	1	1	1	0	1	0.293189	0.000247
8	1	3	0	0	0	0.031836	0.000247
9	1	1	0	1	0	0.019743	0.068115
10	1	0	1	0	1	0.000987	0.062685

In this example, `plant` and `poll` are group identifiers for plant and pollinator species, respectively, `Y` is the count or number of interactions recorded for that plant and pollinator pair, `x1`, `x2`, `x3`, and `x4` are the associated predictor or covariate values for the plant and pollinator pairs, and `z1` is the associated pollinator species covariate used to model dispersion as a function, if needed. To ensure `mydata` is a data frame, type the command:

```
> mydata <- data.frame(mydata)
```

The `dirmultreg` function fits a GCL, which corresponds to a DM model with no dispersion, and the DM regression models described in this paper. Usage of this function is as follows:

```
dirmultreg(formula, data, disp, plant, poll, delta.vars, method)
```

Arguments:

`formula` - a symbolic description of the model to be fitted which has the form `response` $\sim$ `terms`, where response is the (numeric) response vector and terms is a series of terms which specifies a linear predictor for response, e.g., `x1 + x2` indicates all the terms in first together with all the terms in second.

`data` - a data set containing the variables in the model as shown above. Note these data must be an `R` object of class "`data.frame`".

`disp` - dispersion structure to use: "`none`" (GCL), "`dconst`" (δ constant), "`dfunc`" (γ func-

tion of covariates) and "`rconst`" (ρ constant).
`plant` - group identifier for plant species.
`poll` - group identifier for pollinator species.
`delta.vars` - list of predictor variables specified in `formula` to be used to model dispersion as a function of pollinator specific covariates, i.e., `z1`. Default is `NULL`.
`method` - the method to be used for the optimization procedure: "`Nelder- Mead`", "`BFGS`", "`L-BFGS`", "`CG`", and "`SANN`". Default is "`Nelder-Mead`".

Consider the data set above which consists of three linkage rules (`x1`, `x2`, and `x3`), plant species relative abundance (`x4`) and pollinator species abundance (`z1`). If we choose to run a model assuming **no dispersion**, then the command would be:

`dirmultreg(formula = Y∼x1+x2+x3+x4, data = filename, disp =` "`none`", `plant =` "`plant`", `poll =` "`poll`", `delta.vars = NULL, method =` "`BFGS`")

Note that `z1` was not used because we are specifying no dispersion, hence the dispersion parameter cannot be modeled as a function of covariates. Also, the results of the simulation study suggested that method "`BFGS`" worked best for the no dispersion and intra-correlation models while "`Nelder-Mead`" worked best for the constant and function of covariate models.

Alternatively, if we chose to run a model assuming **dispersion as a function of covariates**, then the command would be:
`dirmultreg(formula = Y∼x1+x2+x3+x4+z1, data = filename, disp =` "`dfunc`", `plant =` "`plant`", `poll =` "`poll`", `delta.vars =` "`z1`", `method =` "`Nelder-Mead`")

In this example, we have added `z1` to the `formula` and `delta.vars` arguments and specified "`Nelder-Mead`" for the optimization procedure. A summary of these commands for all DM regression models are summarized in Table B.1.

Table B.1: Commands to fit a GCL model or DM model with over-dispersion.

Dispersion	**R Command**
None ($\delta_g = 0$)	`dirmultreg(formula = Y∼x1+x2+x3+x4, data = filename, disp = "none", plant = "plant", poll = "poll", delta.vars = NULL, method = "BFGS")`
Constant ($\delta_g = \delta$)	`dirmultreg(formula = Y∼x1+x2+x3+x4, data = filename, disp = "dconst", plant = "plant", poll = "poll", delta.vars = NULL, method = "Nelder-Mead")`
Function of covariates ($\delta_g = f(z_g)$)	`dirmultreg(formula = Y∼x1+x2+x3+x4+z1, data = filename, disp = "dfunc", plant = "plant", poll = "poll", delta.vars = "z1", method = "Nelder-Mead")`
Constant correlation ($\rho_g = \rho$)	`dirmultreg(formula = Y∼x1+x2+x3+x4, data = filename, disp = "rconst", plant = "plant", poll = "poll", delta.vars = NULL, method = "BFGS")`

Appendix C

Data generation

This appendix describes the `R` (R Core Team, 2014) code to randomly generate pollination networks according to a Dirichlet-multinomial regression model with four covariates:
x_1 = barrier trait;
x_2 = complementarity trait with narrow range variability;
x_3 = complementarity trait with medium range variability; and
x_4 = plant species relative abundance.

If a network is generated with over-dispersion as a function of covariates, it is in terms of z_1 = pollinator species relative abundance.

As mentioned in Appendix A, the negative multinomial parameterization of DM regression, in which cell counts follow a Poisson distribution conditional on the row sums, is exploited. In particular, we assume that λ_{gj} follow a gamma distribution with parameters $(\delta_g^{-1}\lambda_{gj}, \delta_g^{-1})$. Then, y_{gj} conditional on the row sums follows a Poisson(λ_{gj}) distribution. Using this formulation of the DM regression model, random networks and covariates matrices were simulated for each combination of the model parameters. Users must specify exact values for the model parameters.

There are three utility functions to generate: relative abundances (`simZ`), covariate matrices that include linkage rules (`simX`), and counts for a random network (`simY`). The main function is called `simData` and is available as an accessible file on the GitHub repository "`GCL_DM_Regression`" (Crea, 2015).

Appendix D

Quantitative pollen transfer network for Canterbury data

	Plant Species														
Pollinator Species	*Lolium sp*	*Malva sp*	*Taraxacum officinale*	*Raphanus sp*	*Cirsium vulgare*	*Eucalyptus globulus*	*Eucalyptus nitens*	*Trifolium sp*	*Daucus carota*	*Chenopodium album*	*Cordyline sp*	*Alcea sp*	*Rubus fruticosus*	*alnus serrulate*	*alnus glutinosa*
Apis mellifera	37	0	42	4	21	24	28	63	4	1	3	9	10	0	1
Bombus terrestris	3	0	4	0	1	3	1	3	0	0	1	1	1	0	0
Calliphora stygia	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0
Calliphora vicina	1	0	0	0	0	1	1	3	0	0	0	0	0	0	0
Delia platura	5	1	3	2	4	4	2	14	2	1	4	0	2	1	0
Eristalis tenax	1	0	3	0	4	2	3	5	0	1	2	0	3	0	0
Helophilus hochstetteri	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0
Ichneumon promissorius	3	0	2	0	2	0	0	3	0	0	0	0	0	0	0
Lasioglossum sordidum	12	2	14	4	5	0	1	10	4	0	0	3	0	0	0
Melangyna novaezelandiae	1	0	2	0	0	0	0	0	0	0	1	0	0	1	0
Netelia producta	2	0	2	1	1	0	0	4	1	0	1	1	0	0	0
Odontomyia atrovirens	2	0	1	0	3	0	1	2	1	1	0	0	0	0	0
Oxysarcodexia varia	4	0	1	0	1	0	0	0	0	0	0	0	0	0	0
Pales marginata	5	0	2	0	1	1	1	3	2	1	0	0	0	0	0
Pollenia pseudorudis	3	0	1	0	1	0	0	4	1	0	0	0	0	0	0
Thyreocephalus orthodoxus	5	0	1	0	1	0	0	1	0	1	0	1	0	0	0

Appendix E

Network statistics

The following boxplots display the distributions of the network metrics that were calculated from the networks generated in the simulation study but not presented in the main paper. Note that interaction evenness, Shannons diversity and H_2' (Figure E.3) are collinear, and the latter two are just indices, not measures of true diversities (Jost, 2006). We report on H_2' here and on relative diversity,

$$\text{Relative Diversity} = \frac{\exp(H)}{G \times J}$$

in the main manuscript. The numerator provides an estimate of the effective number of links (or interacting number of plant-pollinator species pairs) while the denominator equals the number of links in a network in which all potential links are equally common/strong. Note that interaction evenness, as defined in bipartite, is a ratio of diversity indices, but not of true densities. Finally, note that due to computational complexities, we could not compute modularity for all networks in the simulation study. However, Dormann and Strauss (2013) show that the Q-index of their modularity algorithm `computeModules` is highly positively correlated with H_2' (see Figure 9 in their paper).

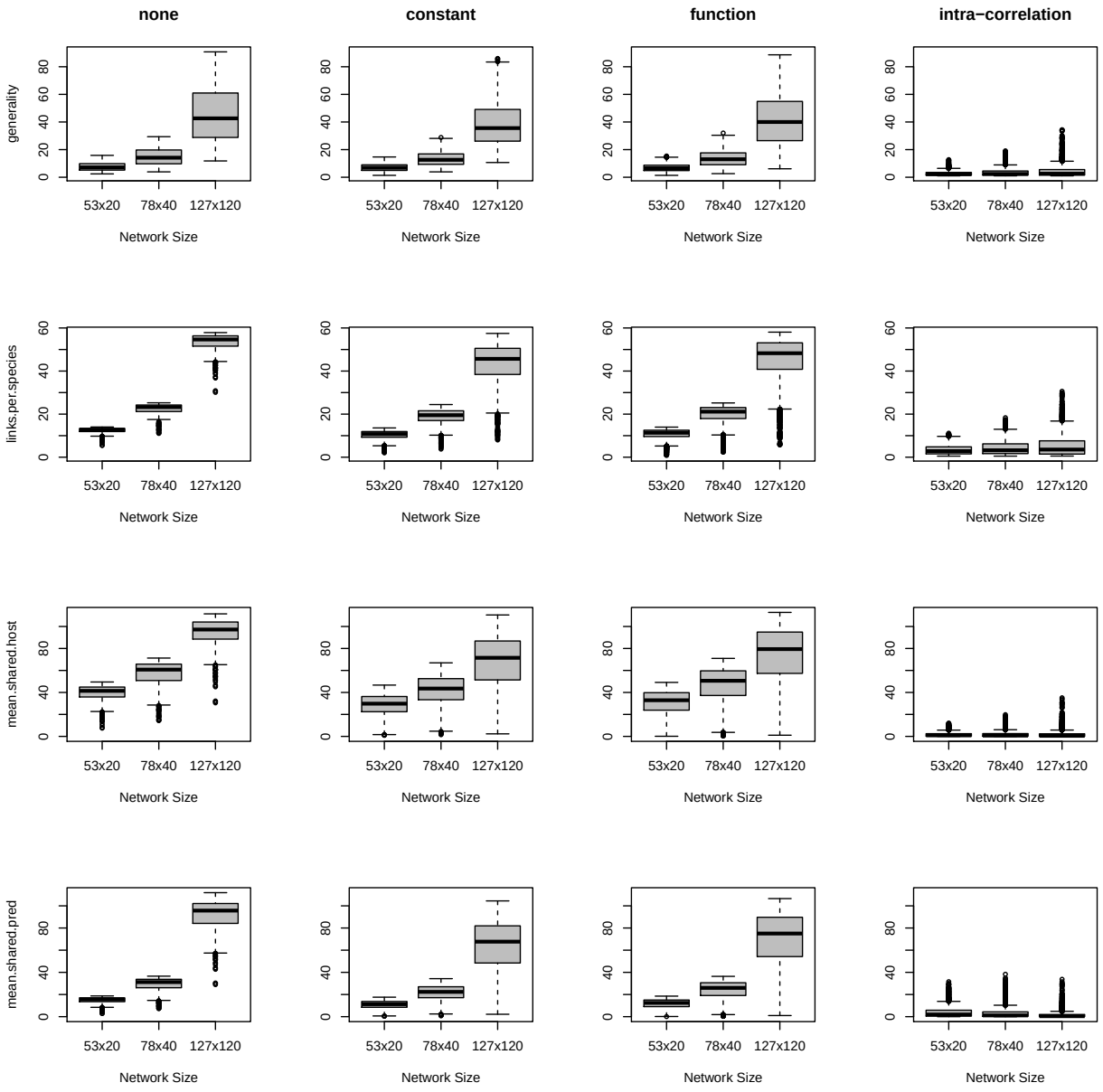

Figure E.1: Boxplots of generality, links per species, mean number of shared partners for host (plant species) and predator (pollinator species) of artificially generated networks by network size and dispersion structure. None: $\delta_g = 0$; Constant: $\delta_g = \delta$; Function: $\delta_g = f(z_g)$; Intra-correlation: $\rho_g = \rho$.

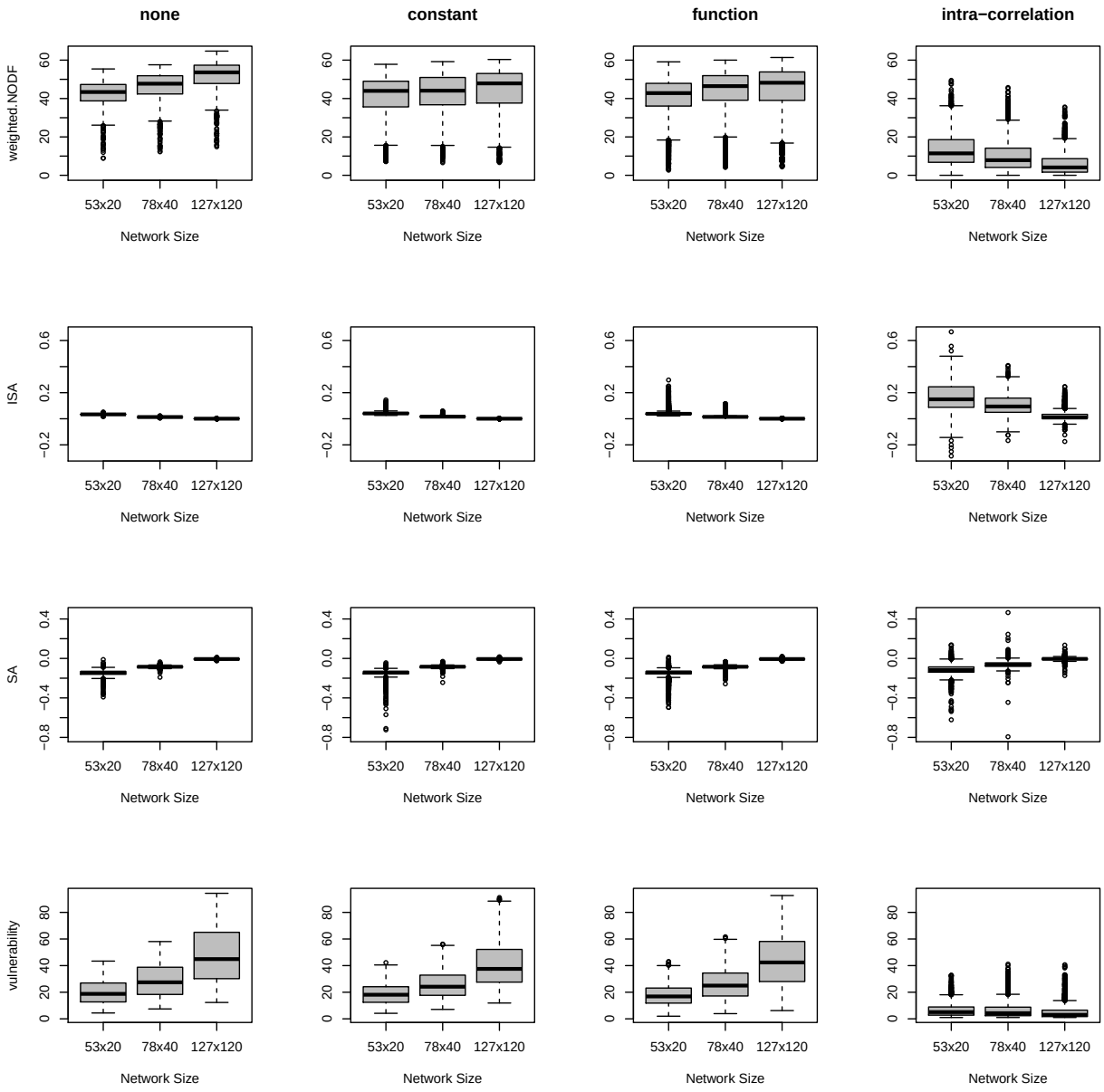

Figure E.2: Box plots of weighted NODF, interaction strength asymmetry (ISA), specialization asymmetry (SA), and vulnerability of artificially generated networks by network size and dispersion structure. None: $\delta_g = 0$; Constant: $\delta_g = \delta$; Function: $\delta_g = f(z_g)$; Intra-correlation: $\rho_g = \rho$.

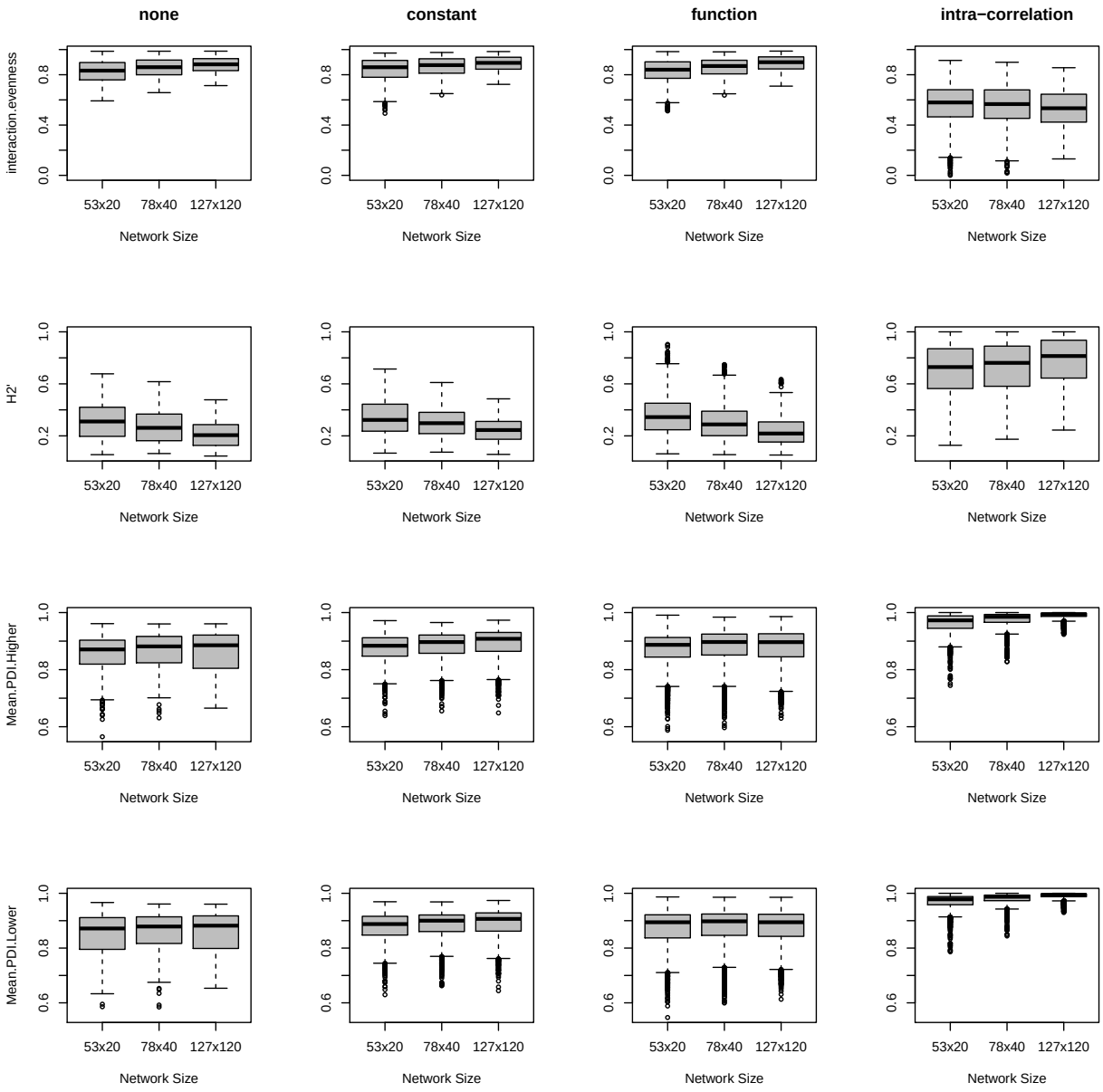

Figure E.3: Box plots of interaction evenness, H_2', and mean PDI (by species level) of artificially generated networks by network size and dispersion structure. None: $\delta_g = 0$; Constant: $\delta_g = \delta$; Function: $\delta_g = f(z_g)$; Intra-correlation: $\rho_g = \rho$.

Appendix F

List of plant-pollinator networks from the Interaction Web Database (IWDB) used to specify network sizes for simulation study

Dataset	Habitat Type	Location	Data Type	# Plants	# Pollinators
Arroyo et al. (1982)	Andean scrub	Chile	binary	87	98
Arroyo et al. (1982)	Andean scrub	Chile	binary	43	62
Arroyo et al. (1982)	Andean scrub	Chile	binary	41	28
Barrett and Helenurm (1987)	Boreal forest	Canada	individuals caught	12	102
Bezerra et al. (2009)	Caatinga (semi-arid vegetation)	Pernambuco State, Brazil	no. visits	13	13
Clements and Long (1923)	Montane forest and grassland	USA	binary	96	276
Dupont et al. (2003)	High-altitude desert	Tenerife, Canary Islands	binary	11	38
Elberling and Olesen (1999)	Alpine subarctic community	Sweden	no. visits	23	118
Hocking (1968)	Arctic community	Canada	binary	29	86
Kaiser-Bunbury et al. (2009)	Heathland habitat	Mauritius	rates	135	74
Kaiser-Bunbury et al. (2009)	Heathland habitat (reduced)	Mauritius	rates	100	64
Kato et al. (1990)	Beech forest	Japan	individuals caught	93	679
Kevan (1970)	Northern Territory, Canada	no. visits	32	115	
Inouye and Pyke (1988)	Montane forest	Australia	individuals caught	42	91
McMullen (1993)	Multiple communities	Galpagos Islands	binary	106	54
Medan et al. (2002)	Xeric scrub	Laguna Diamante, Argentina	binary	21	45
Medan et al. (2002)	Woody riverine vegetation and xeric scrub	Ro Blanco, Argentina	binary	23	72
Memmott (1999)	Medow	Bristol, U.K.	frequency of visits	25	79
Mosquin and Martin (1967)	Arctic community	Canada	individuals caught	11	18
Motten (1982)	Deciduous forest	USA	no. visits	13	44
Olesen et al. (2002)	Coastal forest	Mauritius Island	no. visits	14	13
Olesen et al. (2002)	Rocky cliff and open herb community	Azores Islands	no. visits	10	12
Ollerton et al. (2003)	Upland grassland	South Africa	individuals caught	9	56
Ramrez and Brito (1992)	Palm swamp community	Venezuela	binary	33	53
Robertson (1929)	Agricultural area dominated by crops	USA	binary	456	1429
Santos et al. (2010)	Caatinga (semi-arid vegetation)	Bahia State, Brazil	binary	51	25
Schemske et al. (1978)	Maple-oak woodland	USA	no. visits	7	32
Small (1976)	Peat bog	Canada	individuals caught	13	34
Vázquez and Simberloff (2002)	Evergreen montane forest	Argentina	no. visits	10	29
Vázquez and Simberloff (2002)	Evergreen montane forest	Argentina	no. visits	9	33
Vázquez and Simberloff (2002)	Evergreen montane forest	Argentina	no. visits	9	27
Vázquez and Simberloff (2002)	Evergreen montane forest	Argentina	no. visits	10	29
Vázquez and Simberloff (2002)	Evergreen montane forest	Argentina	no. visits	8	26
Vázquez and Simberloff (2002)	Evergreen montane forest	Argentina	no. visits	7	24
Vázquez and Simberloff (2002)	Evergreen montane forest	Argentina	no. visits	8	27

Appendix G

Derivation of Logit Formulation from the Random Utility Model Used in Discrete Choice Modelling

McFadden's random utility model McFadden (1974) states that individuals are faced with J mutually exclusive and exhaustive choices. The individual assigns a level of utility to each choice and selects the one with maximum utility. The utility assigned to each choice is:

$$U_j = V_j + \epsilon_j, \quad j = 1, \ldots, J \tag{G.1}$$

where $V_j = \beta' x_j$ is the systematic component comprised of unknown regression coefficients in β' and the associated observable covariates in x_j, and ϵ_j is the unobserved component, or error term, independently and identically distributed Extreme Value Type I. The probability density and cumulative distribution functions for the error term are:

$$f(\epsilon_j) = \mathrm{e}^{-\epsilon_j - \mathrm{e}^{-\epsilon_j}},$$

and

$$F(\epsilon_j < \epsilon) = \mathrm{e}^{-\mathrm{e}^{-\epsilon}},$$

respectively. The observed counts, Y_j are defined as:

$$Y_j = \begin{cases} 1, & \text{if } U_j = \max(U_1, \ldots, U_J), \\ 0, & \text{otherwise.} \end{cases}$$

In other words, alternative j will be chosen if and only if $U_j > U_k \ \forall\ j \neq k$. Since ϵ_j are not observed, choices can only be modelled in terms of probabilities:

$$\epsilon_j + V_j > \epsilon_k + V_k \quad \Leftrightarrow \quad \epsilon_k < V_j - V_k + \epsilon_j, \quad \text{for all } k \neq j.$$

Because $\epsilon_1, \ldots, \epsilon_J$ are i.i.d., the probability that choice j is selected, given the observed covariates in X, is given by:

$$\begin{aligned} P(Y_j = 1|X) &= P(\epsilon_k < V_j - V_k + \epsilon_j, k \neq j) \\ &= \prod_{k \neq j} \int_{-\infty}^{\infty} F(V_j - V_k + \epsilon_j) f(\epsilon_j) \mathrm{d}\epsilon_j \\ &= \int_{-\infty}^{\infty} \prod_{k \neq j} \mathrm{e}^{-\mathrm{e}^{-V_j + V_k - \epsilon}} \mathrm{e}^{-\epsilon_j - \mathrm{e}^{-\epsilon_j}} \mathrm{d}\epsilon_j \\ &= \int_{-\infty}^{\infty} \mathrm{e}^{-\epsilon_j - \mathrm{e}^{-\epsilon_j} \left(1 + \sum_{k \neq j} \frac{\mathrm{e}^{V_k}}{\mathrm{e}^{V_j}}\right)} \mathrm{d}\epsilon_j. \end{aligned} \tag{G.2}$$

If we let $\lambda_j = \log\left(1 + \sum_{k \neq j} \frac{\mathrm{e}^{V_k}}{\mathrm{e}^{V_j}}\right) = \log\left(\sum_{j=1}^{J} \frac{\mathrm{e}^{V_k}}{\mathrm{e}^{V_j}}\right)$ and $\epsilon_j^* = \epsilon_j - \lambda_j$, then we can rewrite (G.2) as:

$$\begin{aligned} \int_{-\infty}^{\infty} \mathrm{e}^{-\epsilon_j - \mathrm{e}^{-(\epsilon_j - \lambda_j)}} \mathrm{d}\epsilon_j &= \mathrm{e}^{-\lambda_j} \int_{-\infty}^{\infty} \mathrm{e}^{-\epsilon_j^* - \mathrm{e}^{-\epsilon_j^*}} \mathrm{d}\epsilon_j^* \\ &= \mathrm{e}^{-\lambda_j} \\ &= \frac{\mathrm{e}^{V_j}}{\sum_{j=1}^{J} \mathrm{e}^{V_j}}, \end{aligned}$$

which is the standard logit formulation used to model multinomial probabilities using logistic regression.

Appendix H

DM Model as a Special Case of the Mixed Logit

The mixed logit model (MXL), also known as the random parameters or error components logit, is a generalization of the standard logit, which allows for a more flexible error structure in the utility function (G.1). The standard logit suffers from the independence of irrelevant alternatives (IIA) property, which states that the ratio of probabilities for any two choices is independent of the existence and attributes of any other choice Brownstone and Train (1998). This restrictive substitution pattern may not be practical in some applications; however, allowing the unobserved portion of utility to consist of additional random components effectively eliminates the IIA property and may provide a more realistic substitution pattern Revelt and Train (1998).

MXL models can be derived under a variety of specifications, and each one provides a particular interpretation Train (2009). In particular, the introduction of additional random components can be specified in two ways depending on the application: as an error components logit or as a random parameters logit. The former is analogous to the group-specific random effect in the DM model. In the latter, the regression coefficients β are allowed to vary by group, where the distribution of β can be any discrete or continuous distribution, but the most common is the multivariate normal distribution.
We will focus on a very general specification of the stochastic component of the utility function to simplify the discussion. Consider the utility ascribed to choice j by individual

index i:

$$U_{ij} = \beta' x_{ij} + \eta_{ij} + \epsilon_{ij},$$

where, β' is the vector of fixed unknown regression coefficients associated with observable covariates in x_{ij}; η_{ij} is the first random term with zero mean and may be correlated over the choices and heteroskedastic over individuals and choices depending on the specification; and ϵ_{ij} is the second random term iid Extreme Value Type I. The distribution for η, $f(\eta|\theta)$, can be any discrete or continuous distribution, where θ are the parameters to be estimated. Conditional on η, the probability that individual i selects choice j is the standard logit:

$$L_{ij}(\eta) = \frac{\exp(\beta' x_{ij} + \eta_{ij})}{\sum_{j=1}^{J} \exp(\beta' x_{ij} + \eta_{ij})}.$$

Since the distribution of η is not given, the unconditional probability is the integral of the conditional probability over the distribution of η:

$$P_{ij} = \int \mathrm{L}_{ij} f(\eta|\theta) \mathrm{d}\eta.$$

In other words, the choice probabilities are a mixture of logits with f as the mixing distribution, hence the name "mixed logit" Brownstone and Train (1998). In general, the choice probabilities do not have a closed form solution and the integral is approximated through simulation:

$$\tilde{P}_{ij} = \frac{1}{R} \sum_{r=1}^{R} L_{ij}(\eta^r),$$

where $\tilde{P}_{ij}$ is the simulated probability that individual i selects choice j, R is the number of replications and η^r is value of η on the r^{th} draw. The MXL log-likelihood function is then approximated by the simulated log-likelihood function:

$$l_{MXL} = \sum_{i=1}^{N} \sum_{j=1}^{J} \ln(\tilde{P}_{ij}),$$

and the estimated parameters are those that maximize the simulated log-likelihood.

The DM model is a special case of the MXL when the specification for η reflects unobservable group-specific effects. These random group effects induce correlations across the choices of members in the same group, which translates into overdispersion in the counts Guimarães and Lindrooth (2007). Further, we model the $e^{\eta_{ij}}$'s as independent gamma distributions leading to the DM distribution for the choice probabilities Mosimann (1962). Unlike the MXL model, the DM model is fully parametric, is available in closed form, and can be maximized using standard numerical optimization routines.

Appendix I

First Partial Derivatives for the DM Model

In this appendix we derive the first partial derivatives for the DM model, which are used in the maximization of the associated likelihood. However, the exact form of the derivatives depend on the parameterization of the overdispersion. In Section I.1, we deal with the parameterization in terms of δ and in Section I.2, we deal with the parameterization in terms of ρ.

I.1 DM Parameterized in terms of δ (dconst and dfunc)

The log-likelihood function is:

$$\begin{aligned} l_{DMd} &= \sum_{g=1}^{G} \Big\{ \ln(n_g!) + \ln\Gamma(\delta_g^{-1}\lambda_g) - \ln\Gamma(\delta_g^{-1}\lambda_g + n_g) \\ &\quad + \sum_{j=1}^{J} \ln\Gamma(\delta_g^{-1}\lambda_{gj} + y_{gj}) - \ln\Gamma(\delta_g^{-1}\lambda_{gj}) - \ln(y_{gj}!) \Big\}, \end{aligned}$$

where, $n_g = \sum_{j=1}^{J} y_{gj}$, $\delta_g^{-1} = \mathrm{e}^{\boldsymbol{\gamma}' z_g}$, $\lambda_{gj} = \mathrm{e}^{\boldsymbol{\beta}' x_{gj}}$, and $\lambda_g = \sum_{j=1}^{J} \lambda_{gj}$ for $g = 1, \ldots, G$. When δ_g is constant, then $z_g = 1$ for all g.

The gradient vector $S(\cdot)$ is:

$$S(\cdot) = \begin{pmatrix} \frac{\partial l}{\partial \boldsymbol{\beta}} \\ - - - \\ \frac{\partial l}{\partial \boldsymbol{\gamma}} \end{pmatrix},$$

where $\frac{\partial l}{\partial \boldsymbol{\beta}} = \left(\frac{\partial l}{\partial \beta_1}, \ldots, \frac{\partial l}{\partial \beta_K} \right)^T$ and $\frac{\partial l}{\partial \boldsymbol{\gamma}} = \left(\frac{\partial l}{\partial \gamma_1}, \ldots, \frac{\partial l}{\partial \gamma_L} \right)^T$.

The individual elements of the first partition of $S(\cdot)$ are as follows:

$$\frac{\partial l}{\partial \beta_k} = \sum_{g=1}^{G} \sum_{j=1}^{J} \frac{\partial l}{\partial \lambda_{gj}} \cdot \frac{\partial \lambda_{gj}}{\partial \beta_k}, \; k = 1, \ldots, K,$$

where, for $g = 1, \ldots, G$ and $j = 1, \ldots, J$,

$$\frac{\partial l}{\partial \lambda_{gj}} = \Psi(\delta_g^{-1}\lambda_g) - \Psi(\delta_g^{-1}\lambda_g + n_g) + \Psi(\delta_g^{-1}\lambda_{gj} + y_{gj}) - \Psi(\delta_g^{-1}\lambda_{gj}),$$

and

$$\frac{\partial \lambda_{gj}}{\partial \beta_k} = x_{gjk} \mathrm{e}^{\boldsymbol{\beta}' x_{gj} + \boldsymbol{\gamma}' z_g}.$$

Similarly, the individual elements of the second partition of $S(\cdot)$ are given by:

$$\frac{\partial l}{\partial \gamma_l} = \sum_{g=1}^{G} \sum_{j=1}^{J} \frac{\partial l}{\partial \delta_g^{-1}} \cdot \frac{\partial \delta_g^{-1}}{\partial \gamma_l}, \; l = 1, \ldots, L,$$

where, for $g = 1, \ldots, G$ and $j = 1, \ldots, J$,

$$\frac{\partial l}{\partial \delta_g^{-1}} = \Psi(\delta_g^{-1}\lambda_g) - \Psi(\delta_g^{-1}\lambda_g + n_g) + \Psi(\delta_g^{-1}\lambda_{gj} + y_{gj}) - \Psi(\delta_g^{-1}\lambda_{gj}),$$

and

$$\frac{\partial \delta_g^{-1}}{\partial \gamma_l} = z_{gl} \mathrm{e}^{\boldsymbol{\beta}' x_{gj} + \boldsymbol{\gamma}' z_g}.$$

I.2 DM Parameterization in terms of ρ

The log-likelihood and first partial derivatives for the DM parameterization in terms of ρ are provided separately for ρ as a constant, i.e., $\rho_g = \rho$, and ρ as a function of group covariates, i.e., $\text{logit}(\rho_g) = \text{e}^{\boldsymbol{\gamma}' z_g}$.

I.2.1 ρ as a constant (rconst)

The log-likelihood function is:

$$\begin{aligned} l_{DMr} &= \sum_{g=1}^{G} \Big\{ \ln(n_g!) + \ln\Gamma(\rho_g^{-1} - 1) - \ln\Gamma[(\rho_g^{-1} - 1) + n_g)] \\ &\quad + \sum_{j=1}^{J} \ln\Gamma[(\rho_g^{-1} - 1)p_{gj} + y_{gj}] - \ln\Gamma[(\rho_g^{-1} - 1)p_{gj}] - \ln(y_{gj}!) \Big\}, \end{aligned}$$

where $\rho_g = \rho$, $n_g = \sum_{j=1}^{J} y_{gj}$, and $p_{gj} = \frac{e^{\boldsymbol{\beta}' x_{gj}}}{\sum_{j=1}^{J} e^{\boldsymbol{\beta}' x_{gj}}}$.

The gradient vector $S(\cdot)$ is:

$$S(\cdot) = \begin{pmatrix} \frac{\partial l}{\partial \boldsymbol{\beta}} \\ - - - \\ \frac{\partial l}{\partial \rho} \end{pmatrix},$$

where $\frac{\partial l}{\partial \boldsymbol{\beta}} = \left(\frac{\partial l}{\partial \beta_1}, \ldots, \frac{\partial l}{\partial \beta_K} \right)^T$ and $\frac{\partial l}{\partial \rho}$ is a scalar.

The individual elements of the first partition of $S(\cdot)$ are as follows:

$$\frac{\partial l}{\partial \beta_k} = \sum_{g=1}^{G} \sum_{j=1}^{J} \frac{\partial l}{\partial p_{gj}} \cdot \frac{\partial p_{gj}}{\partial \beta_k}, \; k = 1, \ldots, K,$$

where, for $g = 1, \ldots, G$ and $j = 1, \ldots, J$,

$$\frac{\partial l}{\partial p_{gj}} = (\rho^{-1} - 1) \Big\{ \Psi\big[(\rho^{-1} - 1)p_{gj} + y_{gj}\big] - \Psi\big[(\rho^{-1} - 1)p_{gj}\big] \Big\},$$

and

$$\frac{\partial p_{gj}}{\partial \beta_k} = \frac{e^{\beta' x_{gj}}\left(x_{gjk}\sum_{j=1}^{J} e^{\beta' x_{gj}} - \sum_{j=1}^{J} x_{gjk} e^{\beta' x_{gj}}\right)}{\left(\sum_{j=1}^{J} e^{\beta' x_{gj}}\right)^2}.$$

Similarly, the second partition of $S(\cdot)$ consists of one individual element given by:

$$\begin{aligned}\frac{\partial l}{\partial \rho} = \sum_{g=1}^{G} -\rho^{-2}\Bigg\{&\Psi(\rho^{-1}-1) - \Psi\big[(\rho^{-1}-1)+n_g\big] \\ &+ \sum_{j=1}^{J}\Big(p_{gj}\Big[\Psi\big[(\rho^{-1}-1)p_{gj}+y_{gj}\big] - \Psi\big[(\rho^{-1}-1)p_{gj}\big]\Big]\Big)\Bigg\}.\end{aligned}$$

I.2.2 ρ as a function of group covariates (rfunc)

The log-likelihood function is:

$$\begin{aligned}l_{DMr} = \sum_{g=1}^{G}\Big\{&\ln(n_g!) + \ln\Gamma(\theta_g) - \ln\Gamma(\theta_g+n_g) \\ &+ \sum_{j=1}^{J}\ln\Gamma(\theta_g p_{gj}+y_{gj}) - \ln\Gamma(\theta_g p_{gj}) - \ln(y_{gj}!)\Big\},\end{aligned}$$

where $\theta_g = \rho_g^{-1} - 1 = \mathrm{e}^{-\boldsymbol{\gamma}' z_g}$, $n_g = \sum_{j=1}^{J} y_{gj}$, and $p_{gj} = \frac{e^{\boldsymbol{\beta}' x_{gj}}}{\sum_{j=1}^{J} e^{\boldsymbol{\beta}' x_{gj}}}$.

The gradient vector $S(\cdot)$ is:

$$S(\cdot) = \begin{pmatrix} \frac{\partial l}{\partial \boldsymbol{\beta}} \\ - - - \\ \frac{\partial l}{\partial \boldsymbol{\gamma}} \end{pmatrix},$$

where $\frac{\partial l}{\partial \boldsymbol{\beta}} = \left(\frac{\partial l}{\partial \beta_1}, \ldots, \frac{\partial l}{\partial \beta_K}\right)^T$ and $\frac{\partial l}{\partial \boldsymbol{\gamma}} = \left(\frac{\partial l}{\partial \gamma_1}, \ldots, \frac{\partial l}{\partial \gamma_L}\right)^T$.

The individual elements of the first partition of $S(\cdot)$ are as follows:

$$\frac{\partial l}{\partial \beta_k} = \sum_{g=1}^{G}\sum_{j=1}^{J} \frac{\partial l}{\partial p_{gj}} \cdot \frac{\partial p_{gj}}{\partial \beta_k},\ k = 1, \ldots, K,$$

where, for $g = 1, \ldots, G$ and $j = 1, \ldots, J$,

$$\frac{\partial l}{\partial p_{gj}} = \theta_g \big[\Psi(\theta_g p_{gj} + y_{gj}) - \Psi(\theta_g p_{gj})\big],$$

and,

$$\frac{\partial p_{gj}}{\partial \beta_k} = \frac{e^{\beta' x_{gj}} \big(x_{gjk} \sum_{j=1}^{J} e^{\beta' x_{gj}} - \sum_{j=1}^{J} x_{gjk} e^{\beta' x_{gj}}\big)}{\big(\sum_{j=1}^{J} e^{\beta' x_{gj}}\big)^2}.$$

Similarly, the individual elements of the second partition of $S(\cdot)$ are as follows:

$$\frac{\partial l}{\partial \gamma_l} = \sum_{g=1}^{G}\sum_{j=1}^{J} \frac{\partial l}{\partial \theta_g} \cdot \frac{\partial \theta_g}{\partial \gamma_l},\ l = 1, \ldots, L,$$

where, for $g = 1, \ldots, G$ and $j = 1, \ldots, J$,

$$\frac{\partial l}{\partial \theta_g} = \Psi(\theta_g) - \Psi(\theta_g + n_g) + p_{gj} \big[\Psi(\theta_g p_{gj} + y_{gj}) - \Psi(\theta_g p_{gj})\big],$$

and

$$\frac{\partial \theta_g}{\partial \gamma_l} = -z_{gl} \mathrm{e}^{-\boldsymbol{\gamma}' z_g}.$$

Appendix J

Optimizing the grouped Dirichlet-multinomial regression model: Additional Results Tables

This appendix contains the additional tables summarizing the results of the simulation study as follows:

J.1 - Results of misspecified model fits when there is no dispersion.
J.2 - Results of DM model fits when true dispersion matches modelled dispersion (Network Size 78×40).
J.3 - Results of misspecified DM model fits parameterized in terms of δ (Network Size 78×40).
J.4 - Results of misspecified DM model fits parameterized in terms of ρ (Network Size 78×40).

Table J.1: Results of misspecified model fits when there is no dispersion.

True		Converge	Good Fit	Percent Relative Bias of β		Coverage of 95% Wald CI's	
Disp.[1]	Method	(%)[2]	(%)[3]	Median	IQR[4]	Individual	Family
Network Size 53×20							
dconst	BFGS	100	0	(1.50, 1.03, 1.23, 3.39)	(5.10, 4.13, 4.17, 11.13)	(57, 58, 57, 57)	15
	NM	100	0	(1.50, 1.02, 1.24, 3.42)	(5.10, 4.12, 4.18, 11.18)	(57, 58, 57, 58)	16
	MNR	100	0	(1.50, 1.03, 1.24, 3.43)	(5.10, 4.11, 4.17, 11.16)	(57, 58, 57, 57)	16
rconst	BFGS	100	0	(24.33, 19.18, 21.23, 62.70)	(61.82, 57.90, 47.11, 211.42)	(10, 10, 10, 8)	1
	NM	100	0	(24.31, 19.30, 21.28, 63.07)	(62.20, 58.35, 47.31, 212.41)	(10, 10, 10, 8)	1
	MNR	100	0	(24.37, 19.30, 21.27, 63.12)	(62.22, 58.40, 47.30, 212.95)	(10, 10, 10, 8)	1
dfunc	BFGS	100	3	(1.23, 0.80, 0.87, 2.83)	(4.28, 3.57, 3.00, 9.33)	(66, 68, 67, 67)	35
	NM	100	3	(1.24, 0.81, 0.87, 2.84)	(4.28, 3.59, 3.00, 9.35)	(66, 68, 67, 67)	35
	MNR	100	3	(1.23, 0.80, 0.86, 2.84)	(4.30, 3.57, 3.00, 9.33)	(66, 68, 67, 67)	35
rfunc	BFGS	100	0	(13.38, 10.60, 14.90, 41.51)	(34.15, 27.88, 40.37, 134.65)	(13, 13, 14, 12)	1
	NM	100	0	(13.38, 10.60, 14.91, 41.51)	(34.26, 27.94, 40.40, 134.86)	(13, 13, 14, 12)	1
	MNR	100	0	(13.38, 10.60, 14.91, 41.55)	(34.28, 27.95, 40.41, 134.83)	(13, 13, 14, 12)	1
Network Size 78×40							
dconst	BFGS	100	0	(0.83, 0.48, 0.60, 3.61)	(2.64, 1.72, 1.66, 14.98)	(59, 60, 60, 60)	18
	NM	100	0	(0.83, 0.49, 0.59, 3.74)	(2.63, 1.71, 1.65, 15.72)	(59, 60, 60, 59)	17
	MNR	100	0	(0.83, 0.49, 0.59, 3.67)	(2.63, 1.73, 1.66, 15.50)	(59, 60, 60, 59)	17
rconst	BFGS	100	0	(20.56, 14.74, 17.92, 82.65)	(50.53, 39.97, 43.05, 341.29)	(7, 7, 8, 6)	0
	NM	100	0	(20.55, 14.62, 17.91, 82.59)	(50.59, 39.79, 42.96, 340.05)	(7, 7, 8, 6)	0
	MNR	100	0	(20.57, 14.76, 18.09, 82.85)	(51.01, 40.07, 43.54, 342.90)	(7, 7, 8, 6)	0
dfunc	BFGS	100	2	(0.61, 0.42, 0.48, 2.43)	(2.27, 1.62, 1.69, 8.25)	(65, 67, 68, 68)	34
	NM	100	2	(0.61, 0.42, 0.48, 2.47)	(2.25, 1.62, 1.70, 8.37)	(65, 67, 68, 68)	34
	MNR	100	2	(0.61, 0.42, 0.48, 2.46)	(2.26, 1.61, 1.69, 8.38)	(65, 67, 68, 67)	34
rfunc	BFGS	100	0	(13.32, 10.33, 14.51, 38.30)	(34.21, 27.78, 36.86, 126.00)	(15, 15, 15, 14)	2
	NM	100	0	(13.36, 10.34, 14.51, 38.34)	(34.23, 27.77, 36.87, 126.19)	(15, 15, 15, 14)	2
	MNR	100	0	(13.33, 10.34, 14.52, 38.31)	(34.21, 27.78, 36.87, 126.19)	(15, 15, 15, 14)	2

[1] Total Networks: none=810; dconst/rconst=2430; and dfunc/rfunc=7290.
[2] Percentage of total networks that converged with invertible Hessian matrices.
[3] Percentage of total networks that converged with invertible Hessian matrices that had χ^2 p-values greater than 0.05.
[4] IQR - Interquartile range.

Table J.2: Results of DM model fits when true dispersion matches modelled dispersion (Network Size 78×40)

True		Converge	Good Fit	Percent Relative Bias of β		Coverage of 95% Wald CI's	
Disp.[1]	Method	(%)[2]	(%)[3]	Median	IQR[4]	Individual	Family
none	BFGS	100	93	(0.29, 0.21, 0.27, 1.31)	(0.95, 0.65, 0.84, 5.28)	(89, 95, 94, 94)	74
	NM	100	93	(0.30, 0.21, 0.27, 1.39)	(0.97, 0.66, 0.84, 5.26)	(89, 95, 94, 94)	75
	MNR	100	93	(0.29, 0.22, 0.27, 1.37)	(0.95, 0.66, 0.84, 5.02)	(89, 96, 94, 94)	75
dconst	BFGS	63	84	(0.58, 0.37, 0.41, 2.99)	(1.66, 1.26, 1.14, 11.46)	(94, 96, 94, 94)	51
	NM	100	82	(0.76, 0.49, 0.55, 3.56)	(2.25, 1.58, 1.54, 13.92)	(94, 96, 94, 94)	81
	MNR	97	82	(0.76, 0.48, 0.55, 3.57)	(2.21, 1.58, 1.56, 13.89)	(95, 96, 94, 94)	79
rconst	BFGS	99	72	(7.35, 5.10, 5.87, 33.86)	(17.77, 12.75, 14.07, 102.70)	(94, 95, 95, 96)	81
	NM	99	72	(7.36, 5.03, 5.87, 34.51)	(17.66, 12.97, 14.08, 101.59)	(94, 95, 95, 95)	80
	MNR	99	72	(7.34, 5.10, 5.87, 33.87)	(17.70, 12.73, 14.07, 103.30)	(94, 95, 95, 96)	81
dfunc	BFGS	51	89	(0.42, 0.28, 0.32, 1.92)	(1.41, 1.16, 1.02, 6.69)	(92, 95, 95, 95)	40
	NM	94	88	(0.57, 0.40, 0.46, 2.44)	(2.02, 1.50, 1.47, 8.09)	(93, 95, 95, 95)	75
	MNR	84	87	(0.55, 0.39, 0.45, 2.40)	(1.83, 1.41, 1.38, 7.83)	(93, 95, 95, 95)	67
rfunc	BFGS	57	43	(22.15, 20.38, 20.93, 55.28)	(43.28, 36.75, 37.49, 150.74)	(59, 53, 67, 66)	11
	NM	99	76	(4.96, 4.03, 5.49, 15.45)	(12.56, 9.55, 13.70, 45.82)	(94, 94, 94, 94)	83
	MNR	65	73	(6.23, 5.05, 6.76, 19.35)	(15.46, 12.45, 16.78, 56.46)	(91, 91, 91, 92)	51

[1] Total Networks: none=810; dconst/rconst=2430; and dfunc/rfunc=7290.
[2] Percentage of total networks that converged with invertible Hessian matrices.
[3] Percentage of total networks that converged with invertible Hessian matrices that had χ^2 p-values greater than 0.05.
[4] IQR - Interquartile range.

Table J.3: Results of misspecified DM model fits parameterized in terms of δ (Network Size 78×40)

True		Converge	Good Fit	Percent Relative Bias of β		Coverage of 95% Wald CI's	
Disp.[1]	Method	(%)[2]	(%)[3]	Median	IQR[4]	Individual	Family
Modelled Dispersion: dconst							
none	BFGS	31	98	(0.15, 0.11, 0.13, 0.89)	(0.41, 0.25, 0.33, 3.03)	(86, 93, 92, 94)	23
	NM	73	100	(0.35, 0.23, 0.30, 1.49)	(1.03, 0.69, 1.05, 5.15)	(91, 96, 94, 94)	60
	MNR	66	100	(0.24, 0.17, 0.23, 1.15)	(0.58, 0.48, 0.57, 3.64)	(89, 95, 93, 94)	49
rconst	BFGS	20	89	(31.22, 14.34, 13.76, 45.54)	(23.95, 18.19, 19.64, 124.61)	(48, 66, 70, 91)	7
	NM	100	91	(30.29, 12.98, 13.63, 43.99)	(20.69, 15.15, 16.66, 148.68)	(39, 62, 63, 83)	24
	MNR	57	91	(30.45, 13.89, 13.61, 33.65)	(19.95, 15.76, 14.87, 97.39)	(35, 56, 56, 90)	13
dfunc	BFGS	53	89	(0.41, 0.27, 0.31, 1.93)	(1.40, 1.13, 1.04, 6.45)	(92, 95, 95, 95)	42
	NM	100	88	(0.58, 0.40, 0.46, 2.41)	(2.02, 1.52, 1.50, 7.97)	(93, 95, 95, 95)	80
	MNR	90	87	(0.54, 0.39, 0.43, 2.34)	(1.81, 1.43, 1.37, 7.82)	(92, 95, 95, 95)	71
rfunc	BFGS	37	82	(16.17, 12.99, 26.27, 19.82)	(24.31, 19.55, 30.53, 56.38)	(64, 63, 53, 88)	10
	NM	100	90	(19.06, 16.46, 30.16, 25.05)	(21.97, 19.75, 26.28, 58.86)	(50, 50, 36, 84)	17
	MNR	75	88	(18.10, 15.58, 29.00, 22.30)	(21.20, 19.77, 26.31, 52.41)	(47, 48, 34, 84)	12
Modelled Dispersion: dfunc							
rfunc	BFGS	25	84	(12.57, 9.73, 21.49, 22.37)	(28.05, 20.98, 33.09, 76.84)	(80, 78, 62, 88)	10
	NM	100	91	(11.56, 10.03, 21.92, 19.99)	(18.42, 15.83, 24.58, 56.75)	(66, 65, 43, 87)	26
	MNR	75	91	(10.65, 9.13, 20.71, 17.19)	(17.34, 15.21, 24.24, 44.70)	(64, 63, 42, 89)	19

[1] Total Networks: none=810; dconst/rconst=2430; and dfunc/rfunc=7290.
[2] Percentage of total networks that converged with invertible Hessian matrices.
[3] Percentage of total networks that converged with invertible Hessian matrices that had χ^2 p-values greater than 0.05.
[4] IQR - Interquartile range.

Table J.4: Results of misspecified DM model fits parameterized in terms of ρ (Network Size 78×40)

True		Converge	Good Fit	Percent Relative Bias of β		Coverage of 95% Wald CI's	
Disp.[1]	Method	(%)[2]	(%)[3]	Median	IQR[4]	Individual	Family
Modelled Dispersion: rconst							
none	BFGS	72	99	(0.32, 0.23, 0.29, 1.44)	(1.01, 0.73, 1.00, 4.78)	(89, 96, 94, 94)	54
	NM	76	85	(1.27, 0.53, 0.92, 7.77)	(3.49, 1.71, 2.76, 25.31)	(55, 68, 62, 46)	23
	MNR	32	100	(0.46, 0.30, 0.34, 2.48)	(1.58, 1.39, 1.22, 8.52)	(88, 85, 86, 88)	36
dconst	BFGS	99	13	(1.12, 0.64, 0.70, 4.12)	(3.08, 1.93, 1.92, 16.74)	(81, 92, 90, 94)	62
	NM	99	15	(1.78, 0.82, 1.01, 8.76)	(3.88, 2.32, 2.55, 33.48)	(70, 83, 79, 67)	43
	MNR	88	15	(1.42, 0.81, 0.88, 5.06)	(3.39, 2.18, 2.16, 20.70)	(82, 91, 89, 94)	62
dfunc	BFGS	100	30	(0.82, 0.55, 0.58, 2.80)	(3.05, 2.17, 2.09, 9.07)	(83, 88, 89, 95)	65
	NM	97	29	(1.74, 0.87, 1.14, 7.14)	(5.04, 2.85, 3.19, 19.97)	(66, 74, 72, 66)	38
	MNR	72	28	(1.57, 1.09, 1.12, 4.40)	(4.75, 3.22, 3.03, 12.99)	(82, 84, 86, 95)	60
rfunc	BFGS	100	75	(9.24, 7.67, 9.71, 20.77)	(16.52, 13.64, 17.25, 52.30)	(78, 77, 80, 91)	55
	NM	100	75	(9.24, 7.68, 9.71, 20.86)	(16.61, 13.63, 17.32, 52.54)	(78, 77, 80, 91)	55
	MNR	100	75	(9.23, 7.67, 9.71, 20.82)	(16.53, 13.64, 17.25, 52.31)	(78, 77, 80, 91)	55
Modelled Dispersion: rfunc							
dfunc	BFGS	35	29	(35.11, 46.26, 43.55, 14.04)	(54.55, 38.28, 43.75, 111.06)	(21, 24, 17, 72)	2
	NM	95	29	(0.81, 0.54, 0.57, 2.84)	(3.04, 2.12, 2.04, 9.20)	(99, 98, 99, 99)	91
	MNR	21	34	(2.76, 2.66, 1.88, 5.75)	(8.24, 7.74, 4.69, 18.34)	(96, 98, 98, 99)	20

[1] Total Networks: none=810; dconst/rconst=2430; and dfunc/rfunc=7290.
[2] Percentage of total networks that converged with invertible Hessian matrices.
[3] Percentage of total networks that converged with invertible Hessian matrices that had χ^2 p-values greater than 0.05.
[4] IQR - Interquartile range.

Appendix K

Regularization for the grouped Dirichlet-multinomial regression model: Additional Results Tables

This appendix contains the additional tables summarizing the results of the simulation study as follows:

K.1 - Simulation results for Scenario 1 (Fixed K), Case 1a ($K = 10$), No Dispersion Model.

K.2 - Simulation results for Scenario 1 (Fixed K), Case 1a ($K = 10$), Constant Dispersion Model.

K.3 - Simulation results for Scenario 1 (Fixed K), Case 1a ($K = 10$), Constant Intra-correlation Model.

K.4 - Simulation results for Scenario 1 (Fixed K), Case 1c ($K = 30$), No Dispersion Model.

K.5 - Simulation results for Scenario 1 (Fixed K), Case 1c ($K = 30$), Constant Dispersion Model.

K.6 - Simulation results for Scenario 1 (Fixed K), Case 1c ($K = 30$), Constant Intra-correlation Model.

Table K.1: Simulation results for Scenario 1 (Fixed K), Case 1a ($K = 10$), No Dispersion Model*

		AIC						BIC					
N	Method	Under	Correct	Over	TP	FN	MMSE	Under	Correct	Over	TP	FN	MMSE
250	lasso	0	0	100	0.21	0	0.85	0	0	100	1.06	0	0.82
	alasso ($\tilde{\gamma} = 1$)	0	6	94	4.39	0	0.65	0	32	68	6.58	0	0.44
	alasso ($\tilde{\gamma} = 2$)	0	25	75	6.25	0	0.52	0	72	28	7.63	0	0.32
	alasso ($\tilde{\gamma} = 3$)	0	31	69	6.69	0	0.46	0	86	14	7.83	0	0.29
1000	lasso	0	0	100	0.07	0	0.17	0	0	100	0.47	0	0.17
	alasso ($\tilde{\gamma} = 1$)	0	8	92	4.69	0	0.13	0	52	48	7.29	0	0.07
	alasso ($\tilde{\gamma} = 2$)	0	41	59	6.85	0	0.09	0	81	19	7.8	0	0.06
	alasso ($\tilde{\gamma} = 3$)	0	61	39	7.52	0	0.07	0	92	8	7.92	0	0.05
3000	lasso	0	0	100	0.32	0	0.05	0	0	100	0.95	0	0.05
	alasso ($\tilde{\gamma} = 1$)	0	11	89	5.26	0	0.04	0	64	36	7.56	0	0.03
	alasso ($\tilde{\gamma} = 2$)	0	38	62	6.78	0	0.03	0	98	2	7.98	0	0.02
	alasso ($\tilde{\gamma} = 3$)	0	86	14	7.86	0	0.02	0	99	1	7.99	0	0.02

* alasso - adaptive lasso; Under - percentage of under-fit models; Correct - percentage of correct fit models; Over - percentage of over-fit models; TP - mean number of true positives; FN - mean number of false negatives; MMSE - mean of mean squared errors ($\times 100$).

Table K.2: Simulation results for Scenario 1 (Fixed K), Case 1a ($K = 10$), Constant Dispersion Model*

		AIC						BIC					
N	Method	Under	Correct	Over	TP	FN	MMSE	Under	Correct	Over	TP	FN	MMSE
$\delta = 2$													
250	lasso	0	0	100	0.24	0	2.65	0	1	99	1.95	0	2.43
	alasso ($\tilde{\gamma} = 1$)	0	3	97	3.96	0.01	2.28	0	23	77	6.33	0.01	1.62
	alasso ($\tilde{\gamma} = 2$)	0	14	86	5.71	0.01	1.97	1	56	43	7.33	0.05	1.41
	alasso ($\tilde{\gamma} = 3$)	0	24	76	6.44	0.02	1.77	4	74	22	7.69	0.06	1.25
1000	lasso	0	0	100	0.08	0	0.55	0	0	100	0.5	0	0.54
	alasso ($\tilde{\gamma} = 1$)	0	3	97	4.04	0	0.44	0	26	74	6.7	0	0.29
	alasso ($\tilde{\gamma} = 2$)	0	23	77	6.32	0	0.36	0	85	15	7.84	0	0.21
	alasso ($\tilde{\gamma} = 3$)	0	37	63	6.94	0	0.29	0	92	8	7.92	0	0.2
3000	lasso	0	0	100	0.1	0	0.2	0	0	100	0.42	0	0.2
	alasso ($\tilde{\gamma} = 1$)	0	1	99	3.88	0	0.16	0	30	70	6.78	0	0.1
	alasso ($\tilde{\gamma} = 2$)	0	23	77	6.33	0	0.12	0	86	14	7.85	0	0.08
	alasso ($\tilde{\gamma} = 3$)	0	53	47	7.35	0	0.09	0	93	7	7.93	0	0.07
$\delta = 6$													
250	lasso	0	0	100	0.35	0	4.1	0	1	99	1.81	0	3.84
	alasso ($\tilde{\gamma} = 1$)	0	4	96	3.81	0.01	3.53	4	21	75	6.25	0.08	2.43
	alasso ($\tilde{\gamma} = 2$)	2	13	85	5.66	0.06	3.03	7	42	51	7.16	0.14	2.16
	alasso ($\tilde{\gamma} = 3$)	5	26	69	6.5	0.1	2.68	12	68	20	7.65	0.2	2.1
1000	lasso	0	0	100	0.09	0	0.88	0	0	100	0.67	0	0.87
	alasso ($\tilde{\gamma} = 1$)	0	1	99	3.69	0	0.76	0	25	75	6.46	0	0.51
	alasso ($\tilde{\gamma} = 2$)	0	20	80	6.11	0	0.62	0	69	31	7.61	0	0.42
	alasso ($\tilde{\gamma} = 3$)	0	33	67	6.81	0	0.54	0	82	18	7.79	0	0.39
3000	lasso	0	0	100	0.18	0	0.26	0	0	100	0.35	0	0.26
	alasso ($\tilde{\gamma} = 1$)	0	2	98	3.67	0	0.21	0	34	66	6.78	0	0.12
	alasso ($\tilde{\gamma} = 2$)	0	36	64	6.72	0	0.14	0	85	15	7.8	0	0.09
	alasso ($\tilde{\gamma} = 3$)	0	59	41	7.41	0	0.11	0	95	5	7.95	0	0.08

* alasso - adaptive lasso; Under - percentage of under-fit models; Correct - percentage of correct fit models; Over - percentage of over-fit models; TP - mean number of true positives; FN - mean number of false negatives; MMSE - mean of mean squared errors ($\times 100$).

Table K.3: Simulation results for Scenario 1 (Fixed K), Case 1a ($K = 20$), Constant Intra-correlation Model*

		AIC						BIC					
N	Method	Under	Correct	Over	TP	FN	MMSE	Under	Correct	Over	TP	FN	MMSE
$\rho = 0.2$													
250	lasso	0	0	100	0.49	0.01	4.21	4	16	80	4.41	0.09	3.38
	alasso ($\tilde{\gamma} = 1$)	0	4	96	3.95	0.04	3.41	14	22	64	6.76	0.2	2.24
	alasso ($\tilde{\gamma} = 2$)	2	10	88	5.66	0.08	2.99	18	31	51	7.19	0.26	2.22
	alasso ($\tilde{\gamma} = 3$)	2	12	86	6.09	0.09	2.85	26	43	31	7.54	0.32	2.09
1000	lasso	0	0	100	0.39	0	1.1	0	1	99	2.7	0	0.96
	alasso ($\tilde{\gamma} = 1$)	0	3	97	4.7	0	0.83	1	36	63	6.9	0.01	0.55
	alasso ($\tilde{\gamma} = 2$)	0	15	85	6.29	0	0.68	2	69	29	7.67	0.02	0.43
	alasso ($\tilde{\gamma} = 3$)	0	32	68	6.73	0	0.64	4	82	14	7.85	0.04	0.41
3000	lasso	0	0	100	0.37	0	0.47	0	0	100	1.3	0	0.46
	alasso ($\tilde{\gamma} = 1$)	0	2	98	4.33	0	0.37	0	40	60	7.19	0	0.21
	alasso ($\tilde{\gamma} = 2$)	0	13	87	6.22	0	0.3	0	83	17	7.83	0	0.17
	alasso ($\tilde{\gamma} = 3$)	0	26	74	6.9	0	0.25	0	93	7	7.93	0	0.16
$\rho = 0.5$													
250	lasso	0	3	97	1.09	0.05	8.7	43	23	34	7.06	0.55	6.1
	alasso ($\tilde{\gamma} = 1$)	2	9	89	4.36	0.16	6.96	32	31	37	7.04	0.5	4.58
	alasso ($\tilde{\gamma} = 2$)	6	15	79	5.47	0.28	6.44	37	33	30	7.33	0.54	4.53
	alasso ($\tilde{\gamma} = 3$)	7	23	70	6.1	0.31	6.04	40	33	27	7.4	0.58	4.64
1000	lasso	0	0	100	0.36	0.02	2.84	7	17	76	5.09	0.12	2.31
	alasso ($\tilde{\gamma} = 1$)	1	7	92	4.54	0.04	2.25	12	39	49	7.17	0.18	1.55
	alasso ($\tilde{\gamma} = 2$)	2	19	79	6.01	0.07	2	20	52	28	7.6	0.24	1.53
	alasso ($\tilde{\gamma} = 3$)	2	27	71	6.54	0.08	1.85	29	53	18	7.77	0.31	1.57
3000	lasso	0	0	100	0.26	0	1.16	0	5	95	3.16	0	0.98
	alasso ($\tilde{\gamma} = 1$)	0	6	94	4.68	0	0.88	0	43	57	7.12	0	0.55
	alasso ($\tilde{\gamma} = 2$)	0	21	79	6.22	0	0.76	1	78	21	7.76	0.01	0.46
	alasso ($\tilde{\gamma} = 3$)	0	30	70	6.7	0	0.7	2	87	11	7.89	0.02	0.44

* alasso - adaptive lasso; Under - percentage of under-fit models; Correct - percentage of correct fit models; Over - percentage of over-fit models; TP - mean number of true positives; FN - mean number of false negatives; MMSE - mean of mean squared errors ($\times 100$).

Table K.4: Simulation results for Scenario 1 (Fixed K), Case 1c ($K = 30$), No Dispersion Model*

		AIC						BIC					
N	Method	Under	Correct	Over	TP	FN	MMSE	Under	Correct	Over	TP	FN	MMSE
250	lasso	0	0	100	0.44	0	1.37	0	0	100	3.14	0	1.23
	alasso ($\tilde{\gamma} = 1$)	0	0	100	12.97	0	0.9	0	1	99	19.12	0	0.53
	alasso ($\tilde{\gamma} = 2$)	0	1	99	18.1	0	0.68	0	28	72	22.36	0	0.39
	alasso ($\tilde{\gamma} = 3$)	0	3	97	19.96	0	0.61	0	51	49	23.19	0	0.35
1000	lasso	0	0	100	0.13	0	0.2	0	0	100	0.67	0	0.2
	alasso ($\tilde{\gamma} = 1$)	0	0	100	13.26	0	0.14	0	6	94	20.9	0	0.08
	alasso ($\tilde{\gamma} = 2$)	0	5	95	19.72	0	0.1	0	67	33	23.58	0	0.06
	alasso ($\tilde{\gamma} = 3$)	0	22	78	22.29	0	0.07	0	88	12	23.87	0	0.05
3000	lasso	0	0	100	0.15	0	0.06	0	0	100	0.89	0	0.06
	alasso ($\tilde{\gamma} = 1$)	0	0	100	12.66	0	0.04	0	11	89	21.16	0	0.03
	alasso ($\tilde{\gamma} = 2$)	0	4	96	19.74	0	0.03	0	83	17	23.82	0	0.02
	alasso ($\tilde{\gamma} = 3$)	0	65	35	23.53	0	0.02	0	98	2	23.98	0	0.02

* alasso - adaptive lasso; Under - percentage of under-fit models; Correct - percentage of correct fit models; Over - percentage of over-fit models; TP - mean number of true positives; FN - mean number of false negatives; MMSE - mean of mean squared errors ($\times 100$).

Table K.5: Simulation results for Scenario 1 (Fixed K), Case 1c ($K = 30$), Constant Dispersion Model*

		AIC						BIC					
N	Method	Under	Correct	Over	TP	FN	MMSE	Under	Correct	Over	TP	FN	MMSE
$\delta = 2$													
250	lasso	0	0	100	0.54	0	38.36	0	0	100	5.12	0	36.36
	alasso ($\tilde{\gamma} = 1$)	0	0	100	10.41	0.01	36.52	1	0	99	18.61	0.1	33.4
	alasso ($\tilde{\gamma} = 2$)	0	0	100	15.74	0.05	35.59	3	9	88	21.1	0.11	33.48
	alasso ($\tilde{\gamma} = 3$)	0	3	97	17.86	0.08	35.12	8	24	68	22.36	0.18	33.2
1000	lasso	0	0	100	0.1	0	31.2	0	0	100	0.82	0	31.15
	alasso ($\tilde{\gamma} = 1$)	0	0	100	11.69	0	30.9	0	9	91	20.14	0	30.41
	alasso ($\tilde{\gamma} = 2$)	0	5	95	18.13	0	30.71	0	42	58	23.13	0	30.38
	alasso ($\tilde{\gamma} = 3$)	0	16	84	20.87	0	30.57	0	79	21	23.76	0	30.34
3000	lasso	0	0	100	0.08	0	30.33	0	0	100	0.44	0	30.32
	alasso ($\tilde{\gamma} = 1$)	0	0	100	11.14	0	30.24	0	12	88	21.07	0	30.09
	alasso ($\tilde{\gamma} = 2$)	0	10	90	19.75	0	30.18	0	64	36	23.51	0	30.1
	alasso ($\tilde{\gamma} = 3$)	0	34	66	22.41	0	30.12	0	94	6	23.94	0	30.08
$\delta = 6$													
250	lasso	0	0	100	0.46	0.01	44.33	0	0	100	8.53	0.17	37.75
	alasso ($\tilde{\gamma} = 1$)	0	0	100	8.83	0.11	41.97	1	5	94	17.46	0.32	36
	alasso ($\tilde{\gamma} = 2$)	0	0	100	13.81	0.21	40.4	6	11	83	19.71	0.46	36.14
	alasso ($\tilde{\gamma} = 3$)	0	1	99	16.04	0.27	39.78	9	20	71	21.11	0.6	35.71
1000	lasso	0	0	100	0.25	0	31.53	0	0	100	1.7	0	31.33
	alasso ($\tilde{\gamma} = 1$)	0	0	100	11.91	0	31.13	0	4	96	19.65	0	30.37
	alasso ($\tilde{\gamma} = 2$)	0	4	96	17.77	0	30.89	1	29	70	22.53	0.01	30.36
	alasso ($\tilde{\gamma} = 3$)	0	7	93	19.92	0	30.72	2	65	33	23.54	0.02	30.29
3000	lasso	0	0	100	0.08	0	30.64	0	0	100	0.39	0	30.63
	alasso ($\tilde{\gamma} = 1$)	0	0	100	8.37	0	30.54	0	7	93	20	0	30.2
	alasso ($\tilde{\gamma} = 2$)	0	4	96	18.53	0	30.38	0	53	47	23.31	0	30.23
	alasso ($\tilde{\gamma} = 3$)	0	14	86	21.26	0	30.31	0	82	18	23.81	0	30.2

* alasso - adaptive lasso; Under - percentage of under-fit models; Correct - percentage of correct fit models; Over - percentage of over-fit models; TP - mean number of true positives; FN - mean number of false negatives; MMSE - mean of mean squared errors ($\times 100$).

Table K.6: Simulation results for Scenario 1 (Fixed K), Case 1c ($K = 30$), Constant Intra-correlation Model*

		AIC						BIC					
N	Method	Under	Correct	Over	TP	FN	MMSE	Under	Correct	Over	TP	FN	MMSE
$\rho = 0.2$													
250	lasso	0	0	100	0.6	0	44.49	20	10	70	18.15	0.64	29.56
	alasso ($\tilde{\gamma} = 1$)	0	0	100	10.06	0.18	41.3	4	10	86	19.04	0.44	35.26
	alasso ($\tilde{\gamma} = 2$)	0	1	99	14.37	0.29	40.41	10	14	76	20.91	0.53	35.91
	alasso ($\tilde{\gamma} = 3$)	0	1	99	16.54	0.34	39.79	12	15	73	21.71	0.58	35.89
1000	lasso	0	0	100	0.65	0	32.61	0	1	99	8.03	0	30.62
	alasso ($\tilde{\gamma} = 1$)	0	0	100	12.86	0	31.77	1	5	94	20.3	0.01	30.55
	alasso ($\tilde{\gamma} = 2$)	0	1	99	17.49	0	31.63	2	23	75	22.48	0.03	30.82
	alasso ($\tilde{\gamma} = 3$)	0	2	98	19.25	0	31.53	7	40	53	23.17	0.09	30.84
3000	lasso	0	0	100	0.34	0	30.8	0	0	100	2.95	0	30.49
	alasso ($\tilde{\gamma} = 1$)	0	0	100	11.82	0	30.41	0	10	90	21.21	0	29.87
	alasso ($\tilde{\gamma} = 2$)	0	6	94	19.07	0	30.32	0	47	53	23.08	0	30.04
	alasso ($\tilde{\gamma} = 3$)	0	11	89	20.71	0	30.29	0	76	24	23.68	0	30.03
$\rho = 0.5$													
250	lasso	1	0	99	1.02	0.05	70.43	82	4	14	24.08	3.91	23.04
	alasso ($\tilde{\gamma} = 1$)	0	0	100	8.08	0.29	63.68	35	11	54	20.82	2.13	34.75
	alasso ($\tilde{\gamma} = 2$)	0	1	99	12.43	0.46	60.66	15	8	77	19.48	1.34	42.11
	alasso ($\tilde{\gamma} = 3$)	2	0	98	14.2	0.62	58.29	12	6	82	19.24	1.22	44.96
1000	lasso	0	0	100	0.62	0.01	35.57	3	8	89	18.57	0.3	25.73
	alasso ($\tilde{\gamma} = 1$)	0	0	100	11.48	0.09	33.93	6	3	91	20.42	0.29	30.77
	alasso ($\tilde{\gamma} = 2$)	0	0	100	16.16	0.14	33.62	11	22	67	22.11	0.39	31.71
	alasso ($\tilde{\gamma} = 3$)	0	1	99	18.1	0.15	33.41	21	30	49	22.93	0.48	31.85
3000	lasso	0	0	100	0.56	0	31.84	0	0	100	10.67	0.01	29.22
	alasso ($\tilde{\gamma} = 1$)	0	0	100	12.23	0	31.18	1	6	93	20.99	0.03	30.1
	alasso ($\tilde{\gamma} = 2$)	0	0	100	18.02	0.01	31	3	34	63	22.86	0.05	30.39
	alasso ($\tilde{\gamma} = 3$)	0	2	98	19.62	0.02	30.92	8	62	30	23.63	0.11	30.39

* alasso - adaptive lasso; Under - percentage of under-fit models; Correct - percentage of correct fit models; Over - percentage of over-fit models; TP - mean number of true positives; FN - mean number of false negatives; MMSE - mean of mean squared errors ($\times 100$).