**University of São Paulo
“Luiz de Queiroz” College of Agriculture**

Statistical modelling of data from insect studies

Rafael de Andrade Moral

Thesis presented to obtain the degree of Doctor in Science.
Area: Statistics and Agricultural Experimentation

**Piracicaba
2017**

Rafael de Andrade Moral
Bachelor in Biology

Statistical modelling of data from insect studies

Advisor:

Prof^a Dr^a **CLARICE GARCIA BORGES DEMÉTRIO**

Thesis presented to obtain the degree of Doctor in Science.

Area: Statistics and Agricultural Experimentation

Piracicaba
2017

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Moral, Rafael de Andrade

Statistical modelling of data from insect studies/ Rafael de Andrade Moral.
-- Piracicaba, 2017 .
133 p.

Tese (Doutorado) -- USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Análise de sobrevivência 2. Avaliação da qualidade do ajuste 3.
Estatística ecológica 4. Modelos conjuntos 5. Modelos mistos 6. Software R
. I. Título.

To Mum and Dad

ACKNOWLEDGMENTS

I would like to thank my Mum and Dad, and all my family, for all the love and support throughout the years, ever since I was born.

I would also like to thank my supervisors, Clarice Demétrio, who began supervising me since 2008, when I was a scientific initiation student, and who has always believed in my potential. John Hinde, who supervised me both during my Master's and PhD, and with whom I have learned so much, he who has always been keen to help me in every sense he could. Finally, Wesley Godoy, who has been my co-supervisor since 2009, and who helped to make my transition from a biologist to a statistician so smooth. In fact, I don't really know whether I am a biologist, ecologist, or statistician. Maybe a bit of each.

This work would also not be possible without the help of my friend Carolina Reigada, one of the most brilliant ecologists I know, who did not let me make wrong statements when interpreting statistical models applied to ecology.

I am also very thankful to Richard and Célia Wall, who taught me the English I so much needed throughout my academic life and who were always very helpful whenever I had any questions. They are family to me.

I could not forget to mention my friend Idemauro Lara, who helped me so much throughout the year I was abroad, and who became one of the few true friends a person can have in his life. And also my friend Thiago Oliveira, with whom I shared my home and who has helped me through a very tough time in my life.

Finally, many thanks to FAPESP for providing financial support in Brazil (proc. no. 2014/03310-3) and in Ireland (proc. no. 2014/12903-8). FAPESP has funded me since my scientific initiation, and also throughout my Master's. I would not have been able to conduct my studies abroad without the funding they provided me and I cannot stress enough how thankful I am.

“An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question.”

John Wilder Tukey

“To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.”

Ronald Aylmer Fisher

“Data is not information. Information is not knowledge. And knowledge is certainly not wisdom.”

Clifford Paul Stoll

CONTENTS

RESUMO	8
ABSTRACT	9
1 Introduction	11
References	12
2 Case studies	15
2.1 Introduction	15
2.2 Parasitism data	15
2.3 Mite abundance data	16
2.4 Bald eagle data	16
2.5 Corn and progeny data	16
2.6 Orange data	17
2.7 Leukemia data	18
2.8 Insect attack data	18
References	18
3 Accelerated failure rate mixed models: Applications in insect ecology	21
Abstract	21
3.1 Introduction	21
3.2 Case-study	22
3.3 Background	23
3.3.1 Survival analysis	23
3.3.2 Exploratory analysis	24
3.3.3 Log-rank test	24
3.4 Modelling	25
3.4.1 Exponential model	25
3.4.2 Weibull model	26
3.4.3 Exponentiated family of distributions	26
3.4.4 Maximum likelihood estimation	27
3.4.5 Estimation including random effects	28
3.4.6 Model selection	29
3.4.7 Goodness-of-fit assessment	30
3.4.8 Simulation studies	31
3.5 Results	32
3.5.1 Exploratory analysis	32
3.5.2 Exponential model fit	32
3.5.3 Weibull model fit	35
3.5.4 Exponentiated-Weibull model fit	37
3.5.5 Including day effects as random in the exponentiated-Weibull model	39
3.5.6 Simulation studies	40
3.6 Discussion	40
References	43
4 Models for jointly estimating abundances of two unmarked site-associated species subject to im- perfect detection	47
Abstract	47
4.1 Introduction	47
4.2 Model formulation and assumptions	48

4.2.1	Incorporating covariates	51
4.2.2	Correlation between species	51
4.2.3	Simulation studies	51
4.3	Case study: Mite data	52
4.4	Case study: Bald eagle data	53
4.5	Results	53
4.5.1	Model performance	53
4.5.2	Mite data analysis	54
4.5.3	Bald eagle data analysis	54
4.6	Discussion	56
	References	60
5	Half-normal plots and overdispersed models in R: the hnp package	63
	Abstract	63
5.1	Introduction	63
5.2	Generalized linear models and overdispersion	64
5.3	Half-normal plots with simulated envelopes	66
5.3.1	Implemented model classes	67
5.3.2	Simulation procedures	68
5.3.3	New class implementation	68
5.4	Examples	68
5.4.1	Overdispersed proportion data	68
5.4.2	Overdispersed count data	70
5.4.3	Implementing new model classes	72
5.5	Discussion	79
	References	80
6	Assessing goodness-of-fit for bivariate models using residual plots with simulation polygons	83
	Abstract	83
6.1	Introduction	83
6.2	A bivariate residual plot for assessing goodness-of-fit	84
6.3	Applications	85
6.3.1	Bivariate normal responses	85
6.3.2	Bivariate Poisson responses – attack data	87
6.3.3	Longitudinal counts of two species on different sites	90
6.4	Discussion	93
	References	95
7	Final Considerations	97
	Appendix	99
	Appendix A: Optimization tools and techniques	99
	Appendix B: Derivations for Chapter 3	109
	Appendix C: Derivations for Chapter 4 and simulation studies	112
	Appendix D: Review on techniques used on Chapter 6	117
	Appendix E: R packages from the thesis	122
	Supplementary Material	125
	Supplement S1: Dataset for Chapter 3	125
	Supplement S2: Datasets for Chapter 4	128
	Supplement S3: Dataset for Chapter 6	133

RESUMO

Modelagem estatística de dados provenientes de estudos em entomologia

Dados provenientes de estudos com insetos podem apresentar características diferentes. Respostas univariadas podem ser analisadas utilizando-se modelos lineares generalizados (dados contínuos e discretos), modelos de análise de sobrevivência (dados de tempo até ocorrência de um evento), modelos de efeitos mistos (dados longitudinais), dentre outros métodos. Esses modelos podem ser usados para analisar dados provenientes de experimentos que avaliam processos ecológicos complexos, como competição e predação. Nesse sentido, ferramentas computacionais são úteis para pesquisadores em diversos campos, por exemplo, biologia e fisiologia de insetos, ecologia aplicada e controle biológico. Utilizando diferentes conjuntos de dados entomológicos como motivação, assim como outros tipos de dados para ilustrar os métodos, este trabalho teve como objetivos desenvolver novos modelos e ferramentas para avaliar a qualidade do ajuste. Foram propostos modelos de tempo de vida acelerado mistos, com modelagem simultânea dos parâmetros de locação e de escala com regressores, para analisar dados de tempo até ataque de um experimento que avaliou escolha de predadores. Foram utilizados modelos exponencial, Weibull e Weibull-exponenciado, e a qualidade do ajuste foi avaliada utilizando gráficos meio-normais com envelope de simulação. Esses gráficos são o assunto de um Capítulo inteiro sobre um pacote para o software R, chamado **hnp**, desenvolvido para implementá-los. Foram utilizados conjuntos de dados de diferentes tipos de experimentos para ilustrar o uso desses gráficos e do pacote. Uma extensão bivariada para os modelos chamados “N-mixture” foi proposta para analisar dados longitudinais de contagem para duas espécies pertencentes à mesma teia trófica, que podem interagir direta e indiretamente, e conjuntos de dados provenientes de estudos ecológicos são usados para ilustrar a abordagem. Uma vantagem dessa estratégia de modelagem é a obtenção de um coeficiente de correlação assimétrico, que pode ser utilizado por ecologistas para inferir acerca do grau de associação entre espécies. O pacote **jointNmix** foi desenvolvido para implementar o processo de estimação para esses modelos. Finalmente, foi proposta uma ferramenta de avaliação de qualidade do ajuste para modelos bivariados, análoga ao gráfico meio-normal com envelope de simulação, e a metodologia é ilustrada com dados simulados e dados de competição de insetos. Essa ferramenta está também implementada em um pacote para o R, chamado **bivrp**. Todo o software desenvolvido nesta tese está disponível, gratuitamente, na *Comprehensive R Archive Network* (CRAN).

Palavras-chave: Análise de sobrevivência; Avaliação de qualidade do ajuste; Estatística ecológica; Modelos conjuntos; Modelos mistos; Software R

ABSTRACT

Statistical modelling of data from insect studies

Data from insect studies may present different features. Univariate responses may be analyzed using generalized linear models (continuous and discrete data), survival models (time until event data), mixed effects models (longitudinal data), among other methods. These models may be used to analyse data from experiments which assess complex ecological processes, such as competition and predation. In that sense, computational tools are useful for researchers in several fields, e.g., insect biology and physiology, applied ecology and biological control. Using different datasets from entomology as motivation, as well as other types of datasets for illustration purposes, this work intended to develop new modelling frameworks and goodness-of-fit assessment tools. We propose accelerated failure rate mixed models with simultaneous location and scale modelling with regressors to analyse time-until-attack data from a choice test experiment. We use the exponential, Weibull and exponentiated-Weibull models, and assess goodness-of-fit using half-normal plots with simulation envelopes. These plots are the subject of an entire Chapter on an R package, called **hnp**, developed to implement them. We use datasets from different types of experiments to illustrate the use of these plots and the package. A bivariate extension to the N-mixture modelling framework is proposed to analyse longitudinal count data for two species from the same food web that may interact directly or indirectly, and example datasets from ecological studies are used. An advantage of this modelling framework is the computation of an asymmetric correlation coefficient, which may be used by ecologists to study the degree of association between species. The **jointNmix** R package was also developed to implement the estimation process for these models. Finally, we propose a goodness-of-fit assessment tool for bivariate models, analogous to the half-normal plot with a simulation envelope, and illustrate the approach with simulated data and insect competition data. This tool is also implemented in an R package, called **bivrp**. All software developed in this thesis is made available freely on the Comprehensive R Archive Network.

Keywords: Ecological statistics, Goodness-of-fit assessment, Joint models, Mixed models, R software, Survival analysis

1 INTRODUCTION

Entomological studies, among other research topics, include insect biology, physiology, taxonomy, behaviour, ecology and interdisciplinary research related to biological control (Levins and Wilson, 1980). The latter focus on agroecosystems, which are ecosystems where one or more crops are planted. These ecological systems present simplified food webs given their less heterogeneous structure when compared to a natural ecosystem, e.g., a tropical rainforest. However, multiple and complex trophic relations among two or more species may occur in agroecosystems, such as intraguild predation, parasitism, superparasitism and hyperparasitism (Levins and Wilson, 1980; Rosenheim et al., 1985; Thies et al., 2011).

Insects represent the most diverse taxonomic group and are able to affect the agroecosystems they inhabit (Van Emden and Williams, 1974; Thies et al., 2011). Insects which live in agroecosystems and cause economic damage to producers are referred to as pests. Species which feed on pests, parasitise them or ecologically interact with them where the result of this interaction is a smaller pest population which consequently reduces economic damage, are called natural enemies. When conducting biological control assays it is important to characterize species of interest and their ecological relations so as to develop integrated pest management strategies, which are a set of techniques used to promote pest suppression without relying solely on pesticide application (Kogan, 1998).

Data obtained in entomological studies may present different features. Continuous response variables (such as weight of egg masses, adult and immature insect weights and lengths, weight gain on different diets, longevity, feeding time, etc) and discrete ones that consist of counts (number of eggs, number of matings, number of consumed prey, number of attacks, number of damaged crops, etc) and proportions (mortality, egg viability, prey preference, oviposition substrate preference, proportion of insects which reached a given life stage, etc) are common outcomes of these studies. These data may be collected only once or repeatedly over time, the latter being characterized as a longitudinal study (Verbeke and Molenberghs, 2000). Such studies allow the researcher to observe the behaviour of different biological processes such as growth and survival over time. Studies assessing time until occurrence of a given phenomenon (such as predation, parasitism and death) are also important in entomology and data from these studies are analysed using survival time to event models. In dose-response experiments, on the other hand, interest lies within assessing the effects of different reagents on pest or natural enemy mortality to determine a lethal dose / concentration of this reagent to be applied on the crops.

Studies on insect population dynamics also assess ecological relations among the species. Competition models, predation, parasitism and intraguild predation models, as well as models which consider resource consumption among other factors of interest, may be used to study the biological processes acting in concert in the system.

When analysing data from different entomological studies, such as the previously described ones, finding the best way to interpret results may be challenging. Data transformation is a commonly used variance stabilising technique when analysing discrete response variables (Bolker et al., 2009). However studies show that taking the natural logarithm of counts or transforming proportions using the arcsine are not the most adequate way of analysing these data (O'Hara and Kotze, 2010; Warton and Hui, 2011). Depending on the mean and sample size, these transformations may result in precision loss when estimating the effects of the fitted model and hence the obtained conclusions are less reliable.

Another strategy is to consider longitudinal studies as split plots over time and to use analysis of variance models. However the sphericity condition of the variance-covariance matrix must be met to analyse these data in such a way. Moreover mixed-modelling techniques are widely available nowadays using different software, such as the R packages `nlme` (Pinheiro et al., 2015), `lme4` (Bates et al., 2015) and SAS procedures MIXED and NLMIXED (Liu and Yu, 2008). Because the correlation between responses obtained on the same individual are taken into account, linear and nonlinear mixed effects models may

be used to analyse longitudinal data (Laird and Ware, 1982; Pinheiro and Bates, 2000).

When analysing discrete data, discrete distributions should be used, such as the Poisson distribution for count data and the binomial distribution for proportion data. Even so, one must be careful when fitting these models and verify if there is overdispersion, a common phenomenon in entomological experiments (Demétrio et al., 2014). Fitting inadequate models to overdispersed data may result in underestimation of standard errors and incorrect interpretation of parameter significance. Random effects may also be included in the linear predictor to account for overdispersion, with the advantage of being biologically interpretable in different experiments (Hinde and Demétrio, 1998).

Interest frequently lies in characterizing and assessing complex biological processes which involve more than one response variable. In competition experiments, assessing which competitor is best adapted in the system may involve not only observing the amount of consumed resource but also the occurrence of interference competition, attacks between competitors, time taken to consume the resource, among other variables. A joint modelling approach may be developed and used for this type of research.

Besides developing statistical methods, it is important to implement computational tools to make these methods easily usable. On the Comprehensive R Archive Network (CRAN) several packages for generalized linear modelling, survival modelling and mixed modelling are available for the open-source software R (R Core Team, 2017). However, there are not specific packages for the analysis of entomological data. The development of specific, user-friendly functions, together with package vignettes would be a great utility for researchers of the field.

The data sets obtained from the case-studies that will be explored in the development of the thesis are described in Chapter 2. In Chapter 3, accelerated failure rate models are proposed to analyse data of a preference experiment with different natural enemies and one pest, modelling the location and scale parameters simultaneously. Then, a random effect is included to reflect the experimental design. We present a joint modelling framework to estimate abundance in longitudinal studies involving two co-occurring species in Chapter 4, and illustrate with two examples: one on parasitoid wasps and another with mites and springtails. The fitting algorithms for these models are implemented as the `jointNmix` package on CRAN (Moral et al., 2016). In Chapter 5, we present a review and detailed tutorial on producing half-normal plots with simulated envelopes in the R software, implemented as the `hnp` package on CRAN (Moral et al., 2017), using various examples. A bivariate extension to diagnostic plots with simulated envelopes is developed in Chapter 6, motivated by the joint models proposed in Chapter 3, and the approach is illustrated using simulated examples and real datasets. The framework is implemented as the `bivvrp` package, also available on CRAN (Moral et al., 2016). In Chapter 7, we make our final considerations. Finally, in the Appendix section, we present a review on optimization methods and the mathematical derivations needed for the development of the thesis and the datasets are made available as Supplementary Material. A brief description of the packages is also given in the appendix. Moreover, all R code used in the development of this thesis is available upon request to the author.

References

ATKINSON, A. C. (1985) **Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis**. Oxford: Clarendon. 282 p.

BATES, D.; MAECHLER, M.; BOLKER, B.; WALKER, S. (2015) `lme4`: Linear mixed-effects models using Eigen and S4. R package version 1.1-7, Available at:
<<http://CRAN.R-project.org/package=lme4>>

- BOLKER, B. M.; BROOKS, M. E.; CLARK, C. J.; GEANGE, S.W.; POULSEN, J. R.; STEVENS, M. H. H.; WHITE, J. S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. **Trends in Ecology and Evolution**, v. 24, p. 127–135.
- CASTRO, T. R.; WEKESA, V. W.; MORAL, R. A.; DEMÉTRIO, C. G. B.; DELALIBERA Jr., I.; KLINGEN, I. (2013) The effects of photoperiod and light intensity on the sporulation of Brazilian and Norwegian isolates of *Neozygites floridana*. **Journal of Invertebrate Pathology**, v. 114, p. 230–233.
- HINDE J.; DEMÉTRIO C. G. B. (1998) Overdispersion: models and estimation. **Computational Statistics & Data Analysis**, v. 27, p. 151–170.
- KOGAN, M. (1998) Integrated pest management historical perspectives and contemporary developments. **Annual Review of Entomology**, v. 43, 243–270.
- LAIRD, N. M.; WARE, J. H. (1982) Random effects models for longitudinal data. **Biometrics**, v.38, p. 963–974.
- LEVINS, R.; WILSON, M.. (1980) Ecological theory and pest management. **Annual Review of Entomology**, v. 25, p. 287–308.
- LIU, L.; YU, Z. (2008) A likelihood reformulation method in non-normal random effects models. **Statistics in Medicine**, v. 27, p. 3105–3124.
- MCCULLAGH, P.; NELDER, J. A. (1989) **Generalized linear models**, London: Chapman and Hall. 511 p.
- MORAL, R. A.; DEMÉTRIO, C. G. B.; HINDE, J. (2016) jointNmix: Joint N-mixture models for associated species. R package version 1.0. Available at: <<http://CRAN.R-project.org/package=jointNmix>>.
- MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. B. (2017) hnp: Half-Normal Plots with Simulation Envelopes. R package version 1.2-3. Available at: <<http://CRAN.R-project.org/package=hnp>>.
- MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. B. (2016) bivrp: Bivariate Diagnostic Plots with Simulation Polygons. R package version 1.0. Available at: <<http://CRAN.R-project.org/package=bivrp>>.
- O'HARA, R. B.; KOTZE, D. J. (2010) Do not log-transform count data. **Methods in Ecology and Evolution**, v. 1, p. 118–122.
- PINHEIRO, J. C.; BATES, D. M. (2000) **Mixed-effects models in S and S-PLUS**. New York: Springer. 528 p.
- PINHEIRO, J.; BATES, D.; DEBROY, S.; SARKAR, D.; R CORE TEAM. (2015) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-121. Available at: <<http://CRAN.R-project.org/package=nlme>>
- R CORE TEAM. (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: <<http://www.R-project.org/>>.
- ROSENHEIM, J. A.; KAYA, H. K.; EHLER, L. E.; MAROIS, J. J.; JAFFEE, B. A. (1995) **Biological Control**, v. 5, p. 303–335.

THIES, C.; HAENKE, S.; SCHERBER, C.; BENGTTSSON, J.; BOMMARCO, R.; CLEMENT, L. W.; CERYNGIER, P.; DENNIS, C.; EMMERSON, M.; GAGIC, V.; HAWRO, V.; LIIRA, J.; WEISSER, W. W.; WINQVIST, C.; TSCHARNTKE, T. (2011) The relationship between agricultural intensification and biological control experimental tests across Europe. **Ecological Applications**, v. 21, p. 2187–2196.

VAN EMDEN, H. F.; WILLIAMS, G. F. (1974) Insect stability and diversity in agro-ecosystems. **Annual Review of Entomology**, v. 19, p. 455–475.

VERBEKE, G.; MOLENBERGHS, G. (2000) **Linear mixed models for longitudinal data**. Verlag: Springer. 569 p.

WARTON, D. I.; HUI, F. K. (2011) The arcsine is asinine: the analysis of proportions in ecology. **Ecology**, v. 92, p. 3–10.

2 CASE STUDIES

2.1 Introduction

Here we present the datasets obtained from the case studies that will be explored throughout this work. We describe the case study and then refer to the statistical methodology that was developed to model the data. The case studies are presented in the order they appear in this thesis. Even though this is a statistical thesis with applications to entomology, some examples come from different areas so as to illustrate the broader applications of the methodology developed for data modelling and goodness-of-fit assessment.

2.2 Parasitism data

A major pest of maize worldwide is the larva of *Spodoptera frugiperda* (Oliveira et al., 2004). Three important natural enemies of this pest are the stinkbug *Podisus nigrispinus* (Oliveira et al., 2002) and the earwig *Euborellia annulipes* (Redoan et al., 2010), both predators, and the parasitoid wasp *Campoletis flavicincta* (Cruz et al., 1999). The latter lays eggs inside *S. frugiperda* larvae that hatch and develop inside them, killing the larva upon reaching the adult stage. This process causes substantial metabolic changes in the larvae which may be perceived by potential predators.

In a biological control context, it is important that predators act in synergy to promote pest suppression (Cardinale et al., 2003). In that sense, a desirable outcome when controlling *S. frugiperda* with multiple natural enemies, such as the three species mentioned above, would be weak competition between predators, and preference for non-parasitised larvae so that the parasitoid's population is also maintained in the system. Information on the time taken by different predator species to attack prey and on prey preference are particularly relevant in this context.

To study the predatory behaviour of the stinkbug and the earwig when given the choice between parasitised and non-parasitised prey, Moral (2014) conducted an experiment in a complete randomized block design with 33 blocks and four treatments in a 2×2 (species and gender) factorial design, i.e., male and female stinkbugs and earwigs which were fasted for 24 hours prior to experiment set up. Each block consisted of four Petri dishes with one predator and two *S. frugiperda* larvae, one of which was previously parasitised by *C. flavicincta* and another which was not. Each experimental unit was observed for one hour and the time (in seconds) taken until the predator first attacked a larva was recorded, as well as which larva the predator chose to effectively consume first (predator preference), see the Supplementary Materials section for the complete dataset.

The experiment took three days to be completed and on each day 11 blocks were installed and observed, so there is reason to believe that the day effect is random. However, the block within day effect should be treated as fixed because only up to three blocks could be observed at the same time. In that sense, predators which were used in the first blocks were fasted for 24 hours prior to experiment installation but predators used in the last blocks had a longer fasting time which could systematically influence their behaviour. The data is also right-censored, as some predators did not attack in one hour of observation.

This time-until-attack data is analysed using different accelerated failure rate models (exponential, Weibull, and exponentiated-Weibull), modelling the location and the scale parameter with regressors simultaneously in Chapter 3. Half-normal plots with simulation envelopes are used to assess goodness-of-fit, using the `hnp` package (Moral et al., 2017) for R software (R Core Team, 2017).

2.3 Mite abundance data

The diversity and abundance of soil animals is dominated by the mites, and hence the understanding of soil systems is directly related to the mite fauna (Walter and Proctor, 1999). To study the effect of climate change on mite abundance, Ramírez (2012) conducted a sampling study in Colombia. Mites from the mesostigmata order and springtails (Collembola) were sampled bimonthly throughout 2010 at nine sites in both a forest patch and in a pasture, totaling $6 \times 9 \times 2 = 108$ observations, with many zero counts (see Figure 2.1). Environmental covariates such as humidity and temperature were also recorded, see the Supplementary Materials section for the complete dataset.

Figure 2.1. Number of sampled mites and springtails at nine sites in a forest patch and in a pasture, over six time occasions, from the sampling carried out by Ramírez (2012)

This dataset is explored in Chapter 4 using mixture models to jointly estimate the probability of detection of individuals and the abundance at each sampled site.

2.4 Bald eagle data

As a top predator, the bald eagle may act as an environmental indicator, given its presence in many different food webs (Morrison, 1986). To illustrate the use in ecological studies of the joint model proposed in Chapter 4, we used data on bald eagles and mallards, sampled by the American Breeding Bird Survey (ABBS) in the states of Oregon, Washington and British Columbia for the 2015 campaign, see the Supplementary Materials section for the complete dataset. Every year since 1966, researchers have gone along predefined routes in all of the states in the USA and Canada and recorded the number of birds they see or hear on 50 three-minute stops.

This dataset is not related to entomology, but is also used in Chapter 4 as another illustration of the functionality of the R package `jointNmix` (Moral et al., 2016).

2.5 Corn and progeny data

A very important pest of stored maize in Brazil is *Sitophilus zeamais* (Coleoptera: Curculionidae). Ribeiro et al. (2013) conducted an experiment to assess the insecticide action of organic extracts of *Annona mucosa* (Annonaceae) on *S. zeamais*. They treated Petri dishes containing 10g of corn with extracts prepared with different parts of the plant, i.e. seeds, leaves and branches at concentrations of

1,500mg/kg or just water (as a control), using a completely randomized design with 10 replicates. After that, they transferred 20 *S. zeamais* adults to each Petri dish and, after 60 days, the number of damaged and undamaged corn grains were counted, as well as the number of eggs laid by the insects (progeny).

These datasets are readily available from the `hnp` R package (Moral et al., 2017) using `data(corn)` and `data(progeny)`. They are explored throughout Chapter 5 for the illustration of half-normal plots with simulation envelopes.

2.6 Orange data

In tissue culture experiments, it is frequently of interest to study the effects of different substances on the production of plant embryos. Tomaz et al. (1997) conducted a complete randomized block design experiment with five blocks to study the effect of six sugars, i.e. maltose, glucose, galactose, lactose, sucrose and glycerol, on the stimulation of somatic embryos from callus cultures of the orange variety *Caipira*. The first five sugars were used at dose levels of 18, 37, 75, 110 and 150 μ M and glycerol was used at dose levels of 6, 12, 24, 36 and 50 μ M. The relationship between number of embryos and dose appears to be quadratic (see Figure 2.2).

Figure 2.2. Number of embryos produced at six different doses of six different sugars from the experiment carried out by Tomaz et al. (1997)

Again, this dataset is not explicitly related to entomology, but is used on Chapter 5 as another illustration of the functionality of the R package `hnp` (Moral et al., 2017). The dataset is readily available from this package with `data(orange)`.

2.7 Leukemia data

In a clinical trial to assess the efficacy of maintenance chemotherapy for acute myelogenous leukemia, after reaching a state of remission through treatment with chemotherapy, 23 patients were randomized into two groups. The first group continued receiving maintenance chemotherapy and the second did not. The observed outcome was the time (in weeks) until relapse and it was right-censored, as described by Miller (1997).

Again, this dataset is not related to entomology, but is used in Chapter 5 as another illustration of the functionality of the R package `hnp` (Moral et al., 2017), now with survival analysis models. The dataset is readily available from the `survival` package (Therneau, 2015) using `data(leukemia)`.

2.8 Insect attack data

As part of a study of predation and competition behaviour of *Podisus nigrispinus* and *Euborellia annulipes*, Moral (2014) carried out a randomized complete block design experiment with 50 blocks. Each block consisted of four Petri dishes with one of the four treatments resulting from the combination of two males (one of each species), or two females, competing for one, or three, larvae of *Alabama argillacea*. The Petri dishes were observed for one hour and the number of attacks between the predators was recorded. We used a subset of this data in Chapter 6 to illustrate the use of the `bivrp` R package (Moral et al., 2016). This subset is available in the Supplementary Materials section.

References

- CARDINALE, B. J.; HARVEY, C. T.; GROSS, K.; IVES, A. R. (2003) Biodiversity and biocontrol: emergent impacts of a multi-enemy assemblage on pest suppression and crop yield in an agroecosystem. **Ecology Letters**, v. 6, 857–865.
- CRUZ, I.; FIGUEIREDO, M. L. C.; MATOSO, M. J. (1999) **Controle biológico de *Spodoptera frugiperda* utilizando o parasitismo de ovos *Trichogramma***. Sete Lagoas: EMBRAPA-CNPMS. 40 p.
- MILLER, R. G. (1997) **Survival analysis**. New York: John Wiley & Sons. 238 p.
- MORAL, R. A. (2014) Modelagem estatística e ecológica de relações tróficas em pragas e inimigos naturais. Master's dissertation, ESALQ-USP, Piracicaba, Brazil.
- MORAL, R. A.; DEMÉTRIO, C. G. B.; HINDE, J. (2016) jointNmix: Joint N-Mixture Models for Site-Associated Species. R package version 1.0, Available at: <<http://CRAN.R-project.org/package=jointNmix>>
- MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. B. bivrp: (2016) Bivariate Residual Plots with Simulation Polygons. R package version 1.0, Available at: <<http://CRAN.R-project.org/package=bivrp>>
- MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. B. (2017) hnp: Half-Normal Plots with Simulation Envelopes. R package version 1.2-3, Available at: <<http://CRAN.R-project.org/package=hnp>>
- MORRISON, M. L. (1986) Bird populations as indicators of environmental change. **Current ornithology**, Springer US. p. 429–451.
- OLIVEIRA, J. E. M.; TORRES, J. B.; CARRANO-MOREIRA, A. F.; RAMALHO, F. S. (2002) Biologia de *Podisus nigrispinus* predando lagartas de *Alabama argillacea* em campo. **Pesquisa Agropecuária Brasileira**, v. 37, p. 7–14.

R CORE TEAM. (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, Available at: <<http://www.R-project.org>>.

REDOAN, A. C.; CRUZ, I.; CARVALHO, G. A.; FIGUEIREDO, M. L. C.; SILVA, R. B. (2010) Seletividade de inseticidas registrados no controle de *Spodoptera frugiperda* em milho para ninfas de *Doru luteipes* e *Euborellia annulipes*. In: CONGRESSO NACIONAL DE MILHO E SORGO, 2010, Goiania. Goiânia: ABMS, p. 226–228.

RAMÍREZ, D. M. R. (2012) Ácaros edáficos Mesostigmata de grandes altitudes na Colmbia e os possíveis efeitos de mudanas edafo-climáticas sobre as populaes destes ácaros. Master's dissertation.

THERNEAU, T. (2015) A package for survival analysis in S. R package version 2.38-3. Disponível em <<http://CRAN.R-project.org/package=survival>>.

TOMAZ, M. L.; MENDES, B. M. J.; FILHO, F. A. M.; DEMÉTRIO, C. G. B.; JANSKUL, N.; RODRIGUEZ, A. P. M. (1997) Somatic embryogenesis in *Citrus* spp.: carbohydrate stimulation and histodifferentiation. **In Vitro Cellular & Developmental Biology – Plant**, v. 37, p. 446–452.

WALTER, D. E.; PROCTOR, H. C. (1999) **Mites: Ecology, evolution and behaviour**. Sydney: CABI Publ. 322 p.

3 ACCELERATED FAILURE RATE MIXED MODELS: APPLICATIONS IN INSECT ECOLOGY

Abstract

Survival models have been extensively used to analyse time-until-event data. There is a range of extended models that incorporate different aspects, such as overdispersion/frailty, mixtures, and flexible response functions through semi-parametric models. It is important to assess goodness-of-fit of a fitted model so that statistical inference is accurate and reliable and a relatively simple way of doing this is through half-normal plots with simulation envelopes. In this work, we fitted a range of survival models to time-until-event data, where the event was an insect predator attacking a larva in a biological control experiment. We started with the exponential and Weibull models and then fitted the exponentiated-Weibull location-scale model with regressors both for the location and the scale parameters. We did variable selection for each model and found that the exponentiated-Weibull fits the data better. We also included a random effect in the exponentiated-Weibull model. Finally, we developed useful tools to assess goodness-of-fit, which are implemented in the `hnp` package in R.

Keywords: Biological control; Exponentiated models; Half-normal plots with simulation envelopes; Location-scale modelling; Mixed survival models

3.1 Introduction

In insect studies, outcomes of interest vary widely and hence a broad range of data types are obtained, such as discrete or continuous data, that can be either univariate or multivariate (Demétrio et al., 2014). When studying the ecology of insect species, the interest lies in describing several ecological processes, such as predation, competition, prey-preference, amongst others. In a functional response study, for example, outcomes are in the form of the proportion of eaten prey given a time span (Battell et al., 2012); in competition experiments, it is of interest to understand in which circumstances coexistence between competitors is possible and which competitor prevails under most scenarios (Amarasekare, 2003); in prey-preference studies, information on which prey is preferred, or whether there is no preference, may be of particular interest (e.g. Al-Zyoud and Sengonca, 2004).

These ecological data can be used for decision making, especially for developing pest management strategies. Integrated pest management (IPM) consists of a series of activities with the purpose of reducing damage from insect pests in agroecosystems (Kogan, 1998). IPM involves (i) the choice of an economic damage threshold; (ii) pest identification and monitoring; (iii) prevention through several techniques; and (iv) control, which is done when steps (i) and (ii) indicate that it should be applied and preventive techniques are no longer available. One of the types of control is biological control, which involves the promotion of pest suppression using other species of insects and fungi, referred to as natural enemies. Hence, no chemical pesticides are used, assuring better crop quality and disease risk prevention for humans (Kogan, 1998).

One common type of ecological data is the time until the occurrence of an event, e.g., death of a predator, time taken for a parasitoid to parasitise its prey, time until a competitor attacked another competitor, time until a predator attacked its prey. These data may yield relevant information for biological control strategies. Because this type of variable is continuous and strictly positive, it may be transformed prior to analysis using models with a real support, or analysed through survival analysis models or non-parametric tests (Colosimo and Giolo, 2006). Many parametric models have been proposed in the literature and a common objective is to use a well-fitting but not overly complex model to make

inferences from (Aitkin et al., 2009). It is very important to assess goodness-of-fit of survival models so that any inference is not misleading. However, this may become a difficult task. A good alternative is to use appropriate diagnostic plots, such as half-normal plots with simulation envelopes (Atkinson, 1985).

The main goals of this work are to build and assess goodness-of-fit for different location-scale models, fitted to data from an experiment on prey preference. Predators were given a choice between parasitised and non-parasitised prey and the time until a predator attacked a prey was observed. We propose three location-scale models, starting with the simplest, exponential model, turning to the Weibull model and, finally fitting the exponentiated-Weibull model. Usually, only the location parameter is modelled with covariates. However, here we model the location and scale parameters with covariates, simultaneously. Moreover, we develop functions for the statistical software R (R Core Team, 2017) to produce half-normal plots with simulation envelopes for these models to assess goodness-of-fit, using the framework from the `hnp` package (Moral et al., 2017). In the following sections, we describe the experiment, present the models and estimation methods, as well as the goodness-of-fit assessment techniques, and finally we present and discuss the results.

3.2 Case-study

A major pest of maize worldwide is the larva of *Spodoptera frugiperda* (Oliveira et al., 2004). Three important natural enemies of this pest are the stinkbug *Podisus nigrispinus* (Oliveira et al., 2002) and the earwig *Euborellia annulipes* (Redoan et al., 2010), both predators, and the parasitoid wasp *Campoletis flavicincta* (Cruz et al., 1999). The latter lays eggs inside *S. frugiperda* larvae that hatch and develop inside them, killing the larva upon reaching the adult stage. This process causes substantial metabolic changes in the larvae which may be perceived by potential predators.

In a biological control context, it is important that predators act synergistically to promote pest suppression (Cardinale et al., 2003). In this sense, a desirable outcome when controlling *S. frugiperda* with multiple natural enemies, such as the three species mentioned above, would be weak competition between predators and preference for non-parasitised larvae so that the parasitoid's population is also maintained in the system. Information on the time taken by different predator species to attack prey and on prey preference are particularly relevant in this context.

To study the predatory behaviour of the stinkbug and the earwig when given the choice between parasitised and non-parasitised prey, an experiment was conducted in a complete randomized block design with 33 blocks and four treatments in a 2×2 (species and gender) factorial design, i.e., male and female stinkbugs and earwigs, which were fasted for 24 hours prior to the experiment set up, see Figure 3.1. Each block consisted of four Petri dishes with one predator and two *S. frugiperda* larvae, one of which was previously parasitised by *C. flavicincta* and another which was not. Each experimental unit was observed for one hour and the time (in seconds) taken until the predator first attacked a larva was recorded, as well as which larva the predator chose to effectively consume first (predator preference), see Supplement S1 for the complete dataset. The experiment took three days to complete and on each day 11 blocks were installed and observed, so there is reason to believe that the day effect is random. However, the block within day effect should be treated as fixed because only up to three blocks could be observed at the same time. In that sense, predators which were used in the first blocks were fasted for 24 hours prior to experiment installation but predators used in the last blocks had a larger fasting time which could systematically influence their behaviour. As the data consists of time until an event and in some replicates there was not an attack within one hour of observation there are some censored times, and the proper way of analysing such data is through survival time to event models (Aitkin et al., 2009).

Figure 3.1. Experiment treatments: two *S. frugiperda* larvae, one of which is non-parasitised and the other previously parasitised by *C. flavicincta* with (a) a *P. nigrispinus* male, (b) an *E. annulipes* male, (c) a *P. nigrispinus* female and (d) an *E. annulipes* female

3.3 Background

3.3.1 Survival analysis

Time-until-event data are, in general, analysed by survival models (Aitkin et al., 2009). Let T be a positive random variable representing the time until the occurrence of an event of interest with probability density function (pdf) $f_T(t)$ and distribution function $F_T(t)$. In this work the event of interest is a predator attacking a larva. Then, $S_T(t) = P(t \geq T) = \int_t^\infty f_T(t)dt = 1 - F_T(t)$ is called the survival function of T . Another important function in survival analysis is the hazard function or the failure rate function, which expresses the instantaneous failure rate for an event of interest and is given by

$$h_T(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \frac{f_T(t)}{S_T(t)}. \quad (3.1)$$

Suppose that the observed times t are split into two groups, such as treatments, for example. When the ratio between the hazard functions for the two groups, say $h_1(t)$ and $h_2(t)$, does not depend on t and hence is constant, this feature is called hazard proportionality. It is a very important assumption for several non-parametric tests and for the Cox regression model as well (Harrell Jr., 2015).

An important feature of survival analysis is that there may be censored observations which are cases where the event of interest has not been recorded. Reasons for this include experiments which end after a given amount of time (which is the case of this work), other events that make the predator unable to attack a prey (e.g., a larva escaping the Petri dish), and many other reasons in different types of studies. By treating them as incomplete observations, the censored data may be incorporated into survival analysis models through the likelihood function.

In some cases, it is of interest to build regression models using distributions with positive support to assess the effect of covariates on the process under study. This can be done through location-scale models, also called accelerated failure rate models. The effects of covariates may be interpreted as

acceleration or deceleration of the failure rate function.

3.3.2 Exploratory analysis

A useful tool to explore time-to-event data is the non-parametric Kaplan-Meier estimator $\hat{S}_T(t)$ (Kaplan and Meier, 1958; Borgan, 2005). Let $t_1, t_2, \dots, t_k$ represent k times until the occurrence of an event, $t_{(1)}, t_{(2)}, \dots, t_{(k)}$ be the ordered times, d_j the number of failures at t_j , $j = 1, \dots, k$, and n_j the number of individuals which have not yet failed until $t_{(j)}$. The Kaplan-Meier estimator is given by

$$\hat{S}_T(t) = \prod_{j=t_{(j)} \leq t} \left(\frac{n_j - d_j}{n_j} \right). \quad (3.2)$$

It can be shown that this estimator is a maximum likelihood estimator for $S_T(t)$. It is also an unbiased and consistent estimator, and its distribution converges to a normal distribution (Breslow and Crowley, 1974). The Kaplan-Meier estimator is usually plotted against t to allow for better visualization of the behaviour of the survival function.

Another useful graphical exploration of time-to-event data is the total-time-on-test (TTT) plot, proposed by Aarset (1987). It is obtained by plotting

$$G\left(\frac{r}{n}\right) = \frac{(\sum_{i=1}^r t_{(i)}) + (n-r)t_{(r)}}{\sum_{i=1}^n t_{(i)}}, \quad (3.3)$$

against r/n , where $r = 1, \dots, n$, and $t_{(i)}, i = 1, \dots, n$, are the order statistics of the sample. This is a methodology used for obtaining the empirical hazard rate. The curve may exhibit different patterns and these may provide guidance towards the form of the hazard function that should be used when modelling the data. When it is a diagonal straight line (see Figure 3.2, curve A), the hazard function is constant; when it is a convex curve (see Figure 3.2, curve B) or a concave curve (see Figure 3.2, curve C), the failure rate function monotonically decreases and increases, respectively; when it is firstly convex and then concave (see Figure 3.2, curve D), the hazard function is U-shaped; and when it is firstly concave and then convex (see Figure 3.2, curve E) the hazard function is unimodal.

Figure 3.2. Different shapes of the TTT-plot

3.3.3 Log-rank test

A commonly used test to compare survival curves under the assumption of proportional hazards is the log-rank test, proposed by Mantel and Haenszel (1959). Let $t_{i1} < t_{i2} < \dots < t_{ik}$ represent k times

until the occurrence of an event for group $i, i = 1, 2$, d_{ij} the number of failure times in t_{ij} , and n_{ij} the number of individuals which have not yet failed until t_{ij} . The statistic

$$\frac{\sum_{j=1}^k (d_{2j} - w_{2j})}{\sqrt{\sum_{j=1}^k v_{2j}}} \xrightarrow{D} N(0, 1),$$

where $w_{2j} = \frac{(d_{1j} + d_{2j})n_{2j}}{n_{1j} + n_{2j}}$ and $v_{2j} = \frac{(d_{1j} + d_{2j})n_{1j}n_{2j}(n_{1j} + n_{2j} - d_{1j} - d_{2j})}{(n_{1j} + n_{2j})^2(n_{1j} + n_{2j} - 1)}$. Hence a statistic to test whether two survival curves are equal may be written as

$$\frac{\{\sum_{j=1}^k (d_{2j} - w_{2j})\}^2}{\sum_{j=1}^k v_{2j}},$$

which has an asymptotic χ^2 distribution with 1 degree of freedom. Hence, when the test-statistic is larger than $\chi_{1;0.95}^2 = 3.84$, the null hypothesis that the two survival curves are equal is rejected.

There is no need for fitting a survival model to test for differences in survival curves when using the log-rank test, as it is a non-parametric test. However, when the hazard proportionality assumption fails the log-rank test is not applicable. This test is widely used in many applied areas of science and to publish conclusions based on censored data in high-impact peer-reviewed journals, especially in medical-related research (Matthews and Farewell, 2007).

3.4 Modelling

Let W_i , $i = 1, \dots, n$, be independent random variables representing recorded times until a predator attacked a larva. Over the following sections we will describe how to build location-scale models to analyse these data, starting with a simple model and then adding flexibility with increasing complexity. We then show how random effects may be included to incorporate correlation among observations taken in the same day of experiment.

3.4.1 Exponential model

One of the simplest survival analysis models is the exponential model. Suppose that W_i follows the exponential distribution. The pdf may be written as

$$f_{W_i}(w_i; \lambda_i) = \lambda_i e^{-\lambda_i w_i}, w_i > 0, \quad (3.4)$$

with $\lambda_i > 0$ called the rate parameter (Aitkin et al., 2009). Constructing a location-scale model from (3.4) is easily done by finding the probability density function of $Y_i = \log W_i$ and reparameterising $f_{Y_i}(y_i; \lambda_i)$ in terms of $\mu_i = -\log \lambda_i$, a location parameter, which yields the following result (see Appendix B for calculations):

$$f_{Y_i}(y_i; \mu_i) = e^{y_i - \mu_i - e^{y_i - \mu_i}}, -\infty < y_i < \infty. \quad (3.5)$$

The survival function for model (3.5) is

$$S_{Y_i}(y_i; \mu_i) = e^{-e^{y_i - \mu_i}}. \quad (3.6)$$

Random samples may be generated through the inverse transformation theorem, using the following expression (see Appendix B for calculations):

$$Y_i = \mu_i + \log\{-\log(1 - U_i)\}, U_i \sim U(0, 1). \quad (3.7)$$

3.4.2 Weibull model

The Weibull distribution is an important model in survival analysis (Aitkin et al., 2009). If W_i follows a Weibull distribution, its pdf may be written as

$$f_{W_i}(w_i; \alpha_i, \gamma_i) = \frac{\gamma_i}{\alpha_i^{\gamma_i}} w_i^{\gamma_i-1} e^{-\left(\frac{w_i}{\alpha_i}\right)^{\gamma_i}}, w_i > 0, \quad (3.8)$$

with $\alpha_i > 0$ a scale parameter and $\gamma_i > 0$ a shape parameter. When $\gamma_i = 1$, the Weibull distribution becomes the exponential distribution with parameter $\lambda_i = \frac{1}{\alpha_i}$. Constructing a location-scale model from (3.8) is done by finding the probability density function of $Y_i = \log W_i$ and reparameterising $f_{Y_i}(y_i; \alpha_i, \gamma_i)$ in terms of $\mu_i = \log \alpha_i$, a location parameter, and $\sigma_i = \frac{1}{\gamma_i}$, a scale parameter, which yields the following result (see Appendix B for calculations):

$$f_{Y_i}(y_i; \mu_i, \sigma_i) = \frac{1}{\sigma_i} e^{\frac{y_i - \mu_i}{\sigma_i}} e^{-e^{\frac{y_i - \mu_i}{\sigma_i}}}, -\infty < y_i < \infty, \quad (3.9)$$

which is the pdf of the extreme value (type I), or Gumbel, distribution. The survival function for model (3.9) is

$$S_{Y_i}(y_i; \mu_i) = e^{-e^{\frac{y_i - \mu_i}{\sigma_i}}}. \quad (3.10)$$

Random samples may be generated using the following expression (see Appendix B for calculations):

$$Y_i = \mu_i + \sigma_i \log\{-\log(1 - U_i)\}, U_i \sim U(0, 1). \quad (3.11)$$

3.4.3 Exponentiated family of distributions

If W_i follows a distribution from the family of exponentiated distributions (Nadarajah et al., 2013), its pdf may be written as

$$f_{W_i}(w_i; \theta_i, a) = a[G(w_i; \theta_i)]^{a-1} g(w_i; \theta_i), w_i > 0, \quad (3.12)$$

with $a > 0$ a shape parameter, $g(\cdot)$ a pdf and $G(\cdot)$ its respective distribution function and θ_i the vector of parameters of the distribution. Taking the base distribution as Weibull with $g(u; \alpha_i, \gamma_i) = \frac{\gamma_i}{\alpha_i^{\gamma_i}} u^{\gamma_i-1} e^{-\left(\frac{u}{\alpha_i}\right)^{\gamma_i}}$, $\alpha_i > 0$, $\gamma_i > 0$, and $G(u; \alpha_i, \gamma_i) = 1 - e^{-\left(\frac{u}{\alpha_i}\right)^{\gamma_i}}$, then

$$f_{W_i}(w_i; \alpha_i, \gamma_i, a) = a \left[1 - e^{-\left(\frac{w_i}{\alpha_i}\right)^{\gamma_i}}\right]^{a-1} \frac{\gamma_i}{\alpha_i^{\gamma_i}} w_i^{\gamma_i-1} e^{-\left(\frac{w_i}{\alpha_i}\right)^{\gamma_i}}, w_i > 0, \quad (3.13)$$

is the pdf of the exponentiated-Weibull distribution (Nadarajah et al., 2013). When $\gamma_i = 1$, this is the pdf of the exponentiated-exponential distribution and when $a = 1$, it is the pdf of the Weibull distribution. Consequently, when $a = 1$ and $\gamma_i = 1$, this is the pdf of the exponential distribution. Constructing a location-scale model from (3.13) is straightforward and can be done by finding the probability density function of $Y_i = \log W_i$ and reparameterising $f_{Y_i}(y_i; \alpha_i, \gamma_i, a)$ in terms of $\mu_i = \log \alpha_i$, a location parameter, and $\sigma_i = \frac{1}{\gamma_i}$, a scale parameter, which yields the following result (see Appendix B for calculations):

$$f_{Y_i}(y_i; \mu_i, \sigma_i, a) = \frac{a}{\sigma_i} \left[1 - \exp\left(-e^{\frac{y_i - \mu_i}{\sigma_i}}\right)\right]^{a-1} e^{\frac{y_i - \mu_i}{\sigma_i}} \exp\left(-e^{\frac{y_i - \mu_i}{\sigma_i}}\right), -\infty < y_i < \infty. \quad (3.14)$$

The survival function for model (3.14) is

$$S_{Y_i}(y_i; \mu_i, \sigma_i, a) = 1 - \left[1 - \exp\left(-e^{\frac{y_i - \mu_i}{\sigma_i}}\right)\right]^a. \quad (3.15)$$

The parameter a makes the left tail of the distribution heavier as it approaches zero, making it more symmetric as it increases, see Figure 3.3(a). Together with the parameter σ_i , it provides additional flexibility to the exponential and Weibull models, see Figure 3.3(b).

Random samples for the exponentiated-Weibull distribution may be generated using the following expression (see Appendix B for calculations):

$$Y_i = \mu_i + \sigma_i \log\{-\log(1 - U_i^{\frac{1}{a}})\}, U_i \sim U(0, 1). \quad (3.16)$$

Figure 3.3. Different shapes of the survival functions for the exponentiated-Weibull model (3.14) for different values of a , fixing $\mu_i = 3$ and (a) $\sigma_i = 1$, and (b) for different values of σ_i

3.4.4 Maximum likelihood estimation

For the example described in Section 3.2, there is a type-I censoring scheme taking place, i.e., the Petri dishes were observed for one hour and for those in which the predators did not attack any larva the time until attack was not observed – it is only known that it is larger than 60 minutes. Let C_i represent censoring times. We observe response times $T_i = \min(Y_i, C_i)$, together with a censoring indicator $\delta_i = 1$ if $T_i \leq C_i$ and $\delta_i = 0$ if $T_i > C_i$. The log-likelihood function under non-informative censoring can be written as

$$l(\boldsymbol{\mu}, \boldsymbol{\sigma}, a) = \sum_{i=1}^n \{ \delta_i \log f_{Y_i}(y_i; \mu_i, \sigma_i, a) + (1 - \delta_i) \log S_{Y_i}(y_i; \mu_i, \sigma_i, a) \}, \quad (3.17)$$

To incorporate explanatory variables, let $\mu_i = \mathbf{x}_{1i}^T \boldsymbol{\beta}_1$ and $\log \sigma_i = \mathbf{x}_{2i}^T \boldsymbol{\beta}_2$, with $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ $1 \times p_1$ and $1 \times p_2$ covariate vectors and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ $1 \times p_1$ and $1 \times p_2$ parameter vectors. Then, the log-likelihood function can be written as

$$l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, a) = \sum_{i=1}^n \{ \delta_i \log f_{Y_i}(y_i; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, a) + (1 - \delta_i) \log S_{Y_i}(y_i; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, a) \}. \quad (3.18)$$

Estimation of the $p_1 + p_2 + 1$ parameters can be done by maximizing $l(\cdot)$. Analytical maximization is possible by differentiating $l(\cdot)$ with relation to the parameter vector to be estimated and solving the vector of analytical derivatives. Numerical methods are also available and implemented in a range of statistical software. In this work, we used the BFGS (Broyden et al., 1973) method implemented in the `optim` function in software R (R Core Team, 2017). For an overview of optimization methods, refer to Appendix A.

3.4.5 Estimation including random effects

Because the experiment took three days with 11 blocks installed and observed on each day, the day effect may be treated as random, as there were three different batches of insects, coming from the same population. When including random effects in the linear predictor, the likelihood function must be integrated over the random effects so that only the parameters to be estimated are left in the function. In a sample divided in m groups, let T_{ij} be the time until attack of the i -th predator in the j -th group, with $i = 1, \dots, n_j$ and $j = 1, \dots, m$ (here, $m = 3$ days). Let C_{ij} represent the censoring times. We again observe response times $T_{ij} = \min(Y_{ij}, C_{ij})$, together with a censoring indicator $\delta_{ij} = 1$ if $T_{ij} \leq C_{ij}$ and $\delta_{ij} = 0$ if $T_{ij} > C_{ij}$. To incorporate explanatory variables and a random intercept, let $\mu_{ij} = \mathbf{x}_{1ij}^T \boldsymbol{\beta}_1 + b_j$ and $\log \sigma_{ij} = \mathbf{x}_{2ij}^T \boldsymbol{\beta}_2$, with $\mathbf{x}_{1ij}$ and $\mathbf{x}_{2ij}$ $1 \times p_1$ and $1 \times p_2$ covariate vectors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ $1 \times p_1$ and $1 \times p_2$ parameter vectors, and B_j a normal unobserved random variable, that is, $B_j \sim N(0, \sigma_d^2)$, with pdf $f_{B_j}(b_j; \sigma_d^2)$. Under non-informative censoring, the conditional likelihood function for the n_j individuals in group j is now

$$L_j(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, a | b_j) = \prod_{i=1}^{n_j} f_{Y_{ij}}(y_{ij} | b_j, \mathbf{x}_{1ij}, \mathbf{x}_{2ij})^{\delta_{ij}} S_{Y_{ij}}(y_{ij} | b_j, \mathbf{x}_{1ij}, \mathbf{x}_{2ij})^{1-\delta_{ij}}. \quad (3.19)$$

Assuming independence among groups, to obtain the marginal likelihood function, the product between the conditional likelihood function and the pdf of the random effect must be integrated over the random effect, which yields

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, a, \sigma_d^2) = \prod_{j=1}^m \left\{ \int_{-\infty}^{\infty} L_j(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, a | b_j) f_{B_j}(b_j; \sigma_d^2) db_j \right\}. \quad (3.20)$$

Therefore, the log-likelihood function to be maximized may be written as

$$l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, a, \sigma_d^2) = \sum_{j=1}^m \log \left\{ \int_{-\infty}^{\infty} L_j(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, a | b_j) f_{B_j}(b_j; \sigma_d^2) db_j \right\}. \quad (3.21)$$

There is no analytical solution for the integral in (3.21), so numerical methods must be used to maximize the log-likelihood function. One of these approximations is the Gauss-Hermite quadrature of order q for integrals of the form $\int_{-\infty}^{+\infty} \exp(-x^2) f(x) dx$ which is written as

$$\int_{-\infty}^{+\infty} \exp(-x^2) f(x) dx \approx \sum_{k=1}^q v_k f(s_k), \quad (3.22)$$

with $s_k, k = 1, 2, \dots, q$, the q roots of the Hermite polynomial of degree q , written as

$$H_q(x) = (-1)^q \exp(x^2) \frac{d^q}{dx^q} \exp(-x^2) \quad (3.23)$$

and weights

$$v_k = \frac{2^{(q+1)} n! \sqrt{\pi}}{H'_q(s_k)}. \quad (3.24)$$

In a general form, as

$$g(x) = \exp(-x^2) f(x) \Leftrightarrow f(x) = g(x) \exp(x^2),$$

the approximation of $\int_{-\infty}^{+\infty} g(x) dx = \int_{-\infty}^{+\infty} \exp(-x^2) f(x) dx$ may be obtained by $\sum_{k=1}^q v_k f(s_k) = \sum_{k=1}^q v_k g(s_k) \exp(s_k^2)$. Therefore, the Gauss-Hermite quadrature approximation of the integral of any function $g(x)$ may be written as

$$\int_{-\infty}^{+\infty} g(x) dx \approx \sum_{k=1}^q v_k g(s_k) \exp(s_k^2). \quad (3.25)$$

As the number of integration points q is increased, the better the approximation becomes. This method is implemented in many R packages, such as `npmlreg` (Einbeck et al., 2014) and `g1mmML` (Broström, 2013), and in SAS's procedure NLMIXED (Sheu, 2002).

One variation of this method is the adaptive Gauss-Hermite quadrature, which centers the integration points around the maximum $s^* = \arg \max_x \{\exp(-x^2)f(x)\}$ of the objective function. In this case we have

$$s_k^+ = s^* + s_k \left[-\frac{d^2}{dx^2} \log\{\exp(-x^2)f(x)\}|_{x=s^*} \right]^{-\frac{1}{2}} \quad (3.26)$$

as scaled roots of the Hermite polynomial of degree q and scaled weights

$$v_k^+ = v_k \exp(-s_k^{+2} + s_k^2) \left[-\frac{d^2}{dx^2} \log\{\exp(-x^2)f(x)\}|_{x=s^*} \right]^{-\frac{1}{2}}. \quad (3.27)$$

Hence, the adaptive Gauss-Hermite quadrature of order q approximation may be written as

$$\int_{-\infty}^{+\infty} \exp(-x^2)f(x)dx \approx \sum_{k=1}^q v_k^+ f(s_k^+). \quad (3.28)$$

This method is more precise than the Gauss-Hermite quadrature for functions whose mode is far from zero, however at a computational cost which is maximizing $\exp(-x^2)f(x)$. It is implemented in R packages and in SAS's procedure NLMIXED; a variation is readily available from the R base installation through function `integrate`. For more details see Davis and Rabinowitz (2007), Liu and Pearce (1994), and Rabe-Hesketh et al. (2002).

The posterior distribution $B_j|Y_{ij}$ may be written as

$$f_{B_j}(b_j|y_{ij}) = \frac{L_j(\beta_1, \beta_2, a|b_j)f_{B_j}(b_j; \sigma_d^2)}{\int_{-\infty}^{\infty} L_j(\beta_1, \beta_2, a|b_j)f_{B_j}(b_j; \sigma_d^2)db_j} \quad (3.29)$$

Prediction of random effects may be obtained by finding the posterior distribution $B_j|Y_{ij}$ and then taking $\hat{b}_j = E(B_j|Y_{ij} = y_{ij})$. Hence, the predicted random effects may be computed through

$$\hat{b}_j = \int_{-\infty}^{\infty} b_j f_{B_j}(b_j|y_{ij})db_j. \quad (3.30)$$

Numerical integration techniques, such as the Gauss-Hermite quadrature described above, may be used to compute $\hat{b}_j$.

3.4.6 Model selection

In order to study the effects of the covariates species, gender and type of consumed larva, the following linear predictors were used initially:

$$\mu_{ijkl} = \beta_0 + \delta_l + s_i + g_j + (sg)_{ij} + c_k, \quad (3.31)$$

$$\log(\sigma_{ijkl}) = \beta_0 + s_i + g_j + (sg)_{ij} + c_k, \quad (3.32)$$

where β_0 is the intercept, δ_l is the effect of the l -th block, $l = 1, \dots, 33$, s_i is the effect of the i -th species ($i = 1$ for the earwig and $i = 2$ for the stinkbug), g_j is the effect of the j -th gender ($j = 1$ for females and $j = 2$ for males), $(sg)_{ij}$ is the effect of the interaction between the i -th species and the j -th gender, and c_k is the effect of prey preference ($k = 1$ when no prey was effectively consumed, $k = 2$ when the non-parasitised larva was preferred, and $k = 3$ when the parasitised larva was preferred). We then performed backward selection using likelihood-ratio tests, which compare the maximised likelihood functions under two models. The likelihood-ratio test statistic may be written as

$$\Lambda = -2\{l(\theta^0) - l(\theta^1)\},$$

where $l(\cdot)$ is a log-likelihood function and $\boldsymbol{\theta}^0$ and $\boldsymbol{\theta}^1$ are the vectors of maximum likelihood parameter estimates under the null and alternative hypotheses, respectively. This test statistic has an asymptotic χ^2 distribution with degrees of freedom equal to the difference between the number of parameters of the compared models (Aitkin et al., 2009).

We also studied the combination of c_k levels to reduce the number of parameters and other linear predictors to combine estimated survival curves. After model selection, we then added the day effects $d_m, m = 1, 2, 3$, as random in the linear predictor for the location parameter of the exponentiated-Weibull model as normally distributed, that is, $D_m \sim N(0, \sigma_d^2)$.

3.4.7 Goodness-of-fit assessment

It is important to carry out diagnostic analyses to verify goodness-of-fit because when the model does not fit the data well, statistical inference may lead to incorrect conclusions about the process under study. Several graphical techniques may be used such as plotting different types of residuals or influence measures (e.g. leverage and Cook's distance). Useful diagnostic-checking plots include

- residuals vs. explanatory variables — indicates whether higher-order terms should be included in the linear predictor (for quantitative explanatory variables);
- residuals vs. explanatory variables not included in the model — a systematic relationship indicates that the explanatory variable should be included in the linear predictor
- added-variable plot — detects the relationship of the response variable with an explanatory variable not included in the model without the effects of other variables;
- residuals vs. fitted values — may reveal variance heterogeneity and/or outliers;
- (half-)normal plots of residuals — detects outliers and indicates whether the error distribution was specified properly.

A useful tool to assess model goodness-of-fit is the half-normal plot of deviance residuals with a model-based simulation envelope (Atkinson, 1985; Demétrio et al., 2014). The envelope is such that under the correct model most of the deviance residuals for the fitted model should lie within it. This relatively easy technique consists in plotting the ordered absolute values of a model diagnostic versus the expected order statistics of a half-normal distribution

$$\Phi^{-1} \left(\frac{i + n - \frac{1}{8}}{2n + \frac{1}{2}} \right), \quad (3.33)$$

where i is the i -th order statistic, $1 \leq i \leq n$ and n is the sample size. For a normal plot we use the ordered values (not in absolute value) and plot them versus the expected order statistics of a normal distribution, i.e.,

$$\Phi^{-1} \left(\frac{i + \frac{3}{8}}{n + \frac{1}{4}} \right). \quad (3.34)$$

Obtaining the simulated envelope for a half-normal plot is simple and consists of

1. fitting a model;
2. extracting model diagnostics and calculating sorted absolute values;
3. simulating 99 (or more) response variables using the same model matrix, error distribution and fitted parameters;

4. fitting the same model to each simulated response variable and obtaining the same model diagnostics, again as sorted absolute values;
5. computing the desired percentiles (e.g., 2.5 and 97.5) at each value of the expected order statistic to form the envelope.

In survival analysis theory, there are three commonly used types of residuals. The Cox-Snell residuals (Cox and Snell, 1968) are defined as

$$\hat{e}_i = -\log \hat{S}(y_i), \quad (3.35)$$

where $\hat{S}(\cdot)$ is the estimated survival function. The Martingale (Therneau et al., 1990) residuals are defined as

$$\hat{r}_{m_i} = \delta_i - \hat{e}_i, \quad (3.36)$$

where δ_i is the censoring indicator and $\hat{e}_i$ are the Cox-Snell residuals (3.35). To produce the half-normal plots with simulation envelopes the deviance residuals were used, and are defined as

$$\hat{r}_{D_i} = \text{sign}(\hat{r}_{m_i}) \{-2[\hat{r}_{m_i} + \delta_i \log(\delta_i - \hat{r}_{m_i})]\}^{\frac{1}{2}}, \quad (3.37)$$

where $\hat{r}_{m_i}$ are the Martingale residuals (3.36). It has been demonstrated that the deviance residuals follow, empirically, a standard normal distribution under up to 40% of censored data.

Half-normal plots with simulated envelopes are implemented in the `hnp` package for many standard models (Moral et al., 2017) in R (R Core Team, 2017). It is also possible to use the `hnp` function to produce these plots for models that are not already coded within the function by providing three helper functions, one to extract the deviance residuals, another to simulate response variables, and finally a function to refit the model to the simulated samples, see Annex B for the full R scripts used in this chapter.

3.4.8 Simulation studies

We carried out two simulation studies to help understand the behaviour of the proposed exponentiated-Weibull mixed model under different circumstances. For both studies, three values for the variance of the random effect σ_d^2 were used (0, 1, and 3). Moreover, the location parameter was modelled without covariates and only a random intercept per day of experiment, and the scale parameter was modelled with two dummy covariates, species and whether the predator consumed prey in one hour of observation, that is,

$$\mu_{ijk} = \beta_0 + b_j, \quad (3.38)$$

$$\log(\sigma_{ijk}) = \beta_0 + s_i + c_k, \quad (3.39)$$

where β_0 is the intercept, b_j is the random effect associated with the j -th group of correlated observations, $B_j \sim N(0, \sigma_d^2)$, s_i is the effect of the i -th species and c_k is the effect of consumption, ($k = 1$ when no prey was effectively consumed and $k = 2$ when there was consumption).

The first study was carried out with the same sample size (132) and covariate values as for the actual experiment, as well as same degree of censoring. The second study used three different sample sizes (36, 132, and 480), three proportions of censoring (0%, 15%, and 30% of censored observations), and two numbers of groups of correlated observations (simulating 4 or 12 days of experiment), with the same number of replicates for each combination between the levels of the covariates. True parameter values were fixed as shown in Table 3.1 in all simulated scenarios. For each scenario in each study, 100 simulations were performed and the average of the estimates were computed, as well as the mean squared error.

Table 3.1. True parameter values used in the simulation studies

Parameter	Value
$\log(a)$	2.17
μ – Intercept	2.78
σ – Intercept	1.10
σ – Species: Stinkbug	0.47
σ – Consumption: Yes	−0.51

3.5 Results

3.5.1 Exploratory analysis

The TTT-plots for each treatment show that the failure rate function changes depending on the combination between gender and species, see Figure 3.4. The Kaplan-Meier estimator for the logarithm of the time until attack data suggests that there is a clear difference between species, with earwigs attacking prey in less time than stinkbugs. It also suggests that the interaction between gender and species may not be significant, see Figure 3.5.

Figure 3.4. TTT-plots for (a) male stinkbugs, (b) female stinkbugs, (c) male earwigs, and (d) female earwigs time until attack data

The results from the log-rank test suggest that there is no difference between the genders (log-rank test statistic: 0.0005 on 1 d.f., $p = 0.9830$) but that the survival curves for the species are statistically different (log-rank test statistic: 29.40 on 1 d.f., $p < 0.0001$)

3.5.2 Exponential model fit

We begin by fitting the exponential model to the logarithm of the data. This is, of course, a simple model with only one location parameter which is modelled with covariates, see the linear predictor

Figure 3.5. Kaplan-Meier estimator for each treatment combination

(3.31). There is evidence that the blocks within day effects are not significant ($\lambda = 24.82$, d.f. = 32, $p = 0.8132$). Removing the interaction term yields a non-significant Λ statistic for the likelihood-ratio test between the reduced model and the full model ($\lambda = 2.76$, d.f. = 1, $p = 0.0968$). We then removed the gender main effect and the likelihood-ratio test was also non-significant ($\lambda = 0.46$, d.f. = 1, $p = 0.4978$). Combining levels of the consumption effect such that only two levels were left (i.e., consumption and no consumption in one hour of observation) also yielded a non-significant likelihood-ratio test statistic ($\lambda = 1.96$, d.f. = 1, $p = 0.1615$). The inclusion of a species $\times$ consumption interaction effect was not significant ($\lambda = 2.26$, d.f. = 1, $p = 0.1330$). Any further reductions yielded significant likelihood-ratio test statistics. Therefore, the selected linear predictor was

$$\mu_{ik} = \beta_0 + s_i + c_k, \quad (3.40)$$

where β_0 is the intercept, s_i is the effect of the i -th species and c_k is the effect of consumption, ($k = 1$ when no prey was effectively consumed and $k = 2$ when there was consumption, independent of which prey was consumed – hence the levels are combined), see Table 3.2 for parameter estimates. A global test between the maximal model (3.31) and the selected model (3.40) was also not significant ($\lambda = 30.00$, d.f. = 35, $p = 0.7081$).

Even though the fitted survival curves do not seem to follow the Kaplan-Meier estimates properly, see Figure 3.6, the half-normal plot with simulation envelope suggests that the model fits the data, see Figure 3.7(b). However this may be due to the fact that the deviance residuals are expected to follow a normal distribution and in this case they apparently are well-behaved, see Figure 3.7(a).

The estimated parameters suggest that there is no significant interaction between species and gender at 5% level, as well as no significant gender main effect at 5% level. Moreover, there is a significant

Table 3.2. Parameter estimates and associated standard errors, t values and p -values for the exponential model fitted to the time until attack data

Parameter	Estimate	Std. error	t value	p -value
μ – Intercept	6.24	0.26	23.44	< 0.0001
μ – Species: Stinkbug	1.46	0.18	8.26	< 0.0001
μ – Consumption: Yes	–1.51	0.28	–5.46	< 0.0001

Figure 3.6. Fitted survival curves for the exponential model for (a) each treatment combination and (b) combination between species and consumption levels **Figure 3.7.** (a) Deviance residuals versus indexes and (b) half-normal plot with simulation envelope for the exponential model fit to the time until attack data

consumption effect which separates two groups within each species, the specimens that effectively consume the prey are faster than the ones which only attack and do not entirely consume the larva.

The estimated survival curves for earwigs that did not effectively consume prey in one hour and for stinkbugs that did are almost superimposed (see Figure 3.7b). Fitting a model that merged these two curves yielded the same number of parameters as the previous model (linear predictor (3.40)) but with a more straightforward interpretation (see Table 3.3). However, the fitted curves do not seem to capture well the behaviour of the Kaplan-Meier estimates (see Figure 3.8).

Table 3.3. Parameter estimates and associated standard errors, t values and p -values for the exponential model fitted to the time until attack data

Parameter	Estimate	Std. error	t value	p -value
μ – Species: Earwig x Consumption: Yes	4.78	0.14	35.16	< 0.0001
μ – Species: Earwig x Consumption: No and Species: Stinkbug x Consumption: Yes	6.13	0.12	52.03	< 0.0001
μ – Species: Stinkbug x Consumption: No	8.31	0.58	14.41	< 0.0001

Figure 3.8. Fitted combined survival curves for the exponential model for the combination between species and consumption levels

3.5.3 Weibull model fit

Now we turn to fitting the Weibull model, modelling both the location and the scale parameters with regressors, see linear predictors (3.31) and (3.32). There is evidence that the blocks within day effects are not significant ($\lambda = 37.61$, d.f. = 32, $p = 0.2276$). Removing the interaction term from both linear predictors is not significant according to the likelihood-ratio test ($\lambda = 3.25$, d.f. = 2, $p = 0.1969$). Removing the gender main effect from both linear predictors was also not significant ($\lambda = 0.31$, d.f. = 2, $p = 0.8552$). Combining levels of the consumption effect in both linear predictors also yielded a non-significant likelihood-ratio test statistic ($\lambda = 2.25$, d.f. = 2, $p = 0.3238$). The inclusion of a species $\times$ consumption interaction effect was not significant ($\lambda = 2.39$, d.f. = 2, $p = 0.3025$). Any further reductions yielded significant likelihood-ratio test statistics. Therefore, the selected linear predictor was

$$\mu_{ik} = \beta_0 + s_i + c_k, \quad (3.41)$$

$$\log(\sigma_{ik}) = \beta_0 + s_i + c_k, \quad (3.42)$$

where β_0 is the intercept, s_i is the effect of the i -th species and c_k is the effect of consumption, ($k = 1$ when

no prey was effectively consumed and $k = 2$ when there was consumption), see Table 3.4 for parameter estimates. A global test between the maximal model (3.31 and 3.32) and the selected model (3.41 and 3.42) was also not significant ($\lambda = 43.43$, d.f. = 38, $p = 0.2510$).

Table 3.4. Parameter estimates and associated standard errors, t values and p -values for the Weibull model fitted to the time until attack data

Parameter	Estimate	Std. error	t value	p -value
μ – Intercept	6.16	0.34	17.93	< 0.0001
μ – Species: Stinkbug	1.33	0.18	7.40	< 0.0001
μ – Consumption: Yes	–1.32	0.36	–3.70	0.0002
σ – Intercept	0.20	0.22	0.93	0.3525
σ – Species: Stinkbug	0.26	0.13	1.99	0.0462
σ – Consumption: Yes	–0.43	0.23	–1.88	0.0604

Again, although the fitted survival curves do not seem to follow the Kaplan-Meier estimates properly, see Figure 3.9, the half-normal plot with simulation envelope suggests that the model fits the data, see Figure 3.10(b). However in this case the deviance residuals do not seem to behave as normally distributed, see Figure 3.10(a), because the survival curves do not capture well the curvature of the Kaplan-Meier estimates.

Figure 3.9. Fitted survival curves for the Weibull model for (a) each treatment combination and (b) combination between species and consumption levels

As for the exponential model, the estimated parameters for the location part of the Weibull model suggest that there is no significant interaction between species and gender at 5% level, as well as no significant gender main effect at 5% level, but there is a significant consumption effect. In addition, the estimated parameters for the scale part of the model suggest that the failure rate function accelerates for earwigs and for specimens which have effectively consumed prey. This parameter also models variance heterogeneity.

Here also, the estimated survival curves for earwigs that did not effectively consume prey in one hour and for stinkbugs that did are similar (see Figure 3.10b). Fitting a model that merged these two curves yielded the same number of parameters than the previous model (linear predictors (3.41) and (3.42), see Table 3.5). Again, the fitted curves do not seem to capture well the behaviour of the Kaplan-Meier estimates (see Figure 3.11), but in this case a likelihood-ratio test between this model and the 3-curve exponential model is not significant ($\lambda = 6.76$, d.f. = 3, $p = 0.0799$) and hence this is essentially the same fit.

Figure 3.10. (a) Deviance residuals versus indexes and (b) half-normal plot with simulation envelope for the Weibull model fit to the time until attack data

Table 3.5. Parameter estimates and associated standard errors, t values and p -values for the Weibull model fitted to the time until attack data

Parameter	Estimate	Std. error	t value	p -value
μ – Species: Earwig x Consumption: Yes	4.87	0.11	42.45	< 0.0001
μ – Species: Earwig x Consumption: No and Species: Stinkbug x Consumption: Yes	6.10	0.13	46.11	< 0.0001
μ – Species: Stinkbug x Consumption: No	8.64	1.10	7.86	< 0.0001
σ – Species: Earwig x Consumption: Yes	−0.23	0.10	−2.37	0.0177
σ – Species: Earwig x Consumption: No and Species: Stinkbug x Consumption: Yes	0.06	0.09	0.67	0.5041
σ – Species: Stinkbug x Consumption: No	0.54	0.51	1.06	0.2901

3.5.4 Exponentiated-Weibull model fit

We now fit the exponentiated-Weibull model, modelling both the location and the scale parameters with regressors, see linear predictors (3.31) and (3.32). This model has one additional parameter, a , which is a shape parameter. There is evidence that the blocks within day effects are not significant ($\lambda = 42.21$, d.f. = 32, $p = 0.1071$) Removing the interaction term from both linear predictors is not significant according to the likelihood-ratio test ($\lambda = 2.62$, d.f. = 2, $p = 0.2697$). Removing the gender main effect from both linear predictors was also not significant ($\lambda = 0.24$, d.f. = 2, $p = 0.8850$). Removing all effects for the location parameter linear predictor, leaving only the intercept was also not significant ($\lambda = 3.52$, d.f. = 3, $p = 0.3182$). Combining levels of the consumption effect in the linear predictor for the scale parameter yielded a non-significant likelihood-ratio test statistic ($\lambda = 1.32$, d.f. = 1, $p = 0.2501$). The inclusion of a species $\times$ consumption interaction effect in the linear predictor for the scale parameter was not significant ($\lambda = 0.43$, d.f. = 2, $p = 0.5109$). Any further reductions yielded significant likelihood-ratio test statistics. Therefore, the selected linear predictors were

$$\mu_{ik} = \beta_0, \quad (3.43)$$

$$\log(\sigma_{ik}) = \beta_0 + s_i + c_k, \quad (3.44)$$

where β_0 is the intercept, s_i is the effect of the i -th species and c_k is the effect of consumption, ($k = 1$ when no prey was effectively consumed and $k = 2$ when there was consumption), see Table 3.6 for parameter estimates. A global test between the maximal model (3.31 and 3.32) and the selected model (3.43 and 3.44) was also not significant ($\lambda = 49.91$, d.f. = 40, $p = 0.1354$).

Figure 3.11. Fitted combined survival curves for the Weibull model for the combination between species and consumption levels

Table 3.6. Parameter estimates and associated standard errors, t values and p -values for the exponentiated-Weibull model fitted to the time until attack data

Parameter	Estimate	Std. error	t value	p -value
$\log(a)$	2.17	0.35	6.15	< 0.0001
μ – Intercept	2.78	0.57	4.87	< 0.0001
σ – Intercept	1.10	0.13	8.65	< 0.0001
σ – Species: Stinkbug	0.47	0.10	4.64	< 0.0001
σ – Consumption: Yes	−0.51	0.14	−3.58	0.0003

Now, the fitted survival curves follow the Kaplan-Meier estimates quite well, see Figure 3.12, and the half-normal plot with simulation envelope also suggests that the model fits the data well, see Figure 3.13(b). Also, there is no significant effect whatsoever for the location part of the model and the only significant effects (which are the same for the exponential and Weibull models) are now for the scale part of the model.

Again, the estimated survival curves for earwigs that did not effectively consume prey in one hour and for stinkbugs that did are fairly close (see Figure 3.13b). Fitting a model that merged these two curves yielded the same number of parameters than the previous model (linear predictors (3.43) and (3.44)) but with a more straightforward interpretation (see Table 3.7). The survival curve for earwigs which effectively consumed prey in one hour is more accelerated, the survival curve for earwigs that did not and stinkbugs that did so is less accelerated and the survival curve for stinkbugs that did not consume prey effectively in one hour is the least accelerated (see Figure 3.14).

Figure 3.12. Fitted survival curves for the exponentiated-Weibull model for (a) each treatment combination and (b) combination between species and consumption levels

Figure 3.13. (a) Deviance residuals versus indexes and (b) half-normal plot with simulation envelope for the exponentiated-Weibull model fit to the time until attack data

Table 3.7. Parameter estimates and associated standard errors, t values and p -values for the exponentiated-Weibull model fitted to the time until attack data

Parameter	Estimate	Std. error	t value	p -value
$\log(a)$	2.19	0.36	6.03	< 0.0001
μ – Intercept	2.75	0.59	4.62	< 0.0001
σ – Species: Earwig x Consumption: Yes	0.61	0.18	3.40	0.0007
σ – Species: Earwig x Consumption: No	1.06	0.10	10.56	< 0.0001
and Species: Stinkbug x Consumption: Yes				
σ – Species: Stinkbug x Consumption: No	1.69	0.22	7.50	< 0.0001

3.5.5 Including day effects as random in the exponentiated-Weibull model

Now we turn to including the day effects in the location linear predictor for the exponentiated-Weibull model. These should be included as random, because three different batches of insects from the same population were used and there is random variation in the conditions of the three days when the experiment was conducted.

The maximum likelihood estimates for the fixed effects are exactly the same as shown in Table 3.6. The estimate for $\sigma_d^2 = 0.0005$ suggests that there is no significant day effect in the experiment. The

Figure 3.14. Fitted combined survival curves for the exponentiated-Weibull model for the combination between species and consumption levels

hypothesis test $H_0 : \sigma_d^2$ versus $H_1 : \sigma_d^2 > 0$ is on the boundary of the parameter space and hence an asymptotic likelihood-ratio test here is done with reference to a mixture of χ^2 distributions: one with 0 degrees of freedom and another with 1 degree of freedom, with a mixing proportion of $\frac{1}{2}$, and for this case $\lambda = 1.54$, with $p = 0.1073$. This was expected since the blocks within day effects were also not significant.

3.5.6 Simulation studies

From the first simulation study, which used the same sample size and covariate values as in the actual experiment, it is possible to see that the model performs reasonably well and estimates σ_d^2 with good precision, see Table 3.8. As expected, when the variance of the random effects is larger, the precision is lower. In the second simulation study, whose mean squared errors for each parameter are shown in Tables 3.9, 3.10, and 3.11, the model performed less well as the proportion of censored observations increased and for smaller sample sizes, as expected. As the variance of the random effects increased, for sample size 36 the precision for the estimates of the location (μ) and shape (a) parameters was very low. When there were more random effects (or groups – days of the experiment) the model seemed to estimate σ_d^2 with better precision.

3.6 Discussion

The conclusions from inference were practically the same for all three models fitted to the data, that is, there is no significant interaction between gender and species and the main difference is between genders. Moreover, when specimens consume prey in one hour this is associated with faster attack times. However, it is clear that the fitted survival curves from the exponentiated-Weibull model describes the

Table 3.8. Mean of parameter estimates and associated mean squared errors (MSE) for the mixed exponentiated-Weibull model fitted to 1000 simulated data sets in the first simulation study

Parameter	$\sigma_d^2 = 0$		$\sigma_d^2 = 1$		$\sigma_d^2 = 3$	
	Mean	MSE	Mean	MSE	Mean	MSE
$\log(a)$	2.46	0.27	2.48	0.32	2.44	0.72
μ – Intercept	2.36	0.77	2.36	0.83	2.45	1.81
σ – Intercept	1.09	0.02	1.08	0.02	0.99	0.07
σ – Species: Stinkbug	0.41	0.01	0.40	0.01	0.39	0.02
σ – Consumption: Yes	−0.40	0.02	−0.40	0.02	−0.39	0.04
σ_d^2	0.07	0.59	0.86	0.71	2.33	0.73

Table 3.9. Mean squared errors (MSE) for each parameter of the mixed exponentiated-Weibull model fitted to 1000 simulated data sets in the second simulation study, with $\sigma_d^2 = 0$

Censoring	Parameter	4 random effects			12 random effects		
		$n = 36$	$n = 132$	$n = 480$	$n = 36$	$n = 132$	$n = 480$
0%	$\log(a)$	2.47	0.30	0.12	3.36	0.34	0.11
	μ – Intercept	10.96	0.89	0.26	12.30	0.96	0.28
	σ – Intercept	0.09	0.01	0.00	0.11	0.01	0.00
	σ – Sp.: Stinkbug	0.04	0.02	0.01	0.05	0.02	0.01
	σ – Cons.: Yes	0.04	0.02	0.02	0.05	0.02	0.02
	σ_d^2	0.03	0.01	0.06	0.13	0.03	0.01
15%	$\log(a)$	3.48	0.40	0.17	3.73	0.43	0.17
	μ – Intercept	19.07	1.33	0.40	16.13	1.34	0.44
	σ – Intercept	0.12	0.03	0.01	0.12	0.02	0.01
	σ – Sp.: Stinkbug	0.04	0.02	0.02	0.05	0.02	0.02
	σ – Cons.: Yes	0.05	0.03	0.02	0.05	0.03	0.02
	σ_d^2	0.04	0.02	0.05	0.16	0.04	0.01
30%	$\log(a)$	4.73	0.55	0.22	4.98	0.53	0.22
	μ – Intercept	28.47	2.17	0.59	23.78	1.94	0.66
	σ – Intercept	0.18	0.05	0.03	0.17	0.05	0.03
	σ – Sp.: Stinkbug	0.06	0.03	0.02	0.06	0.03	0.02
	σ – Cons.: Yes	0.06	0.03	0.02	0.08	0.03	0.02
	σ_d^2	0.05	0.02	0.03	0.22	0.05	0.02

data best (see Figure 3.12). Another important feature of this model is that the significant effects are in the scale parameter, which means that heterogeneity of variance is being modelled. Furthermore, the scale parameter affects the acceleration of the survival function, see Figure 3.15, and hence there is a practical interpretation.

It was clear that the earwig usually attacks earlier than the stinkbug and this is important information from a biological control viewpoint. Previous experiments have shown that this earwig species presents more aggressive behaviour than the stinkbug (Moral et al., 2017). Also, the fact that prey consumption affected the survival curves in the same way for both the parasitised and the non-parasitised prey shows that there is a clear distinction of two groups: specimens that effectively consumed prey in one hour and specimens that did not. This separation between “fast” and “slow” specimens suggests that there is individual variability that must be taken into account when planning biological control strategies. Therefore, further studies on mechanisms of recognition of parasitised prey are particularly relevant in this context. In addition, experiments assessing the competition behaviour of these two predators would also enrich the understanding of the ecological relations among these three species.

Assessing goodness-of-fit for these models is not an easy task and the use of half-normal plots with simulation envelopes was a good alternative, especially through the `hnp` package in R, which allows

Table 3.10. Mean squared errors (MSE) for each parameter of the mixed exponentiated-Weibull model fitted to 1000 simulated data sets in the second simulation study, with $\sigma_d^2 = 1$

Censoring	Parameter	4 random effects			12 random effects		
		$n = 36$	$n = 132$	$n = 480$	$n = 36$	$n = 132$	$n = 480$
0%	$\log(a)$	4.78	0.31	0.11	5.12	0.35	0.11
	μ – Intercept	23.57	0.89	0.25	21.18	1.08	0.27
	σ – Intercept	0.13	0.02	0.00	0.15	0.02	0.00
	σ – Sp.: Stinkbug	0.05	0.02	0.02	0.06	0.02	0.02
	σ – Cons.: Yes	0.05	0.02	0.02	0.07	0.03	0.02
	σ_d^2	0.23	0.18	0.12	0.10	0.06	0.05
15%	$\log(a)$	4.57	0.47	0.15	8.60	0.42	0.15
	μ – Intercept	27.06	1.66	0.47	45.39	1.53	0.37
	σ – Intercept	0.16	0.03	0.01	0.24	0.03	0.01
	σ – Sp.: Stinkbug	0.05	0.03	0.02	0.07	0.03	0.02
	σ – Cons.: Yes	0.06	0.03	0.02	0.08	0.03	0.02
	σ_d^2	0.23	0.19	0.14	0.12	0.06	0.05
30%	$\log(a)$	7.35	0.63	0.20	10.66	0.59	0.20
	μ – Intercept	53.48	2.52	0.74	72.75	2.58	0.57
	σ – Intercept	0.25	0.05	0.02	0.36	0.05	0.02
	σ – Sp.: Stinkbug	0.07	0.03	0.02	0.09	0.03	0.02
	σ – Cons.: Yes	0.07	0.03	0.02	0.10	0.04	0.02
	σ_d^2	0.26	0.19	0.15	0.17	0.07	0.06

Table 3.11. Mean squared errors (MSE) for each parameter of the mixed exponentiated-Weibull model fitted to 1000 simulated data sets in the second simulation study, with $\sigma_d^2 = 3$

Censoring	Parameter	4 random effects			12 random effects		
		$n = 36$	$n = 132$	$n = 480$	$n = 36$	$n = 132$	$n = 480$
0%	$\log(a)$	6.88	0.66	0.06	8.40	0.50	0.06
	μ – Intercept	37.04	2.38	0.23	43.63	1.94	0.27
	σ – Intercept	0.20	0.05	0.02	0.24	0.03	0.02
	σ – Sp.: Stinkbug	0.06	0.02	0.01	0.08	0.03	0.02
	σ – Cons.: Yes	0.06	0.03	0.02	0.09	0.03	0.02
	σ_d^2	1.76	1.23	1.24	0.76	0.70	0.56
15%	$\log(a)$	8.53	0.84	0.09	13.26	0.59	0.08
	μ – Intercept	53.41	3.96	0.34	80.24	2.85	0.28
	σ – Intercept	0.23	0.05	0.02	0.36	0.04	0.01
	σ – Sp.: Stinkbug	0.07	0.03	0.02	0.11	0.03	0.02
	σ – Cons.: Yes	0.07	0.03	0.02	0.11	0.04	0.02
	σ_d^2	1.78	1.23	1.12	0.84	0.69	0.56
30%	$\log(a)$	10.99	1.49	0.26	13.17	0.76	0.12
	μ – Intercept	83.99	8.21	1.21	106.37	5.00	0.52
	σ – Intercept	0.31	0.06	0.03	0.49	0.05	0.01
	σ – Sp.: Stinkbug	0.08	0.03	0.03	0.13	0.04	0.02
	σ – Cons.: Yes	0.09	0.04	0.02	0.15	0.04	0.02
	σ_d^2	1.75	1.23	1.08	0.89	0.71	0.52

for relatively easy implementation, even for more complex models. It was clear that the use of half-normal plots with simulation envelopes alone was not the best way to choose between the three models and that observing how the fitted survival curves accommodated to the data was crucial on deciding which model to use.

When including a random effect in accelerated failure rate models, a common issue that arises is that of convergence problems due to the numerical integration methods used. The most commonly used

Figure 3.15. Survival function for the exponentiated-Weibull location-scale model with $a = 2, \mu = 2$, and different values for σ

methods are the Laplace transformation, Gaussian quadrature and Monte Carlo integration (Ribeiro Jr. et al., 2012). There is, in this context, a trade-off between numerical precision and computation time. The Laplace method is the least computationally intensive but also the least precise. Gaussian quadrature depends upon the number of points used (the higher the number, the higher the precision and computation time). Monte Carlo integration is the most precise method, and also the most computationally intensive option. The main issue with numerical integration methods arise when used concomitantly with different optimization methods. The R software offers a range of optimization methods, see documentation for function `optim` and package `bbmle` (Bolker and R Development Core Team, 2016). See Appendix A for an overview of optimization methods.

References

- AARSET, M. V. How to identify bathtub hazard rate. (1987) **IEEE Transactions on Reliability**, v. 36, p. 106–108.
- AITKIN, M.; FRANCIS, B.; HINDE, J.; DARNELL, R. (2009) **Statistical modelling in R**. Oxford: Oxford University Press. 576 p.
- AL-ZYOD, F.; SENGONCA, C. (2004) Prey consumption preferences of *Serangium parcesetosum* Sicard (Col., Coccinelidae) for different prey stages, species and parasitized prey. **Journal of Pest Science**, v. 77, p. 197–204.
- AMARASEKARE, P. (2003) Competitive coexistence in spatially structured environments: a synthesis. **Ecology Letters**, v. 6, p. 1109–1122.
- ATKINSON, A. C. (1985) **Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis**. Oxford: Clarendon. 282 p.
- BATTEL, A. P. M. B.; MORAL, R. A.; GODOY, W. A. C. (2012) Modelos matemáticos predador-presa e aplicaes ao manejo integrado de pragas. **Oecologia Australis**, v. 16, p. 43–62.
- BOLKER, B; R DEVELOPMENT CORE TEAM. (2016) `bbmle`: Tools for general maximum likelihood estimation. R package version 1.0.18, Available at: <<http://CRAN.R-project.org/=bbmle>>

- BORGAN, Ø. (2005) Kaplan-Meier estimator. In: ARMITAGE, P.; COLTON, T. (Eds.). **Encyclopedia of biostatistics**, John Wiley & Sons.
- BRESLOW, N.; CROWLEY, J. (1974) A large sample study of the life table and product limit estimates under random censorship. **The Annals of Statistics**, v. 2, p. 437–453.
- BROSTRM, G. (2013) glmmML: Generalized linear models with clustering. R package version 1.0, Available at: <<https://CRAN.R-project.org/package=glmmML>>
- BROYDEN, C. G.; DENNIS, J. E.; MORÉ, J. J. (1973) On the local and superlinear convergence of quasi-Newton methods. **Journal of Applied Mathematics**, v. 12, 223–245.
- CARDINALE, B. J.; HARVEY, C. T.; GROSS, K.; IVES, A. R. (2003) Biodiversity and biocontrol: emergent impacts of a multi-enemy assemblage on pest suppression and crop yield in an agroecosystem. **Ecology Letters**, v. 6, 857–865.
- COLOSIMO, E. A.; GIOLO, S. R. (2006) **Análise de sobrevivência aplicada**. São Paulo: Edgard Blücher. 367 p.
- COX, D. R.; SNELL, E. J. (1968) A general definition of residuals. **Journal of the Royal Statistical Society. Series B**, v. 30, p. 248–275.
- CRUZ, I.; FIGUEIREDO, M. L. C.; MATOSO, M. J. (1999) **Controle biológico de *Spodoptera frugiperda* utilizando o parasitoide de ovos *Trichogramma***. Sete Lagoas: EMBRAPA-CNPMS. 40 p.
- DAVIS, P. J.; RABINOWITZ, P. (2007) **Methods of numerical integration**. New York: Academic Press. 612 p.
- DEMÉTRIO, C. G. B.; HINDE, J.; MORAL, R. A. (2014) Models for overdispersed data in entomology. In: FERREIRA, C. P.; GODOY, W. A. C. (Eds.). **Ecological modeling applied to entomology**. Springer. 262 p.
- EINBECK, J.; DARNELL, R.; HINDE, J. (2014) npmlreg: Nonparametric maximum likelihood estimation for random effect models. R package version 0.46-1, Available at: <<https://CRAN.R-project.org/package=npmlreg>>
- HARRELL, F. E. (2014) Cox proportional hazards regression model. In: HARRELL, F. E. (Ed.) **Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis**. Springer. 582 p.
- KAPLAN, E. L.; MEIER, P. (1958) Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, v. 53, p. 457–481.
- KOGAN, M. (1998) Integrated pest management historical perspectives and contemporary developments. **Annual Review of Entomology**, v. 43, 243–270.
- LIU, Q.; PEARCE, D. A. (1994) A note on Gauss-Hermite quadrature. **Biometrika**, v. 81, p. 624–629.
- LIU, L.; YU, Z. (2008) A likelihood reformulation method in non-normal random effects models. **Statistics in Medicine**, v. 27, p. 3105–3124.
- MANTEL, N.; HAENSZEL, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, v. 22, 716–748.

- MATTHEWS, D. E.; FAREWELL, V. T. (2007) The log-rank or Mantel-Haenszel test for the comparison of survival curves. In: MATTHEWS, D. E.; FAREWELL, V. T. (Eds.) **Using and understanding medical statistics**. Karger. 322 p.
- MORAL, R. A.; DEMÉTRIO, C. G. B.; HINDE, J.; GODOY, W. A. C.; FERNANDES, F. S. (2017) Parasitism-mediated prey selectivity in laboratory conditions and implications for biological control. **Basic and Applied Ecology**, v. 19, 67–75.
- MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. B. (2017) hnp: Half-Normal Plots with Simulation Envelopes. R package version 1.2-3, Available at: <<http://CRAN.R-project.org/=hnp>>
- NADARAJAH, S.; CORDEIRO, G. M.; ORTEGA, E. M. M. (2013) The exponentiated Weibull distribution: A survey. **Statistical Papers**, v. 54, p. 839–877.
- OLIVEIRA, H. N.; PRATISSOLI, D.; PEDRUZZI, E. P.; ESPINDULA, M. C. (2004) Desenvolvimento do predador *Podisus nigrispinus* alimentado com *Spodoptera frugiperda* e *Tenebrio molitor*. **Pesquisa Agropecuária Brasileira**, v. 39, p. 947–951.
- OLIVEIRA, J. E. M.; TORRES, J. B.; CARRANO-MOREIRA, A. F.; RAMALHO, F. S. (2002) Biologia de *Podisus nigrispinus* predando lagartas de *Alabama argillacea* em campo. **Pesquisa Agropecuária Brasileira**, v. 37, p. 7–14.
- R CORE TEAM. (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, Available at: <<http://www.R-project.org>>
- RABE-HESKETH, S.; SKRONDAL, A.; PICKLES, A. (2002) Reliable estimation of generalized linear mixed models using adaptive quadrature. **The Stata Journal**, v. 2, p. 1–21.
- REDOAN, A. C.; CRUZ, I.; CARVALHO, G. A.; FIGUEIREDO, M. L. C.; SILVA, R. B. (2010) Seletividade de inseticidas registrados no controle de *Spodoptera frugiperda* em milho para ninfas de *Doru luteipes* e *Euborellia annulipes*. In: CONGRESSO NACIONAL DE MILHO E SORGO, 2010, Goiânia. Goiânia: ABMS. p. 226–228.
- RIBEIRO JR., P. J.; BONAT, W. H.; KRAINSKI, E. T.; ZEVIANI, W. M. (2012) **Métodos computacionais em inferência estatística**. Joo Pessoa: 20º SINAPE. 289 p.
- THERNEAU, T. M.; GRAMBSCH, P. M.; FLEMING, T. R. (1990) Martingale-based residuals for survival models. **Biometrika**, v. 77, p. 147–160.

4 MODELS FOR JOINTLY ESTIMATING ABUNDANCES OF TWO UNMARKED SITE-ASSOCIATED SPECIES SUBJECT TO IMPERFECT DETECTION

Abstract

In ecological field surveys it is often of interest to estimate the abundance of species. It is frequently the case that unmarked animals are counted on different sites over several time occasions. A natural starting point to model these data, while accounting for imperfect detection, is by using Royle's N-mixture model (*Biometrics*, 2004, **60**, 108–115). Subsequently, many multivariate extensions have been proposed to model communities as a whole. However, these approaches are used to study species richness and other community-level variables and do not focus on the relationship between two site-associated species. Here, we extend the N-mixture modelling framework to model two site-associated species abundances jointly and propose to measure the influence of one species' abundance on the populations of the other, and study how this changes over time and space. By including a new parameter in the abundance distribution of one of the species, linking it to abundance of the other, our proposed model treats extra-variability as an effect induced by an associated species' abundance and allows one to study how environmental covariates may affect this. Using results from simulation studies, we show that the model is able to recover true parameter estimates. We illustrate our approach using data from mites and springtails in Colombia and from bald eagles and mallards obtained in the 2015 survey of the North American Breeding Bird Survey. By using the joint model we were able to separate overdispersion from variability induced by an associated species and hence what would be accounted for with a dispersion parameter in the univariate framework was explained by covariates related to the associated species' abundance in the joint model. Our approach represents an attractive, yet simple, way of modelling site-associated species populations jointly. Conservation ecologists can use the approach to devise management strategies based on the strength of association between species, which may be due to direct interactions and/or environmental effects affecting both species' populations. Also, mathematical ecologists can use this framework to develop tools for studying population dynamics under different scenarios.

Keywords: Bald eagle; Binomial-Poisson mixture; Bivariate models; Mite ecology; N-mixture models; Replicated counts; Site association

4.1 Introduction

Ecology researchers emphasize how important it is to study the abundance of species and how it changes over space and time (Hanski, 1998; Ehrlén and Morris, 2015). When dealing with animal counts from unmarked populations, the problem of imperfect detection arises and this has to be accommodated into the modelling framework when analysing these data (IKNAYAN et al., 2014). A simple way of tackling this problem was proposed by Royle (2004), and uses a binomial-Poisson mixture to model species abundance and detection probability for repeated counts, assuming population closure. When the data do not meet this assumption, extensions to this model have been proposed by including a hidden Markov process to model species survival and fecundity rates (see Dail and Madsen, 2011; Hostetler and Chandler, 2015). Over the past ten years, this modelling framework has attracted the attention of conservation and statistical ecologists (80 papers with a total of 508 citations, according to *Web of Science*, access in July 21st, 2016, search for the key-word “N mixture”). This shows that it has been applied to data from a wide range of taxa and has served as a building block for the development of different statistical models and tools to analyse this type of data.

Specific multivariate extensions to N-mixture models have been the focus of many publications

in statistical ecology, such as the joint modelling of a community of unmarked species. For example, a multispecies N-mixture model was used to estimate habitat associations with different species of fishes (Stewart and Long, 2016), another type of model formulation was used to study bird species richness in agroecosystems (Riffell et al., 2015), and models were also proposed to study the abundance and richness of early-successional insect, bird and plant species in Japan (Yamaura et al., 2012). As pointed out by Dénes et al. (2015), there are still few applications of multispecies abundance models and they are generally related to the study of species richness, diversity, similarity and other community- and functional group-level variables and, frequently, use a Bayesian formulation of the model due to problematic maximisation of the likelihood function in many cases.

In ecological studies, species living in a same site can interact directly or indirectly by predation or competition, for example. However, the strength of association between two species can be reduced if they are part of a complex food web (MacCan et al., 1998; Berlow et al., 2004). Changes in abundances of associated species can indicate important aspects of habitat quality (i.e. occurrence of disturbance), community species composition and environmental conditions in different sample sites. In this context, a simple, yet flexible modelling framework that allows for investigating the influence of one species' abundance on another site-associated species, and provides a metric for this influence, can be useful.

To jointly model two site-associated species, we have developed an extension to Royle's N-mixture model that allows for joint estimation of detection probabilities and abundance parameters, and we propose to use the correlation between the two species abundances, subject to environmental covariates, as an influence metric. Knowledge on this could aid on the development of conservation strategies and wildlife management policies, and help to understand specific abundance patterns, given the notable interest in making inferences on ecological relationships for these purposes (Yoccoz et al., 2001). By incorporating the detection probability, this approach can be used to analyse data from many studies in which site-associated species are counted on different sites over several time occasions and provide estimates of abundance and the correlation between the species.

This paper presents a simple framework for modelling two co-occurring species when the interest does not lie in studying community dynamics and richness as a whole, as opposed to other multispecies N-mixture models proposed in the literature (Dénes, 2015). We begin by describing the joint model's formulation and assumptions and show how covariates may be included in the model in such a way that it is possible to study their influence on species abundance, which may also be used to incorporate overdispersion, as discussed later. We then carry out simulation studies to assess the influence of different sample sizes and parameter value settings on model performance. Finally, we apply our proposed model to data on mites and springtails, obtained in Colombia (Ramírez, 2012) and to another dataset on bald eagles and mallards, obtained from the American Breeding Bird Survey (ABBS, Pardieck et al., 2016) for the 2015 campaign on the states of British Columbia, Washington and Oregon, and study the model's predictive power.

4.2 Model formulation and assumptions

Consider a study in which counts are taken of two site-associated species, y_{1it} and y_{2it} , where y_{sit} is the number of individuals of species s , recorded at site i on occasion t . Suppose these observations are made in R different sites and T occasions over time and that the true, latent abundances for each site and species are N_{si} . Assuming sites are independent and the populations are closed, *i.e.* no migration and constant birth and death rates, and also that there is no double-counting of the animals, the standard N-mixture models proposed by Royle (2004) allow for the problem of imperfect detection and hence we assume that we observe N_{si} with probability p_{sit} . Taking the abundances as Poisson distributed, these

models may be written, separately, for each species as

$$\begin{aligned} Y_{sit}|N_{si} &\sim \text{Bin}(N_{si}, p_{sit}) \\ N_{si} &\sim P(\lambda_{si}). \end{aligned}$$

A straightforward extension for modelling two site-associated species data would link the models for both species in such a way that the density of one species is influenced by the other species' population, as well as other factors. By writing

$$\begin{aligned} Y_{1it}|N_{1i} &\sim \text{Bin}(N_{1i}, p_{1it}) \\ N_{1i} &\sim P(\lambda_{1i}) \\ Y_{2it}|N_{1i}, N_{2i} &\sim \text{Bin}(N_{2i}, p_{2it}) \\ N_{2i}|N_{1i} &\sim P(\psi_i + \lambda_{2i}N_{1i}), \end{aligned}$$

with $\psi_i > 0$ and $\lambda_{2i} > 0$, we now have a hierarchical model that allows us to model the abundances of two unmarked site-associated species jointly through the linking parameter of proportionality λ_{2i} .

Abundances for each site can be estimated by obtaining the posterior distribution of N_{1i} and N_{2i} given the observed data using Bayes Rule,

$$P(N_{si} = n_{si} | \{y_{sit}\}) = \frac{P(\{Y_{sit}\} = \{y_{sit}\} | N_{si} = n_{si}) P(N_{si} = n_{si})}{\sum_{n_{si}=0}^{\infty} \{P(\{Y_{sit}\} = \{y_{sit}\} | N_{si} = n_{si}) P(N_{si} = n_{si})\}}. \quad (4.1)$$

Then, the estimates for the abundance at each site are obtained by taking the expected value of the posterior distribution in (4.1), *i.e.*

$$\hat{N}_{si} = E(N_{si} | \{y_{sit}\}) = \sum_{n_{si}=0}^{\infty} n_{si} P(N_{si} = n_{si} | \{y_{sit}\}). \quad (4.2)$$

For the P-P model, (4.1) results in

$$P(N_{1i} = n_{1i} | \{y_{1it}\}) = \frac{\lambda_{1i}^{n_{1i}} \prod_{t=1}^T \left[\frac{p_{1it}^{y_{1it}} (1 - p_{1it})^{n_{1i} - y_{1it}}}{y_{1it}! (n_{1i} - y_{1it})!} \right]}{\sum_{n_{1i}=0}^{\infty} \left\{ \lambda_{1i}^{n_{1i}} \prod_{t=1}^T \left[\frac{p_{1it}^{y_{1it}} (1 - p_{1it})^{n_{1i} - y_{1it}}}{y_{1it}! (n_{1i} - y_{1it})!} \right] \right\}}$$

and

$$P(N_{2i} = n_{2i} | \{y_{2it}\}) = \frac{\prod_{t=1}^T \left[\frac{p_{2it}^{y_{2it}} (1 - p_{2it})^{n_{2i} - y_{2it}}}{y_{2it}! (n_{2i} - y_{2it})!} \right] \sum_{n_{1i}=0}^{\infty} \left[\frac{\lambda_{1i}^{n_{1i}} e^{-(\psi_i + \lambda_{2i}n_{1i})} (\psi_i + \lambda_{2i}n_{1i})^{n_{2i}}}{n_{1i}!} \right]}{\sum_{n_{2i}=0}^{\infty} \left\{ \prod_{t=1}^T \left[\frac{p_{2it}^{y_{2it}} (1 - p_{2it})^{n_{2i} - y_{2it}}}{y_{2it}! (n_{2i} - y_{2it})!} \right] \sum_{n_{1i}=0}^{\infty} \left[\frac{\lambda_{1i}^{n_{1i}} e^{-(\psi_i + \lambda_{2i}n_{1i})} (\psi_i + \lambda_{2i}n_{1i})^{n_{2i}}}{n_{1i}!} \right] \right\}}.$$

Note that the model for the latent influenced species' abundance derives from a Poisson-Poisson mixture, and hence is automatically able to incorporate overdispersion that may be induced from the other species' abundance.

Integrating over the latent abundances, we obtain the unconditional likelihoods for species 1's observations

$$L_1(\mathbf{p}_1, \boldsymbol{\lambda}_1 | \{y_{1it}\}) = \prod_{i=1}^R \left\{ \sum_{N_{1i}=\max_t \{y_{1it}\}}^{\infty} \left(\prod_{t=1}^T \text{Bin}(y_{1it}; N_{1i}, p_{1it}) \right) \text{Pois}(N_{1i}; \lambda_{1i}) \right\}$$

and for the influenced species 2 observations

$$L_2(\mathbf{p}_2, \boldsymbol{\psi}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 | \{y_{2it}\}, \{y_{1it}\}) = \prod_{i=1}^R \left\{ \sum_{N_{2i}=\max_t \{y_{2it}\}}^{\infty} \left(\prod_{t=1}^T \text{Bin}(y_{2it}; N_{2i}, p_{2it}) \right) \right. \\ \left. \times \sum_{N_{1i}=\max_t \{y_{1it}\}}^{\infty} \text{Pois}(N_{2i}; \psi_i, \lambda_{2i}) \text{Pois}(N_{1i}; \lambda_{1i}) \right\}.$$

The joint likelihood for both species observations is

$$L(\mathbf{p}_1, \mathbf{p}_2, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\psi} | \{y_{1it}\}, \{y_{2it}\}) = L_1(\mathbf{p}_1, \boldsymbol{\lambda}_1 | y_{1it}) \times L_2(\mathbf{p}_2, \boldsymbol{\psi}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 | y_{2it}, y_{1it}), \quad (4.3)$$

which, for the case where the latent abundances for both species are distributed as Poisson (hereafter referred to as the P-P model), may be written as

$$L(\mathbf{p}_1, \mathbf{p}_2, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\psi} | \{y_{1it}\}, \{y_{2it}\}) \\ = \prod_{i=1}^R e^{-(2\lambda_{1i} + \psi_i)} \left\{ \sum_{N_{1i}=\max_t \{y_{1it}\}}^{\infty} \left(\lambda_{1i}^{N_{1i}} \left[\prod_{t=1}^T \frac{p_{1it}^{y_{1it}} (1 - p_{1it})^{N_{1i} - y_{1it}}}{y_{1it}! (N_{1i} - y_{1it})!} \right] \right) \right. \\ \left. \times \sum_{N_{1i}=\max_t \{y_{1it}\}}^{\infty} \left(\frac{e^{-\lambda_{2i} N_{1i}} \lambda_{1i}^{N_{1i}}}{N_{1i}!} \sum_{N_{2i}=\max_t \{y_{2it}\}}^{\infty} (\psi_i + \lambda_{2i})^{N_{2i}} \left[\prod_{t=1}^T \frac{p_{2it}^{y_{2it}} (1 - p_{2it})^{N_{2i} - y_{2it}}}{y_{2it}! (N_{2i} - y_{2it})!} \right] \right) \right\}$$

For overdispersed counts, we may also use a negative binomial distribution to model the latent abundances, taking

$$N_{1i} \sim \text{NB}(\lambda_{1i}, \phi_1) \\ N_{2i} | N_{1i} \sim \text{NB}(\psi_i + \lambda_{2i} N_{1i}, \phi_2)$$

and hence we have, initially, four possible combinations for the abundance models, namely P-P (Poisson abundance models for both species), P-NB (Poisson for N_1 and negative binomial for N_2), NB-P (negative binomial for N_1 and Poisson for N_2), and NB-NB (negative binomial for both). Maximising (4.3) requires the approximation of the infinite summations in the likelihood by a finite sum, truncated to K . To avoid estimation problems, Dennis et al. (2015) recommend choosing K such that the Poisson or negative binomial upper tail probability is $< 10^{-10}$. For example, for the P-P model with a maximum observed count of 15, setting $K = 46$ would suffice as this yields an upper tail probability of $7 \times 10^{-11} < 10^{-10}$. Dennis et al. (2015) have also devised a diagnostic based on the sample covariance which is negative when the model produces infinite abundance estimates associated with low detection probabilities. Here, the diagnostic is 0.86 for the springtail dataset, 5.51 for the mite dataset, 0.04 for the mallard dataset, and 0.37 for the bald eagle dataset and, hence, indicates that the N-mixture model should produce reasonable estimates. However, this diagnostic does not take dependence on covariates into account and further research would be needed to study more complex situations.

Our model formulation captures positive ($\lambda_{2i} > 0$) or neutral ($\lambda_{2i} = 0$) impact of one species' abundance on the abundance of the other. The inclusion of parameters ψ_i and λ_{2i} add enough flexibility to the model such that if $\lambda_{2i} = 0$, the species may be independently modelled, while if $\psi_i = 0$ then there is a strong association between them.

The estimates for the detection probabilities p_{1it} are the same as for the univariate model and hence it is assumed that only one species influences the abundance of the other. Although this might seem like a strong assumption, we believe it may serve as a good approximation under many circumstances, such as when predators are generalist and may be found cohabiting with their prey (see Wright, 1953, for an example with bald eagles and breeding ducks), and when the occurrence of disturbances in the habitat

reduces the resources of one or both interacting species (see Carvalho et al., 2010). The abundance parameter λ_{1i} , however, is also part of the contribution of the influenced species' observations to the likelihood (4.3).

4.2.1 Incorporating covariates

In the joint model proposed here it is also possible to study the effects of covariates on the detection probability and abundance parameters for each species through a generalized linear modelling framework. A logit link function may be used to model detection probabilities, as they lie between zero and one, and abundance parameters may be modelled using a log link, as the parameters of the count distributions used are non-negative, *i.e.*

$$\text{logit}(\mathbf{p}_s) = \mathbf{X}_{p_s} \boldsymbol{\beta}_{p_s}, \quad (4.4)$$

$$\log(\boldsymbol{\lambda}_s) = \mathbf{X}_{\lambda_s} \boldsymbol{\beta}_{\lambda_s}, \quad (4.5)$$

$$\log(\boldsymbol{\psi}) = \mathbf{X}_{\psi} \boldsymbol{\beta}_{\psi}, \quad (4.6)$$

with $\mathbf{X}_{p_s}$ the $RT \times q_{p_s}$ design matrices for the detection probabilities with associated $q_{p_s} \times 1$ parameter vectors $\boldsymbol{\beta}_{p_s}$, $\mathbf{X}_{\lambda_s}$ the $R \times q_{\lambda_s}$ design matrices for the abundance parameters λ_{s_i} with associated $q_{\lambda_s} \times 1$ parameter vectors $\boldsymbol{\beta}_{\lambda_s}$, and $\mathbf{X}_{\psi}$ the $R \times q_{\psi}$ design matrix for the abundance parameter ψ_i with associated $q_{\psi} \times 1$ parameter vector $\boldsymbol{\beta}_{\psi}$.

Although we have not done so in the present paper, random observer effects may also be included in the linear predictors (4.4)–(4.6), but in this case the likelihood in (4.3) would now have to be integrated over the random effects to obtain the maximum likelihood estimates of the parameters (for a frequentist approach).

4.2.2 Correlation between species

As the species are linked in the joint model through the parameter λ_{2i} , the correlation between observed abundances $\rho(Y_{1it}, Y_{2it})$ may be derived by using theorems of conditional expectation and variance (see Appendix C) and is presented in Table 4.1. The correlation between the latent abundances N_{1i} and N_{2i} is obtained by setting $p_{1it} = p_{2it} = 1$. Note that, because of the model assumptions and specification, ρ_{it} lies between 0 and 1 and represents the level of influence of one species' density on the abundance of the other. Ecologically, ρ_{it} may be interpreted as a measurement of how one species' abundance induces variability on the abundance of the other species, due to different ecological relations, subject to detection probability, which may change for different covariate values. These relations may be resulted from direct/indirect interactions or environmental influence on trophic cascades, which are the changes in population size of one species driven by changes in the strength of interaction between other species within the same food web.

4.2.3 Simulation studies

To assess model performance simulation studies were carried out under different scenarios. We fixed the true values for the detection probabilities as $p_{1it} = 0.60$ and $p_{2it} = 0.20$. The abundance parameters were fixed as $\lambda_{1i} = 8.00$ and $\psi_i = 2.00$. We chose these values to emulate a study where one species is more abundant and easier to detect than the other. We studied three scenarios: no correlation between the species ($\lambda_{2i} = 0.00$), low correlation, by setting $\lambda_{2i} = 0.25$ ($\rho_{it} = 0.12$), and approximately double this correlation with $\lambda_{2i} = 0.80$ ($\rho_{it} = 0.26$). We simulated under $(R, T) = (100, 2)$ and $(R, T) = (100, 10)$ for the P–P combination of abundance models. We used the root mean square error (RMSE) as a measure of estimator quality and all scenarios were simulated 1000 times.

Table 4.1. Correlation (ρ) between observed species' abundances, Y_{1it} and Y_{2it} , and between latent abundances N_{1i} and N_{2i} , given different abundance model choices

Distribution		$\rho(Y_{1it}, Y_{2it})$
N_{1i}	N_{2i}	
P	P	$\lambda_{2i} \sqrt{\frac{\lambda_{1i} p_{1it} p_{2it}}{\psi_i + \lambda_{1i} \lambda_{2i} (1 + \lambda_{2i} p_{2it})}}$
P	NB	$\lambda_{2i} \sqrt{\frac{\lambda_{1i} p_{1it} p_{2it}}{\psi_i + \lambda_{1i} \lambda_{2i} (1 + \lambda_{2i} p_{2it}) + \frac{p_{2it}}{\phi_2} [\lambda_{1i} \lambda_{2i}^2 + (\psi_i + \lambda_{1i} \lambda_{2i})^2]}}$
NB	P	$\lambda_{2i} \left(1 + \frac{\lambda_{1i}}{\phi_1}\right) \sqrt{\frac{\lambda_{1i} p_{1it} p_{2it}}{\left(1 + \frac{\lambda_{1i} p_{1it}}{\phi_1}\right) [\psi_i + \lambda_{1i} \lambda_{2i} (1 + \lambda_{2i} p_{2it} + \frac{\lambda_{1i} \lambda_{2i} p_{2it}}{\phi_1})]}}$
NB	NB	$\lambda_{2i} \left(1 + \frac{\lambda_{1i}}{\phi_1}\right) \sqrt{\frac{\lambda_{1i} p_{1it} p_{2it}}{\left(1 + \frac{\lambda_{1i} p_{1it}}{\phi_1}\right) \left\{ \psi_i + \lambda_{1i} \lambda_{2i} (1 + \lambda_{2i} p_{2it} + \frac{\lambda_{1i} \lambda_{2i} p_{2it}}{\phi_1} + \frac{\lambda_{1i} \lambda_{2i} p_{2it}}{\phi_1 \phi_2}) + \frac{p_{2it}}{\phi_2} [\lambda_{1i} \lambda_{2i}^2 + (\psi_i + \lambda_{1i} \lambda_{2i})^2] \right\}}}$
Distribution		$\rho(N_{1i}, N_{2i})$
N_{1i}	N_{2i}	
P	P	$\lambda_{2i} \sqrt{\frac{\lambda_{1i}}{\psi_i + \lambda_{1i} \lambda_{2i} (1 + \lambda_{2i})}}$
P	NB	$\lambda_{2i} \sqrt{\frac{\lambda_{1i}}{\psi_i + \lambda_{1i} \lambda_{2i} (1 + \lambda_{2i}) + \frac{1}{\phi_2} [\lambda_{1i} \lambda_{2i}^2 + (\psi_i + \lambda_{1i} \lambda_{2i})^2]}}$
NB	P	$\lambda_{2i} \left(1 + \frac{\lambda_{1i}}{\phi_1}\right) \sqrt{\frac{\lambda_{1i}}{\left(1 + \frac{\lambda_{1i}}{\phi_1}\right) [\psi_i + \lambda_{1i} \lambda_{2i} (1 + \lambda_{2i} + \frac{\lambda_{1i} \lambda_{2i}}{\phi_1})]}}$
NB	NB	$\lambda_{2i} \left(1 + \frac{\lambda_{1i}}{\phi_1}\right) \sqrt{\frac{\lambda_{1i}}{\left(1 + \frac{\lambda_{1i}}{\phi_1}\right) \left\{ \psi_i + \lambda_{1i} \lambda_{2i} (1 + \lambda_{2i} + \frac{\lambda_{1i} \lambda_{2i}}{\phi_1} + \frac{\lambda_{1i} \lambda_{2i}}{\phi_1 \phi_2}) + \frac{1}{\phi_2} [\lambda_{1i} \lambda_{2i}^2 + (\psi_i + \lambda_{1i} \lambda_{2i})^2] \right\}}}$

P, Poisson; NB, negative binomial.

We logit- and log-transformed the detection probabilities and abundance-related parameters to bound the estimation process in the parameter space and prevent nonsensical estimates, i.e. probabilities smaller than zero or larger than one or negative abundance parameters. All implementations were made in the R software (R Core Team, 2017) and optimization was performed using the quasi-Newton algorithm BFGS implemented in the function `optim`. Here we present a selection of the scenarios that we actually explored which highlight typical results; for a complete overview of the full simulation study, see Appendix C.

4.3 Case study: Mite data

We first illustrate our modelling framework by fitting the joint models to data on mites and springtails, interacting species commonly found in soil. We used data obtained by Ramírez (2012), who conducted a sampling study in Colombia. Mites from the mesostigmata order and springtails (Collembola) were sampled bimonthly throughout the year 2010 at nine sites in both a forest patch and in a pasture, totaling $6 \times 9 \times 2 = 108$ observations.

First, we fitted univariate N-mixture models for each species including the effects of mean temperature (centigrade) and relative humidity (%) in both linear predictors for the detection probability and an interaction between ecosystem and area type in the linear predictors for the abundance parameters. All covariates were scaled to have zero mean and unit variance to avoid numerical problems when fitting

the models. We then proceeded to fit the joint model for the four combinations of abundance distributions, *i.e.* P–P, NB–P, P–NB, and NB–NB. Finally, we performed model selection based on likelihood-ratio (LR) tests and obtained the correlation estimates from the selected model. For all fitting procedures, as the maximum number of observed individuals at a given occasion was 145, we set $K = 230$, which yields a tail probability of 1.8×10^{-11} for the Poisson abundance model.

4.4 Case study: Bald eagle data

We also illustrate the applications of our proposed joint model by fitting it to data on bald eagles (*Haliaeetus leucocephalus*) and mallards (*Anas platyrhynchos*), which is one of the many species that is preyed upon by the eagle (Gerrard and Bortolotti, 1998). We used the data obtained by the American Breeding Bird Survey (ABBS, Pardieck et al., 2016) in the states of Oregon, Washington and British Columbia for the 2015 campaign. The ABBS began in 1966 and up until now records bird counts in predefined routes in all states of continental United States and Canada. These datasets are available from the ABBS website (<https://www.pwrc.usgs.gov/bbs/rawdata/>, consulted on 3 June 2016). Environmental variables are also available.

We used the total number of individuals counted every 10 stops per route and we included in our analysis only routes where at least one bald eagle or mallard was observed, as routes with zero counts for both species would be uninformative regarding the correlation between their abundances. This resulted in the use of data from 103 routes, out of which 34 were in British Columbia, 32 in Washington, and 37 in Oregon. These species are breeding or permanent residents in these territories (Ridgely et al., 2003). As we are modelling the totals every 10 stops, the data is aggregated and hence may also be overdispersed (Demétrio et al., 2014).

First, we fitted univariate N-mixture models for each species including the effects of mean temperature (centigrade) and maximum wind speed (Beaufort scale) in both linear predictors for the detection probability and a response surface for latitude and longitude in the linear predictors for the abundance parameters. All covariates were scaled to have zero mean and unit variance to avoid numerical problems when fitting the models. We then proceeded to fit the joint model for the four combinations of abundance distributions, *i.e.* P–P, NB–P, P–NB, and NB–NB. Finally, we performed model selection based on likelihood-ratio (LR) tests, obtained the correlation estimates from the selected model and carried out a leave-one-out cross-validation study to assess the model’s predictive power. For all fitting procedures, as the maximum number of observed individuals at a given occasion was 71, we set $K = 150$, which yields a tail probability of 1.3×10^{-16} for the Poisson abundance model.

4.5 Results

4.5.1 Model performance

The results of the simulation studies showed that the proposed joint model is able to recover the true parameter values from simulated data reasonably well (see Table 4.2). When the number of time occasions was small and the true value of λ_2 was zero, however, ψ was overestimated in several samples (greater than 5 in 13.5% of simulations). When this happened the estimates were always associated with infinite standard errors, as well as a very low $\hat{p}_2$, which suggests some sort of weak identifiability or boundary estimation problem with small samples. As the true value of λ_2 increased, first to 0.25 and then to 0.80, this type of bias was reduced. When the number of time occasions increased, the bias for all estimates was notably reduced, as the root mean square error was smaller. In some simulations there was a problem with flat likelihood surfaces, particularly associated with true parameter values set on the boundary of the parametric space ($\lambda_2 = 0$). With an increasing sample size, there were fewer cases in

which this happened. Moreover, this did not induce greater bias in the correlation estimates, which were close to the true value for all simulated scenarios.

Table 4.2. Estimates and correlation (ρ) mean, median and RMSE for the proposed joint model under six scenarios, which varied true parameter values and number of sites (R) and sampling occasions (T), considering the latent abundances distributed as Poisson for both species (P-P case). Results are based on from 1000 simulations

Parameter	True	Mean	Median	RMSE	True	Mean	Median	RMSE	True	Mean	Median	RMSE
$(R, T) = (100, 2)$												
p_1	0.60	0.60	0.60	0.06	0.60	0.60	0.60	0.06	0.60	0.60	0.60	0.06
p_2	0.20	0.18	0.18	0.11	0.20	0.18	0.18	0.10	0.20	0.19	0.19	0.11
λ_1	8.00	8.08	7.96	0.86	8.00	8.07	7.95	0.90	8.00	8.05	7.96	0.84
ψ	2.00	7.17	1.29	21.84	2.00	5.73	2.31	15.90	2.00	3.38	1.58	8.22
λ_2	0.00	0.24	0.10	0.66	0.25	0.62	0.23	1.30	0.80	1.58	0.80	2.16
ρ	0.00	0.07	0.03	0.10	0.12	0.11	0.10	0.10	0.26	0.23	0.26	0.11
$(R, T) = (100, 10)$												
p_1	0.60	0.60	0.60	0.02	0.60	0.60	0.60	0.02	0.60	0.60	0.60	0.02
p_2	0.20	0.19	0.19	0.03	0.20	0.20	0.20	0.03	0.20	0.20	0.20	0.03
λ_1	8.00	8.01	8.01	0.34	8.00	8.00	7.99	0.35	8.00	8.01	7.99	0.33
ψ	2.00	1.26	1.54	1.02	2.00	2.15	2.23	1.28	2.00	2.34	2.15	1.62
λ_2	0.00	0.10	0.06	0.14	0.25	0.25	0.22	0.20	0.80	0.79	0.77	0.33
ρ	0.00	0.07	0.04	0.09	0.12	0.11	0.11	0.08	0.26	0.24	0.25	0.07

RMSE, root mean square error.

4.5.2 Mite data analysis

The likelihood-ratio test is significant when comparing the NB-P joint model to the NB-P univariate models ($LR = 99.43$ on 4 d.f., $p < 0.01$) and the improvement in the log-likelihood is not significant when compared to the NB-NB joint model ($LR = 1.47$ on 1 d.f., $p = 0.23$). Hence, the model that fits the data best is the NB-P joint model (see Table 4.3). However, looking in isolation at the univariate models, the model that best fits the data is the NB-NB model. This shows that if modelled separately, the effect of springtail density on the abundance of mites is incorporated by the additional dispersion parameter. In contrast, the joint models explain the extra-variability through covariates included in the linear predictor for the linking parameter λ_2 .

The correlation between mite and springtail counts is approximately zero in pasture ecosystems, while for forest patches it is moderately large and does not present a clear trend over time (Figure 4.1). This may be due to the fact that in pastures, the soil usually presents less nutrients, which can interfere in the community dynamics that take place there. Moreover, these changes may induce trophic cascades that can weaken the interaction strength between these species.

4.5.3 Bald eagle data analysis

The likelihood-ratio test is significant when comparing the NB-P joint model to the NB-P univariate models ($LR = 365.94$ on 4 d.f., $p < 0.01$) and the improvement in the log-likelihood is not significant at the 5% level when compared to the NB-NB joint model ($LR = 3.50$ on 1 d.f., $p = 0.06$). Hence, the model that fits the data best is the NB-P joint model (see Table 4.3). However, looking in isolation at the univariate models, the model that best fits the data is the NB-NB model. This shows that if modelled separately, the effect of mallard density on the abundance of eagles is incorporated by the additional dispersion parameter. In contrast, the joint models explain the extra-variability through covariates included in the linear predictor for the co-abundance parameter λ_2 . Both species' abundances are clearly overdispersed and hence the infinite estimates for ψ_i in the P-P model and for λ_{2i} in the P-NB

Table 4.3. Parameter estimates (standard errors) and $-2 \times$ the log-likelihood for the analysis of the mite and springtail data with univariate N-mixture models and joint predator-prey models with the four different abundance distribution combinations P-P, NB-P, NP-NB, and NB-NB. Detection probability covariates and abundance covariates were modelled using a logit and a log link, respectively

Parameter	Univariate models				Joint models			
	P-P	NB-P	P-NB	NB-NB	P-P	NB-P	P-NB	NB-NB
Detection (Springtail - p_{1it})								
Intercept	-3.20 (0.01)	-2.43 (0.01)	-3.20 (0.01)	-2.43 (0.01)	-3.06 (0.01)	-2.43 (0.01)	-3.20 (0.01)	-2.43 (0.01)
Humidity	0.79 (0.00)	0.76 (0.00)	0.79 (0.00)	0.76 (0.00)	0.82 (0.00)	0.76 (0.00)	0.79 (0.00)	0.76 (0.00)
Temperature	-0.34 (0.00)	-0.07 (0.00)	-0.34 (0.00)	-0.07 (0.00)	-0.40 (0.00)	-0.07 (0.00)	-0.34 (0.00)	-0.07 (0.00)
(Mite - p_{2it})								
Intercept	-2.85 (0.01)	-2.85 (0.01)	-2.53 (0.01)	-2.53 (0.01)	-2.66 (0.01)	-2.49 (0.01)	-2.53 (0.01)	-2.50 (0.01)
Humidity	-0.01 (0.00)	-0.01 (0.00)	0.00 (0.00)	0.00 (0.00)	-0.01 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Temperature	0.34 (0.01)	0.34 (0.01)	0.27 (0.01)	0.27 (0.01)	0.31 (0.01)	0.29 (0.01)	0.27 (0.01)	0.28 (0.01)
Abundance (Springtail - λ_{1i})								
Intercept	3.34 (0.03)	2.95 (0.16)	3.34 (0.03)	2.95 (0.16)	3.12 (0.02)	2.95 (0.16)	3.35 (0.03)	2.95 (0.15)
Ecosystem	-0.22 (0.05)	-0.70 (0.33)	-0.22 (0.05)	-0.70 (0.33)	-0.11 (0.05)	-0.70 (0.32)	-0.22 (0.05)	-0.70 (0.32)
Area	0.81 (0.02)	0.85 (0.30)	0.81 (0.02)	0.85 (0.30)	0.82 (0.02)	0.85 (0.29)	0.81 (0.02)	0.85 (0.29)
Ecosystem $\times$ Area	0.78 (0.05)	0.55 (0.63)	0.78 (0.05)	0.55 (0.63)	0.82 (0.05)	0.53 (0.58)	0.78 (0.05)	0.55 (0.59)
(Mite - ψ_i)								
Intercept	4.58 (0.01)	4.58 (0.01)	4.23 (0.04)	4.23 (0.04)	-11.73 (10^6)	3.60 (0.06)	3.03 (37)	3.61 (0.06)
Ecosystem	-2.21 (0.05)	-2.21 (0.05)	-2.09 (0.10)	-2.09 (0.10)	13.95 (10^6)	-1.52 (0.03)	-4.66 (188)	-1.64 (0.35)
Area	0.59 (0.01)	0.59 (0.01)	0.56 (0.06)	0.56 (0.06)	-5.26 (10^7)	0.00 (0.14)	0.63 (76)	0.00 (0.14)
Ecosystem $\times$ Area	-0.15 (0.06)	-0.15 (0.06)	-0.11 (0.17)	-0.11 (0.17)	5.66 (10^7)	0.44 (0.12)	0.12 (1154)	0.06 (0.66)
(Mite - λ_{2i})								
Intercept	—	—	—	—	1.29 (0.04)	0.48 (0.37)	0.54 (6.85)	0.48 (0.36)
Ecosystem	—	—	—	—	-8.54 (10^4)	-4.88 (328)	-1.53 (6.80)	-2.53 (9.65)
Area	—	—	—	—	-0.25 (0.03)	0.12 (0.64)	-0.28 (19)	0.12 (0.64)
Ecosystem $\times$ Area	—	—	—	—	2.0 (10^4)	-3.66 (10^4)	-0.87 (20)	-0.03 (10)
(Dispersion)								
ϕ_1	—	0.78 (0.04)	—	0.78 (0.04)	—	0.82 (0.04)	—	0.83 (0.04)
ϕ_2	—	—	4.34 (0.46)	4.34 (0.46)	—	—	4.57 (0.43)	8746 (10^7)
$-2 \times \text{loglik}$	3789.50	3214.49	3689.00	3113.99	3733.58	3115.06	3689.17	3113.59

P, Poisson; NB, negative binomial.

model when fitting the joint model, but well-behaved estimates for the NB-P and NB-NB models. When infinite parameter estimates are obtained these typically have infinite standard errors, and other numerical instabilities can arise from the computation of the inverse of the information matrix. Here, these infinite values may be due to model misspecification, as the two affected models fail to take into account the clear additional overdispersion in the mallard data, and this yields unstable parameter estimates. However, although the parameter estimates are infinite, the fitted values remain sensible in both situations.

The correlation between bald eagle and mallard counts is larger in Oregon and Washington. (Figure 4.2). This may be due to the greater sparseness of routes in British Columbia, given that we are using a response surface for latitude and longitude in the linear predictor for the abundance parameters. Moreover, the influence of mallard abundance increases as the latitude rises (see Figure 4.3(a)), as the number of species and complexity of food webs decrease due to the cold weather (Gaston, 2000; Begon et al., 2005) and hence our model captures this effect and we are able to explain eagle extra-variability using the abundance of an associated species. However, as expected mallard influence does not seem to change as strongly across longitudes (see Figure 4.3(b)).

The results of the leave-one-out cross-validation (LOOCV) study showed that the selected model (NB-P) has good predictive power (see Figure 4.4). The cases in which the LOOCV model estimated a much smaller correlation than the full model were particularly associated with routes where many animals

Figure 4.1. Correlation between mite and springtail counts, predicted by the NB-P joint model, for each ecosystem and area combination

were observed for one of the species, *e.g.* for the single point lying outside of the confidence bands, nine mallards were counted as opposed to 68 bald eagles. Therefore, these points may be acting as leverage for the estimated correlation in the full model for the surrounding points. To assess how large their influence was, we removed them from the dataset and refitted all models, but the results were qualitatively similar.

Also, we observe that some of the estimates for Lat×Long present large standard errors, when compared to the estimate. We proceeded to refit all models without the Lat×Long terms and compare them using likelihood-ratio tests. All tests, except for the P-P joint model, were non-significant. However, when carrying out model selection without these terms, the NB-P joint model was selected again as the one which best fits the data.

4.6 Discussion

We have proposed an extension for the basic N-mixture model (Royle, 2004) for jointly modelling two site-associated species' abundances, by including a co-abundance parameter, λ_2 , which is affected by the abundance of one of the species in the distribution of the latent abundances of the other. We presented a simple extension for closed populations that allows for the computation of the correlation between the species at different sites and time occasions while accounting for the effect of covariates.

Our approach represents an attractive and simple framework for studying the effect of a site-associated species' abundance on unmarked populations of another. As a feature of model formulation, λ_2 represents part of the variation in the influenced species' abundance explained by another species' density. If $\lambda_2 = 0$ there is no influence and the correlation is also equal to zero. This is easily tested through likelihood-ratio tests of nested models. If $\psi_i = 0$ this means that the abundance of one species may be modelled as a proportion of the abundance of the other. Note, however, that this does not necessarily imply a specialist association between them. Moreover, interesting extensions of the proposed model may arise if we allow for the distribution of N_1 to depend on N_2 as well, yielding in a symmetric relationship, *i.e.* both species influence each other. This could be useful in a predator-prey context such as the one presented in this paper, specially when developing dynamic models to study temporal trends in the species' populations. We may also consider a model formulation including a multiplication of N_1 and N_2 to model abundance, when extending this modelling framework to contemplate the law of mass

Table 4.4. Parameter estimates (standard errors) and $-2 \times$ the log-likelihood for the analysis of the bald eagle and mallard data with univariate N-mixture models and joint predator-prey models with the four different abundance distribution combinations P-P, NB-P, NP-NB, and NB-NB. Detection probability covariates and abundance covariates were modelled using a logit and a log link, respectively

Parameter	Univariate models				Joint models			
	P-P	NB-P	P-NB	NB-NB	P-P	NB-P	P-NB	NB-NB
Detection (Mallard - p_{1it})								
Intercept	-4.49 (0.03)	-3.17 (0.02)	-4.49 (0.03)	-3.17 (0.02)	-2.69 (0.02)	-3.14 (0.02)	-4.49 (0.03)	-3.15 (0.02)
Temperature	0.25 (0.00)	0.36 (0.02)	0.25 (0.00)	0.36 (0.02)	0.28 (0.01)	0.36 (0.02)	0.25 (0.00)	0.36 (0.02)
Wind speed	-0.15 (0.00)	-0.28 (0.02)	-0.15 (0.00)	-0.28 (0.02)	-0.11 (0.00)	-0.29 (0.02)	-0.15 (0.00)	-0.28 (0.02)
(Bald eagle - p_{2it})								
Intercept	-4.09 (0.05)	-4.09 (0.05)	-2.88 (0.02)	-2.88 (0.02)	-3.07 (0.02)	-2.87 (0.02)	-2.89 (0.02)	-2.88 (0.02)
Temperature	0.00 (0.00)	0.00 (0.00)	-0.15 (0.03)	-0.15 (0.03)	-0.04 (0.01)	-0.13 (0.03)	-0.19 (0.03)	-0.11 (0.03)
Wind speed	0.14 (0.00)	0.14 (0.00)	0.34 (0.02)	0.34 (0.02)	0.11 (0.01)	0.33 (0.02)	0.34 (0.02)	0.37 (0.02)
Abundance (Mallard - λ_{1i})								
Intercept	3.92 (0.03)	2.72 (0.04)	3.92 (0.03)	2.72 (0.04)	2.10 (0.02)	2.63 (0.03)	3.92 (0.03)	2.67 (0.03)
Latitude	0.22 (0.00)	0.06 (0.02)	0.22 (0.00)	0.06 (0.02)	0.37 (0.00)	0.05 (0.03)	0.22 (0.00)	0.05 (0.02)
Longitude	0.32 (0.00)	0.17 (0.03)	0.32 (0.00)	0.17 (0.03)	0.30 (0.01)	0.15 (0.03)	0.32 (0.00)	0.15 (0.02)
Lat×Long	0.04 (0.00)	0.09 (0.02)	0.04 (0.00)	0.09 (0.02)	0.14 (0.01)	0.08 (0.02)	0.04 (0.00)	0.09 (0.02)
(Bald eagle - ψ_i)								
Intercept	3.22 (0.06)	3.22 (0.06)	2.14 (0.05)	2.14 (0.05)	-22.50 (350)	-0.76 (0.75)	2.04 (0.05)	-0.48 (0.78)
Latitude	-0.86 (0.01)	-0.86 (0.01)	-0.86 (0.04)	-0.86 (0.04)	-1.86 (3169)	-1.59 (0.47)	-0.99 (0.04)	-1.65 (0.61)
Longitude	-0.13 (0.01)	-0.13 (0.01)	-0.23 (0.04)	-0.23 (0.04)	12.37 (118)	-0.67 (0.20)	-0.42 (0.05)	-0.94 (0.27)
Lat×Long	-0.12 (0.00)	-0.12 (0.00)	0.13 (0.04)	0.13 (0.04)	0.91 (1215)	-0.16 (0.12)	0.05 (0.04)	-0.01 (0.16)
(Bald eagle - λ_{2i})								
Intercept	—	—	—	—	0.12 (0.04)	-0.53 (0.06)	-28.96 (2216)	-0.69 (0.10)
Latitude	—	—	—	—	-1.37 (0.02)	-0.83 (0.07)	-6.58 (14776)	-0.78 (0.09)
Longitude	—	—	—	—	-0.59 (0.02)	-0.34 (0.06)	13.05 (686)	-0.27 (0.08)
Lat×Long	—	—	—	—	-0.31 (0.02)	0.04 (0.07)	3.49 (5507)	0.07 (0.08)
(Dispersion)								
ϕ_1	—	0.51 (0.02)	—	0.51 (0.02)	—	0.39 (0.02)	—	0.46 (0.02)
ϕ_2	—	—	0.46 (0.03)	0.46 (0.03)	—	—	0.48 (0.02)	1.71 (0.05)
$-2 \times \loglik$	3350.28	2986.90	2994.67	2631.29	3196.88	2620.96	2990.69	2617.46

P, Poisson; NB, negative binomial.

action, such as in an epidemiology setting (*e.g.* a SIR model) or standard predator-prey dynamic models (*e.g.* the Lotka-Volterra model).

The method can also be applied to study two interacting species in local communities. The association strength between two species can help to understand the effects of habitat quality, food web complexities and the relationship between biodiversity and ecosystem function. For example, in low quality habitats, the biodiversity loss may propagate through a web of species interactions influencing the abundance correlations between two particular species (Schmitz et al., 2010). Loss of predator species can lead to higher herbivore abundance and thus lower plant biomass or plant productivity. In this case, a higher correlation between herbivore-plant abundances can be understood as a consequence of a variable number of predators across different sample sites. The joint model may also be applied to analyse data from studies in which predators and prey from the same food web are sampled (*e.g.* Brodie and Giordano, 2013; Hua et al., 2016).

The choice of the abundance model is flexible and there are many ways of accommodating overdispersion. An attractive way of capturing extra variability, unexplained by covariates, is by including a dispersion parameter such as in the negative binomial or Neyman type-A distributions. This approach brings a direct interpretation from the ecological point of view regarding the distribution of the abundances of species. The joint model, however, treats part of the variability for the influenced species' counts as

Figure 4.2. Correlation between bald eagle and mallard counts, predicted by the NB-P joint model, on each route where at least one of them was counted, in the 2015 ABBS survey for the states of Oregon, Washington and British Columbia

induced by another species and hence provides a way of studying this influence through covariates, instead of incorporating it all in the dispersion parameter. By including the co-abundance parameter λ_2 in our proposed model, extra-Poisson variability may be incorporated (for the P-P model, for example, the model formulation resembles the Neyman type-A overdispersion model). For that reason, when fitting the joint model, we may find that the Poisson model for the influenced species' abundance is sufficient to incorporate extra variability even though when fitting a univariate model for the same data, the negative binomial abundance model would provide a significantly better fit.

One drawback of our approach is that identifiability issues may arise for small sample sizes and hence the estimates may be biased. The same might happen if the closure assumption is violated. This is actually a feature of the basic N-mixture model (see Dennis et al., 2015) that may have an even more pronounced influence on the parameter estimates of the joint model due to the greater flexibility introduced by the combined use of parameters ψ and λ_2 . An interesting alternative is to use the criterion devised by Dennis et al. (2015) so as to check whether model estimates can be considered reasonable.

Figure 4.3. Natural logarithm of estimates of λ_2 (red) and ψ (blue) by the NB-P joint model fitted to the bald eagle and mallard data over (a) latitude and (b) longitude

Figure 4.4. Estimated correlations from full model fitted to the bald eagle data (solid black line), sorted from lowest to highest, with 95% Wald confidence interval bands and correlations estimated from leave-one-out cross-validation (LOOCV) models (solid blue line and points), based on the joint NB-P model

Also, the computational time for fitting the joint model exceeds the combined time of fitting two univariate models. Depending on the sample size and especially on the truncation threshold K , the optimization procedure may take from tens of minutes to several hours of computation time on a machine equipped with an Intel i5 processor and 4GB RAM. This, however, is also the case for MCMC sampling under the Bayesian framework, especially when working with highly flexible hierarchical models.

Alternatively to what we have proposed, one might include the latent abundances of one species

as a covariate in the linear predictor for the abundance model of the other species (see, for example, Roth et al., 2016). This induces correlations that range from -1 to 1 and is able to capture the effect of species displacement given the abundance of a competitor and avoidance behavior, for example. Likelihood maximisation for this problem becomes more complex due to the fact that the latent abundance random variables are included in the likelihood in a nonlinear fashion and hence Roth et al. (2016) opted for assuming prior distributions for the parameters and sample the posterior distribution using MCMC algorithms in a Bayesian framework. This has also been the case when fitting multispecies models for count or occupancy data (see, for example, Yamaura et al., 2012; Dorazio and Connor, 2014; Riffell et al., 2015; Broms et al., 2016; Ovaskainen et al., 2016a, b; Yamaura et al., 2016). The objective, however, when fitting the so-called multispecies occupancy or abundance models is to estimate species diversity, richness and other types of metrics from a community perspective and is hence different from our approach. Dénes et al. (2015) and Guillera-Arroita (2017) provide an overview of these methods and their usage.

Finally, natural extensions to the proposed joint model are the inclusion of random effects to reflect ecological assumptions, such as the use of different observers per site/occasion, and the relaxation of the closure assumption (see Dail and Madsen, 2011; Hostetler and Chandler, 2015). There is a trade-off, however, between relaxing assumptions and computation time, which has to be considered for these high-dimensional optimization problems. Another extension would be to condition the probability of detection of one species on the latent abundance of the other, which may be true in many predator-prey scenarios in which prey hide when predators are around. Moreover, the double-counting issue is not accounted for in our approach. While we may assume that the rate of double-counting in the ABBS is low because of their well-established observation protocols, this can become an issue and Johnson (1995) provides an interesting discussion on the matter. In this context, the use of diagnostic measures such as standardized residuals can become an important tool for assessing goodness-of-fit of a particular model. A normal plot with a simulation envelope could be helpful to check whether, for a particular model fit, the observed residuals are within what we would expect, should that be the true model describing the ecological process. This is subject of ongoing work, as well as bivariate plots to identify possible outliers in the joint modelling approach, as proposed in this paper.

References

- BEGON, M.; TOWNSEND, C. R.; HARPER, J. L. (2005) **Ecology: from individuals to ecosystems**. Blackwell Publishing, Oxford.
- BERLOW, E. L.; NEUTEL, A.; COHEN, J. E.; DE RUITER, P. C.; EBENMAN, B.; EMMERSON, M.; FOX, J. W.; JANSEN, V. A. A.; JONES, J. I.; KOKKORIS, G. D.; LOGOFET, D. O.; MCKANE, A. J.; MONTOYA, J. M.; PETCHEY, O. (2004) Interaction strengths in food webs: issues and opportunities. **Journal of Animal Ecology**, v. 75, p. 585–598.
- BRODIE, J. F.; GIORDANO, A. (2013) Lack of trophic release with large mammal predators and prey in Borneo. **Biological Conservation**, v. 163, p. 58–67.
- BROMS, K. M.; HOOTEN, M. B.; FITZPATRICK, R. M. (2016) Model selection and assessment for multi-species occupancy models. **Ecology**, v. 97, p. 1759–1770.
- CARVALHEIRO, L. G.; BUCKLEY, Y. M.; MEMMOTT, J. (2010) Diet breadth influences how the impact of invasive plants is propagated through food webs. **Ecology**, v. 91, p. 1063–1074.
- DAIL, D.; MADSEN, L. (2011) Models for estimating abundance from repeated counts of an open metapopulation. **Biometrics**, v. 67, p. 577–587.

- DEMÉTRIO, C. G. B.; HINDE, J.; MORAL, R. A. (2014) Models for overdispersed data in entomology. In: FERREIRA, C. P.; GODOY, W. A. C. (2014) **Ecological modelling applied to entomology**. Springer International Publishing.
- DÉNES, F. V.; SILVEIRE, L. F.; BEISSINGER, S. R. (2015) Estimating abundance of unmarked animal populations: accounting for imperfect detection and other sources of zero inflation. **Methods in Ecology and Evolution**, v. 6, p. 543–556.
- DENNIS, E. B.; MORGAN, B. J. T.; RIDOUT, M. S. (2015) Computational aspects of N-mixture models. **Biometrics**, v. 71, p. 237–246.
- DORAZIO, R. M.; CONNOR, E. F. (2014) Estimating abundances of interacting species using morphological traits, foraging guilds, and habitat. **PloS one**, v. 9, e94323.
- EHRLÉN, J.; MORRIS, W. F. (2015) Predicting changes in the distribution and abundance of species under environmental change. **Ecology Letters**, v. 18, p. 303–314.
- GASTON, K. J. (2000) Global patterns in biodiversity. **Nature**, v. 405, p. 220–227.
- GUILLERA-ARROITA, G. (2017) Modelling of species distributions, range dynamics and communities under imperfect detection: Advances, challenges and opportunities. **Ecography**, v. 40, p. 281–295.
- HANSKI, I. (1998) Metapopulation dynamics. **Nature**, v. 396, p. 41–49.
- HUA, F.; FLETCHER Jr., R. J.; SIEVING, K. E.; DORAZIO, R. M. (2013) Too risky to settle: avian community structure changes in response to perceived predation risks on adults and offspring. **Proceedings of the Royal Society B**, v. 280, 20130762.
- HOSTETLER, J. A.; CHANDLER, R. B. (2015) Improved state-space models for inference about spatial and temporal variation in abundance from count data. **Ecology**, v. 96, p. 1713–1723.
- GERRARD, J. M.; BORTOLOTTI, G. R. (1988) **The bald eagle: Haunts and habits of a wilderness monarch**. Smithsonian books, Washington.
- IKNAYAN, K. J.; TINGLEY, M. W.; FURNAS, B. J.; BEISSINGER, S. R. (2014) Detecting diversity: emerging methods to estimate species diversity. **Trends in Ecology and Evolution**, v. 29, p. 97–106.
- JOHNSON, D. H. (1995) Point counts of birds: What are we estimating? In: RALPH, C. J.; SAUER, J. R.; DROEGE, S. (1995) **Monitoring bird populations by point counts**. United States Forest Service General Technical Report PSW-GTR-149.
- MACCAN, K.; HASTINGS, A.; HUXEL, G. R. (1998) Weak trophic interactions and the balance of nature. **Nature**, v. 395, p. 794–798.
- OVASKAINEN, O.; ABREGO, N.; HALME, P.; DUNSON, D. (2016a) Using latent variable models to identify large networks of species-to-species associations at different spatial scales. **Methods in Ecology and Evolution**, v. 7, p. 549–555.
- OVASKAINEN, O.; ROY, D. B.; FOX, R.; ANDERSON, B. J. (2016b) Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. **Methods in Ecology and Evolution**, v. 7, p. 428–436.
- PARDIECK, K. L.; ZIOLKOWSKI Jr., D. J.; HUDSON, M. A. R.; CAMPBELL, K. (2016) North American breeding bird survey dataset 1966 - 2015, version 2015.0. U.S. Geological Survey, Patuxent Wildlife Research Center. URL <www.pwrc.usgs.gov/BBS/RawData/>; doi:10.5066/F71R6NK8.

R CORE TEAM (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. URL <http://www.R-project.org>

REFAEILZADEH, P.; TANG, L.; LIU, H. (2009) Cross-validation. **Encyclopedia of Database Systems**, p. 532–538.

RIDGELY, R. S.; ALLNUTT, T. F.; BROOKS, T.; MCNICOL, D. K.; MEHLMAN, D. W.; YOUNG, B. E.; ZOOK, J. R. (2003) Digital distribution maps of the birds of the Western hemisphere, version 1.0. NatureServe, Arlington, Virginia, USA.

RIFFELL, S. K.; MONROE, A. P.; MARTIN, J. A.; EVANS, K. O.; BURGER Jr., L. W.; SMITH, M. D. (2015) Response of non-grassland avian guilds to adjacent herbaceous field buffers: testing the configuration of targeted conservation practices in agricultural landscapes. **Journal of Applied Ecology**, v. 52, p. 300–309.

ROTH, T.; BÜHLER; AMRHEIN, V. (2016) Estimating effects of species interactions on populations of endangered species. **The American Naturalist**, v. 187, p. 457–467.

ROYLE, J. A. (2004) N-mixture models for estimating population size from spatially replicated counts. **Biometrics**, v. 60, p. 108–115.

SCHMITZ, O. J.; HAMBACK, P. A.; BECKERMAN, A. P. (2000) Trophic cascades in terrestrial systems: a review of the effects of carnivore removals on plants. **American Naturalist**, v. 155, p. 141–153.

STEWART, D. R.; LONG, J. M. (2016) Using hierarchical Bayesian multispecies mixture models to estimate tandem hoop-net-based habitat associations and detection probabilities of fishes in reservoirs. **Transactions of the American Fisheries Society**, v. 145, p. 450–461.

WRIGHT, B. S. (1953) The relation of bald eagles to breeding ducks in New Brunswick. **The Journal of Wildlife Management**, v. 17, p. 55–62.

YAMAURA, Y.; ROYLE, J. A.; SHIMADA, N.; ASANUMA, S.; SATO, T.; TAKI, H.; MAKINO, S. (2012) Biodiversity of man-made open habitats in an underused country: a class of multispecies abundance models for count data. **Biodiversity Conservation**, v. 21, p. 1365–1380.

YAMAURA, Y.; KÉRY, M.; ROYLE, J. A. (2016) Study of biological communities subject to imperfect detection: bias and precision of community N-mixture abundance models in small-sample situations. **Ecological Research**, v. 31, p. 289–305.

YOCCOZ, N. G.; NICHOLS, J. D.; BOULINIER, T. (2001) Monitoring of biological diversity in space and time. **Trends in Ecology and Evolution**, v. 16, p. 446–453.

5 HALF-NORMAL PLOTS AND OVERDISPERSED MODELS IN R: THE HNP PACKAGE

Abstract

Count and proportion data may present overdispersion, i.e., greater variability than expected by the Poisson and binomial models, respectively. Different extended generalized linear models that allow for overdispersion may be used to analyse this type of data, such as models that use a generalized variance function, random-effects models, zero-inflated models and compound distribution models. Assessing goodness-of-fit and verifying assumptions of these models is not an easy task and the use of half-normal plots with a simulated envelope is a possible solution for this problem. These plots are a useful indicator of goodness-of-fit that may be used with any generalized linear model and extensions. For GLIM users, functions that generated these plots were widely used, however, in the open-source software R, these functions were not yet available on the Comprehensive R Archive Network (CRAN). We describe a new package in R, **hnp**, that may be used to generate the half-normal plot with a simulated envelope for residuals from different types of models. The function **hnp()** can be used together with a range of different model fitting packages in R that extend the basic generalized linear model fitting in **glm()** and is written so that it is relatively easy to extend it to new model classes and different diagnostics. We illustrate its use on a range of examples, including continuous and discrete responses, and show how it can be used to inform model selection and diagnose overdispersion.

Keywords: Goodness-of-fit; Generalized linear models; Mixed models; R

5.1 Introduction

An important step of statistical modelling of any sort is to perform diagnostic analyses to assess goodness-of-fit. Several problems arise when model assumptions are not met, such as misleading estimates, standard errors and p -values and/or wrong conclusions about the process being studied. Among the reasons for a poorly-fitted model, the most common ones are

- measurement error of observed and/or explanatory variables (including typos);
- incomplete or inadequate linear predictor to describe the systematic structure of the data;
- incorrect specification of the error distribution or link function;
- unmodelled overdispersion;
- combination of one or more of the above.

When fitting linear models under the normality assumption, goodness-of-fit can be checked using formal tests, such as the Shapiro-Wilk test for residual normality (Shapiro and Wilk, 1965), or the Bartlett test for variance homogeneity (Bartlett, 1937). However, these tests may fail under many circumstances, such as small sample sizes, and usually graphical techniques provide a better assessment for model goodness-of-fit. These techniques include plotting different types of residuals or influence measures (e.g., leverage and Cook's distance). Several types of residuals may be used (e.g., standardized residuals, studentized residuals, deletion residuals, Pearson residuals, deviance residuals, etc). However, for other types of model, diagnostic checking may be problematic. Useful diagnostic-checking plots include

- residuals vs. explanatory variables — indicates whether higher-order terms should be included in the linear predictor or the need for transformation of the response and/or explanatory variables (for quantitative explanatory variables);
- residuals vs. explanatory variables not included in the model — a systematic relationship indicates that the explanatory variable should be included in the linear predictor;
- added-variable plot — detects the relationship of the response variable with an explanatory variable not included in the model allowing for the effects of other variables;
- residuals vs. fitted values — may reveal variance heterogeneity and/or outliers;
- (half-)normal plots of residuals — detects outliers and indicates whether the error distribution was specified appropriately.

Under a normally distributed error assumption, when a model fits the data well, it is expected that studentized residuals follow a t distribution on the residual degrees of freedom, which for large samples converges to the standard normal distribution. In this case, the half-normal plot of these residuals should show a straight 45° line. However, it is hard to interpret whether points are sufficiently aligned or if the inevitable irregularities are caused by something other than random fluctuations. Another difficulty posed by these plots is that the ordering of the observations may induce dependence. For generalized linear models, this plot may have several forms depending on the variance and link functions and response variable distribution. Atkinson (1985) proposes the addition of a simulated envelope so that interpretation is more straightforward. For a well-fitted model the envelope is such that model diagnostics are likely to fall within it. The purpose is not to provide a region for acceptance or rejection of observations but to serve as a guide of what to expect under a well-fitted model. These plots are also useful for detecting possible outliers, overdispersion, and if the link function and/or error distribution were properly specified (Demétrio et al., 2014).

When fitting generalized linear models, or different types of extended models (e.g., zero-inflated models and mixed models), half-normal plots with simulated envelopes are useful to assess goodness-of-fit, especially when analyzing overdispersed data. Demétrio and Hinde (1997) wrote **GLIM4** macros to produce these plots for different overdispersion models. We developed the R package **hnp** that provides functions for generating half-normal plots with a simulated envelope for a range of generalized linear models and extensions. The scope of the **hnp()** function can be easily extended to include different diagnostics and models by the user-specification of appropriate simulation, model fitting, and diagnostic extraction codes.

5.2 Generalized linear models and overdispersion

The generalized linear models (GLM) framework of statistical modelling, formulated by Nelder and Wedderburn (1972), provides a unified theory for the application of normal, binomial, Poisson, gamma and inverse Gaussian regression models and brings together methods for the analysis of count, proportion, continuous measurement and time to event data, that until then were studied separately. As well as building on the standard normal regression model, this theory also generalized the ideas of analysis of variance through the analysis of deviance.

For random variables $Y_1, \dots, Y_n$, with probability function (pf) (Y_i discrete) or probability density function (pdf) (Y_i continuous) $f(y_i; \theta_i)$, where θ_i is associated with explanatory variables $x_1, \dots, x_p$, statistical modelling will typically be based on a random sample of n observations $(y_i; \mathbf{x}_i)$, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$. A GLM is defined in terms of three basic components. The first component, called the random component, is represented by the random variables $Y_1, \dots, Y_n$ following a distribution which belongs to the

exponential family of distributions and differs only in terms of a parameter θ_i . Writing the exponential family in canonical form, the pf or pdf is given by:

$$f(y_i; \theta_i, \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\}, \quad (5.1)$$

where $b(\cdot)$ and $c(\cdot)$ are known functions, $\phi > 0$ is a known dispersion parameter and θ_i is called canonical parameter. The normal, Poisson, binomial, gamma, inverse Gaussian and negative binomial distributions (with suitable definitions of dispersion parameters) all belong to the exponential family and can be expressed in the canonical form (5.1). For this family of distributions we have that $E(Y_i) = \mu_i = b'(\theta_i)$ and $\text{Var}(Y_i) = \phi b''(\theta_i) = \phi V_i$, where $V_i = V(\mu_i) = d\mu_i/d\theta_i$ is called variance function and depends only on the mean μ_i .

Writing $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ as the design (model) matrix and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)^\top$ as the vector of unknown parameters, the second component is a linear predictor $\eta = \mathbf{X}\boldsymbol{\beta}$, called the systematic component. Finally, the third component, the link function $g(\cdot)$, which is a monotonic and differentiable function, links the random and systematic components by relating the mean to the linear predictor through $\eta_i = g(\mu_i)$.

Nelder and Wedderburn (1972) proposed the scaled deviance as an overall discrepancy quantity to measure the fit of a GLM. This is given by

$$S_p = 2(\hat{l}_n - \hat{l}_p), \quad (5.2)$$

where $\hat{l}_p$ and $\hat{l}_n$ are the maximum of the log-likelihood functions for the current and saturated models, respectively, and the saturated model is one whose fit reproduces the observed data and typically has n parameters. These maximized likelihoods may be written as

$$\hat{l}_n = \phi^{-1} \sum_{i=1}^n [y_i \tilde{\theta}_i - b(\tilde{\theta}_i)] + \sum_{i=1}^n c(y_i, \phi)$$

and

$$\hat{l}_p = \phi^{-1} \sum_{i=1}^n [y_i \hat{\theta}_i - b(\hat{\theta}_i)] + \sum_{i=1}^n c(y_i, \phi),$$

where $\tilde{\theta}$ and $\hat{\theta}$ are the maximum likelihood estimates of the canonical parameter for the saturated and current models, respectively, hence the scaled deviance we may also write as

$$S_p = 2\phi^{-1} \sum_{i=1}^n [y_i(\tilde{\theta}_i - \hat{\theta}_i) + b(\hat{\theta}_i) - b(\tilde{\theta}_i)]. \quad (5.3)$$

Another important overall discrepancy measure of a fitted model is the generalized Pearson statistic X_p^2 , given by

$$X_p^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}. \quad (5.4)$$

For the true model and a known dispersion parameter ϕ both the scaled deviance and the scaled generalized Pearson statistic follow, asymptotically, a χ^2 distribution with $n - p$ degrees of freedom, although Jørgensen (2002) recommends the use of the scaled X_p^2 statistic as convergence to the reference distributions is more rapidly achieved as the sample size n increases. For further details on GLM theory, see McCullagh and Nelder (1989). When ϕ is unknown, there is no formal goodness-of-fit test available based on S_p or scaled X_p^2 and in this case X_p^2 is often used to estimate ϕ (Jørgensen, 2002), making it useless as a measure of goodness-of-fit.

To analyse count and proportion data the Poisson and binomial models are, respectively, reasonable first choices. For these distributions the dispersion parameter is fixed and equal to 1. For a

well-fitted model it is expected that the residual deviance and the generalized Pearson statistic should be approximately equal to the residual degrees of freedom, the expected value of the χ^2 reference distribution. When this does not happen, the model may not fit the data well (for various reasons, such as a misspecified linear predictor or link function, outliers, etc), or simply, the variation in the data may be larger than is predicted by these models, a phenomenon often referred to as overdispersion, see (Hinde and Demétrio, 1998). There are different causes of overdispersion and failure to take it into account may result in misleading inferences. For a more thorough discussion, see Demétrio, Hinde and Moral (2014).

These basic models can be extended to incorporate overdispersion in several ways. A relatively simple one involves the specification of a more general variance function. For example, to analyse count data, suppose that $Y_i \sim P(\mu_i)$, then the Poisson model assumes that $E(Y_i) = \mu_i$ and that $\text{Var}(Y_i) = \mu_i$. Then, a simple extension is to take $\text{Var}(Y_i) = \phi\mu_i$ and use a quasi-likelihood approach to estimate $\phi > 1$, as the variance is greater than expected under the Poisson model. This specification of the variance function is called constant overdispersion. This may be generalized further by taking

$$\text{Var}(Y_i) = \mu_i(1 + \phi\mu_i^\delta). \quad (5.5)$$

For $\phi = 0$, this is simply the variance function of the standard Poisson model; for $\delta = 0$ it gives the constant overdispersion case; while if $\delta = 1$ we get the same type of quadratic variance function as for the negative binomial type-II model.

A similar approach may also be adopted for proportion data. Suppose that $Y_i \sim B(m_i, \pi_i)$, then the standard binomial model assumes that $E(Y_i) = m_i\pi_i$ and $\text{Var}(Y_i) = m_i\pi_i(1 - \pi_i)$. If we now take an extended variance function

$$\text{Var}(Y_i) = m_i\pi_i(1 - \pi_i)\{1 + \phi(m_i - 1)^{\delta_1}[\pi_i(1 - \pi_i)]^{\delta_2}\}, \quad (5.6)$$

for $\phi = 0$, the expression is the same as for the standard binomial model; for $\delta_1 = \delta_2 = 0$, we obtain the constant overdispersion variance function; while if $\delta_1 = 1$ and $\delta_2 = 0$, this is the same form of variance function as the beta-binomial model.

Another way to model overdispersion consists in adding an observation-level random effect in the linear predictor, i.e.,

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{z},$$

where $\mathbf{z} = (z_1, z_1, \dots, z_n)^\top$ is the vector of random effects, and typically we assume that $Z_i \sim N(0, 1)$. For count data, this leads to the Poisson-normal model. For proportion data, when the logit link is used, this results in the binomial-logit-normal model. This approach is closely related to the ideas in two-stage models where the basic parameter of interest is assumed to be an individual level random variable. This leads to negative binomial models for count data (by assuming the Poisson mean to have a gamma distribution) and the beta-binomial model for counted proportions (by assuming the binomial probability to have a beta distribution). Further details on overdispersion models for count and proportion data may be found in Hinde and Demétrio (1998) and Demétrio et al. (2014).

5.3 Half-normal plots with simulated envelopes

This relatively easy technique consists in plotting the ordered absolute values of a model diagnostic versus the expected order statistics of a half-normal distribution, which can be approximated as

$$\Phi^{-1}\left(\frac{i + n - \frac{1}{8}}{2n + \frac{1}{2}}\right), \quad (5.7)$$

where i corresponds to the i -th order statistic, $1 \leq i \leq n$ and n is the sample size, as in McCullagh and Nelder (1989), p. 407, following the results from Blom (1958) and Royston (1982). For a normal plot we use the ordered values versus the expected order statistics of a normal distribution, approximated as

$$\Phi^{-1}\left(\frac{i - \frac{3}{8}}{n + \frac{1}{4}}\right). \quad (5.8)$$

These order statistics are easily obtained using the R function `qnorm`, e.g., for $n = 7$ the expected order statistics of the half-normal distribution are

```
R> i <- 1:7
R> n <- 7
R> qnorm((i + n - 1/8)/(2 * n + 1/2))

[1] 0.1082554 0.2847156 0.4705935 0.6744898 0.9114298 1.2155984 1.7157550
```

Obtaining the simulated envelope for a half-normal plot is simple and consists of

1. fitting a model;
2. extracting model diagnostics and calculating sorted absolute values;
3. simulating 99 (or more) response variables using the same model matrix, error distribution and parameter estimates;
4. fitting the same model to each simulated response variable and extracting the same model diagnostics, and again sorting the absolute values;
5. computing the desired percentiles (e.g., 2.5 and 97.5) of the simulated diagnostic values at each value of the expected order statistic and using these to form the envelope.

5.3.1 Implemented model classes

The `hnp` function handles many different model classes and more will be implemented as time goes by. So far, a range of generalized linear models and extensions are included:

- models for continuous data — Gaussian (`lm`, `aov` and `glm` functions), gamma (`glm` function), and inverse Gaussian (`glm` function) models;
- models for count data — Poisson and quasi-Poisson (`glm` function), negative binomial type-II (`glm.nb` function in `MASS` (Venables and Ripley, 2002) or `aodml` function in `aods3` (Lesnoff and Lancelot, 2013)), and hurdle Poisson and negative binomial (`hurdle` function in `zeroinfl` (Zeileis et al., 2008)) models;
- models for proportion data — binomial and quasi-binomial (`glm` function), beta-binomial (`vglm` function in `VGAM` (Yee, 2010), `aodml` function in `aods3`, `gamlss` function in `gamlss` (Rigby and Stasinopoulos, 2005), `glmmadmb` function in `glmmADMB` (Skaug et al., 2014), and multinomial (`multinom` function in `nnet` (Venables and Ripley, 2002)) models;
- models for zero-inflated data — zero-inflated Poisson and negative binomial (`zeroinfl` function in `pscl`), zero-inflated binomial (`vglm` function in `VGAM`, `gamlss` function in `gamlss`, `glmmadmb` function in `glmmADMB`), and zero-inflated beta-binomial (`gamlss` function in `gamlss`, `glmmadmb` function in `glmmADMB`) models;
- mixed models — Gaussian (`lmer` function in `lme4` (Bates et al., 2015)), Poisson-normal and binomial-normal (`glmer` function in `lme4`) models.

5.3.2 Simulation procedures

For most of the implemented model classes, the simulation procedures are already implemented in the R base packages or in the packages being used to fit the models, such as the `rbinom`, `rpois`, `rmultinom`, `rgamma` and `rnorm` functions for binomial, Poisson, multinomial, gamma and normal models. Package `mgcv`'s (Wood, 2006) function `rig` was used for inverse gaussian models; package `gamlss`'s (Rigby and Stasinopoulos, 2005) functions `rBB`, `rZIBI` and `rZIBB` were used for beta-binomial, zero-inflated binomial and zero-inflated beta-binomial models fitted using `gamlss`; package `VGAM`'s (Yee, 2010) functions `rbetabinom` and `rzibinom` were used for beta-binomial and zero-inflated binomial models fitted using `vglm`; package `lme4`'s (Bates et al., 2015) function `simulate` was used for generalized linear mixed models fitted using `lmer` and `glmer`; package `MASS`'s (Venables and Ripley, 2002) function `rneqbin` for negative binomial models.

For the quasi-binomial model the random samples were simulated using `rbinom` and then multiplied by ϕ (`summary(model)$dispersion`) and the residuals were then scaled by $1/\sqrt{\phi}$. For the quasi-Poisson model, the function `rnbinom` was adapted with the `size` argument set to $\mu/(\phi - 1)$. New functions were written for zero-inflated binomial and zero-inflated beta-binomial models fitted with `glm-mADMB` (Skaug et al., 2014) and for zero-inflated and hurdle Poisson and negative binomial models fitted with `pscl` (Zeileis et al., 2008). The simulation procedures are all accessible in the `hnp` code.

5.3.3 New class implementation

To produce the half-normal plot with a simulated envelope, three procedures are required: (i) one to extract a diagnostic measure from the fitted model; (ii) one to simulate response variables using information from the model (error distribution, model matrix and fitted parameters); and finally (iii) one to refit the same model to the simulated data. The `hnp` function firstly recognizes the model class of the fitted object. If this class is not yet implemented, it returns an error. However, users may opt to supply their own diagnostic extraction, simulation and model fitting functions so that the half-normal plot is produced, see Section 5.4 for a practical guide.

5.4 Examples

Package `hnp` provides a range of examples drawn from Demétrio, Hinde and Moral (2014) and two of them will be discussed below for overdispersed proportion and count data. An example on orange tree embryogenesis (Tomaz et al., 2001) will be used to discuss new class implementation and an example on leukemia recurrence times (Miller, 1997) will be used to discuss the survival analysis models implementation. All data sets are presented in Supplement S3.

5.4.1 Overdispersed proportion data

A major pest of stored maize in Brazil is *Sitophilus zeamais*. In an experiment to assess the insecticide action of organic extracts of *Annona mucosa* (Annonaceae), Petri dishes containing 10g of corn were treated with extracts prepared with different parts of the plant (seeds, leaves and branches) at a concentration of 1500mg/kg or just water (control), using a completely randomized design with 10 replicates. Then 20 *S. zeamais* adults were placed in each Petri dish and, after 60 days, the numbers of damaged and undamaged corn grains were counted, see Ribeiro et al. (2013).

We begin by fitting a standard binomial model, i.e., $Y_{ij} \sim B(m_{ij}, \pi_{ij})$, to the data using the logit link with the following linear predictor:

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_0 + e_i, \quad (5.9)$$

where β_0 is the intercept and e_i is the effect of the i -th extract, $i = 1, \dots, 4$. We can fit this model in R and produce a simple analysis of deviance table using `glm()` and `anova()`:

```
R> require("hnp")
R> data("corn")
R> fit1_b <- glm(cbind(y, m - y) ~ extract, family = binomial, data = corn)
R> anova(fit1_b, test = "Chisq")
```

Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(y, m - y)

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			39	801.96	
extract	3	636.04	36	165.92	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The residual deviance is much larger than the number of residual degrees of freedom, indicating that the model does not fit the data well. We can easily check this by producing the half-normal plot with simulation envelope, see Figure 5.1(a) and it is clear that this is not a good model fit.

```
R> hnp(fit1_b, xlab = "Half-normal scores", ylab = "Deviance residuals",
+     main = "(a) Binomial model", pch = 4)
```

By default the deviance residuals have been used. Users may change colors, sizes or point character symbols freely using the same arguments passed to `plot()` and `par()`. We now turn into fitting overdispersion models, and begin by considering a quasi-likelihood approach:

```
R> fit2_b <- glm(cbind(y, m - y) ~ extract, family = quasibinomial, data = corn)
R> summary(fit2_b)$dispersion

[1] 4.409755
```

```
R> hnp(fit2_b, xlab = "Half-normal scores", ylab = "Deviance residuals",
+     main = "(b) Quasi-binomial model", pch = 4)
```

We observe that the estimated dispersion parameter $\hat{\phi} = 4.41$ and now the half-normal plot indicates that this model fit was satisfactory with most of the deviance residuals lying within the simulated envelope, see Figure 5.1(b). We may also easily fit other overdispersion models, such as the beta-binomial model using the same linear predictor (5.9) or the binomial-logit-normal model in which a random effect is included in the linear predictor:

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_0 + e_i + \sigma Z_{ij}, \quad (5.10)$$

where $Z_{ij} \sim N(0, \sigma^2)$. To fit these models, we make use of the packages `aods3` (Lesnoff and Lancelot, 2013) and `lme4` (Bates et al., 2015):

Figure 5.1. Half-normal plots with simulated envelopes for the deviance residuals from (a) the binomial model, (b) the quasi-binomial model, (c) the beta-binomial model, and (d) the binomial-logit-normal model, fitted to the corn data.

```
R> require("aods3")
R> fit3_b <- aodml(cbind(y, m - y) ~ extract, family = "bb", data = corn)
R> hnp(fit3_b, xlab = "Half-normal scores", ylab = "Deviance residuals",
+     main = "(c) Beta-binomial model", pch = 4)
R> require("lme4")
R> x <- factor(1:nrow(corn))
R> fit4_b <- glmer(cbind(y, m - y) ~ extract + (1 | x),
+     family = binomial, data = corn)
R> hnp(fit4_b, xlab = "Half-normal scores", ylab = "Deviance residuals",
+     main = "(d) Binomial-logit-normal model", pch = 4)
```

Both resulting half-normal plots with simulation envelopes show that these are also satisfactory models for this dataset, see Figures 5.1(c) and 5.1(d).

The three overdispersion models seem to fit the data equally well, which is not a surprise because the binomial sample sizes do not vary much (from 32 to 39) and when they are equal the beta-binomial and binomial-logit-normal variance functions both reduce to the constant overdispersion form, which may be a reasonable choice for this case, see Ribeiro et al. (2013).

5.4.2 Overdispersed count data

For the same experiment described in the previous section we now turn to focus on the number of emerged insects (progeny) after 60 days, see Ribeiro et al. (2013). We begin by fitting a standard

Poisson model, i.e., $Y_{ij} \sim P(\mu_{ij})$, using a log link with the following linear predictor:

$$\log(\mu_{ij}) = \beta_0 + e_i, \quad (5.11)$$

where β_0 is the intercept and e_i is the effect of the i -th extract, $i = 1, \dots, 4$. We can fit this model in R and produce a simple analysis of deviance as before, using `glm()` and `anova()`:

```
R> data("progeny")
R> fit1_p <- glm(y ~ extract, family = poisson, data = progeny)
R> anova(fit1_p, test = "Chisq")
```

Analysis of Deviance Table

Model: poisson, link: log

Response: y

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			39	534.44	
extract	3	444.68	36	89.77	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The residual deviance is again much larger than the number of residual degrees of freedom, indicating that the model does not fit the data well. This is confirmed by the half-normal plot with simulated envelope, see Figure 5.2(a):

```
R> hnp(fit1_p, xlab = "Half-normal scores", ylab = "Deviance residuals",
+     main = "(a) Poisson model", pch = 4)
```

We now fit overdispersion models, and begin with a constant overdispersion quasi-Poisson model:

```
R> fit2_p <- glm(y ~ extract, family = quasipoisson, data = progeny)
R> summary(fit2_p)$dispersion
```

```
[1] 2.365385
```

```
R> hnp(fit2_p, xlab = "Half-normal scores", ylab = "Deviance residuals",
+     main = "(b) Quasi-Poisson model", pch = 4)
```

We observe that the estimated dispersion parameter $\hat{\phi} = 2.37$ and now the half-normal plot indicates that this model fit was satisfactory with most of the deviance residuals lying within the simulated envelope, see Figure 5.2(b). We now make use of package `MASS` to fit the negative binomial type-II model using the same linear predictor (5.11) and `lme4` to fit the Poisson-normal model with a random effect in the linear predictor:

$$\log(\mu_{ij}) = \beta_0 + e_i + \sigma Z_{ij}, \quad (5.12)$$

where $Z_{ij} \sim N(0, \sigma^2)$.

Figure 5.2. Half-normal plots with simulated envelopes for the deviance residuals from (a) the Poisson model, (b) the quasi-Poisson model, (c) the negative binomial type-II model, and (d) the Poisson-normal model, fitted to the progeny data.

```
R> require("MASS")
R> fit3_p <- glm.nb(y ~ extract, data = progeny)
R> hnp(fit3_p, xlab = "Half-normal scores", ylab = "Deviance residuals",
+   main = "(c) Negative binomial model", pch = 4)
R> require("lme4")
R> x <- factor(1:nrow(progeny))
R> fit4_p <- glmer(y ~ extract + (1 | x), family = poisson, data = progeny)
R> hnp(fit4_p, xlab = "Half-normal scores", ylab = "Deviance residuals",
+   main = "(d) Poisson-normal model", pch = 4)
```

Both half-normal plots with simulation envelopes show that these are also satisfactory models for this dataset, see Figures 5.2(c) and 5.2(d), although there is perhaps some suggestion that the Poisson-normal model seems to capture the bulk of the variability better than the others.

5.4.3 Implementing new model classes

Negative binomial type-I model using package `gamlss`

We now turn to a dataset on a tissue-culture experiment using the orange variety *Caipira*. To study the effect of six sugars (maltose, glucose, galactose, lactose, sucrose and glycerol) on the stimulation of somatic embryos from callus cultures, the number of embryos after approximately four weeks was observed. The experiment was set up in a completely randomized block design with five blocks and the six sugars at dose levels of 18, 37, 75, 110 and 150 μM for the first five and 6, 12, 24, 36 and 50 μM for

glycerol, see Tomaz et al. (2001). The main interest was in the dose-response relationship and the data shows high variability. In their analysis, Tomaz et al. (2001) used a quasi-Poisson model. An alternative is the negative binomial type-I model (Jansakul and Hinde, 2004), which has the same variance function as the quasi-Poisson, that is, it is also a constant overdispersion model (see (5.5), with $\delta = 0$). However, as the negative binomial type-I is a fully specified probability model (albeit not in the exponential family) it is possible to obtain standard maximum likelihood parameter estimates.

For sugars lactose and galactose there seems to be a quadratic relationship when we look at the scatter plot, see Figure (5.4), which justifies the use of the following linear predictor:

$$\log(\mu_{ijk}) = \beta_0 + b_j + \beta_{1k} d_i + \beta_{2k} d_i^2, \quad (5.13)$$

where β_0 is the intercept, b_j is the effect of the j -th block, $j = 1, \dots, 5$, d_i is the i -th dose, $i = 1, \dots, 5$, and β_{1k} and β_{2k} are the linear and quadratic dose effects for the k -th sugar, $k = 1, \dots, 6$. Package `gamlss` allows for simple implementation of the negative binomial type-I model using `family = NBII()`¹:

```
R> data("orange")
R> require("gamlss")
R> fit_nbI <- gamlss(embryos ~ block + poly(dose, 2) * sugar,
+   family = NBII(), data = orange)
```

```
GAMLSS-RS iteration 1: Global Deviance = 1381.312
GAMLSS-RS iteration 2: Global Deviance = 1289.437
GAMLSS-RS iteration 3: Global Deviance = 1271.898
GAMLSS-RS iteration 4: Global Deviance = 1267.792
GAMLSS-RS iteration 5: Global Deviance = 1267.293
GAMLSS-RS iteration 6: Global Deviance = 1267.23
GAMLSS-RS iteration 7: Global Deviance = 1267.222
GAMLSS-RS iteration 8: Global Deviance = 1267.222
```

The `hnp` function does not handle `gamlss` objects with negative binomial type-I family. Attempting to use it yields an error. In order to use it we must pass three helper functions to `hnp`. The first function, passed to argument `diagfun`, must extract the desired model diagnostics. In this case, we will use the z-scores (see Rigby and Stasinopoulos, 2005, for further details), which are the default standardised residuals computed from the function `resid` on a `gamlss` object:

```
R> d.fun <- function(obj) resid(obj)
```

The second function, passed to argument `simfun`, is used to simulate random samples using the same model matrix and distribution used as when fitting the model to the original data:

```
R> s.fun <- function(n, obj) {
+   mu <- obj$mu.fv
+   sigma <- obj$sigma.fv
+   rNBII(n, mu, sigma)
+ }
```

A final function, passed to argument `fitfun`, is used to refit the simulated data using the same model as fitted to the original data:

¹Note that in package `gamlss` (as of version 4.3-8) families `NBI()` and `NBII()` correspond to the negative binomial type-II and type-I models, respectively, i.e., the types are switched.

```
R> f.fun <- function(y.) {
+   gamlss(y. ~ block + poly(dose, 2) * sugar, family = NBII(), data = orange)
+ }
```

By setting `newclass = TRUE` we can now pass the three helper functions and the dataset to the `hnp` function to produce the half-normal plot with simulated envelope:

```
R> hnp(fit_nbI, newclass = TRUE,
+   diagfun = d.fun, simfun = s.fun, fitfun = f.fun,
+   xlab = "Half-normal scores", ylab = "z-scores", main = "",
+   pch = 4, cex = 1, cex.lab = .8, cex.axis = .8)
```

As expected, this model also fits the data well, see Figure 5.3.

Figure 5.3. Half-normal plot with simulated envelope for the z-scores from the negative binomial type-I model fitted to the orange data using `gamlss`.

The fitted mean curves are easily obtained using the `predict` function and are displayed with the raw data in Figure 5.4:

```
R> fit_pred <- gamlss(embryos ~ poly(dose, 2) * sugar,
+   family = NBII(), data = orange)
R> orange.pred <- rbind(expand.grid(sugar = levels(orange$sugar)[-6],
+   dose = 18:150),
+   expand.grid(sugar = "Glycerol", dose=6:50))
R> orange.pred$pred <- predict(fit_pred, newdata = orange.pred,
+   type = "response")
R> require("latticeExtra")
R> trellis.par.set(strip.background = list(col = "lightgrey"))
R> xyplot(embryos ~ dose | sugar, scales = list(relation = "free"),
+   layout = c(3, 2), data = orange, col = 1, xlab = "Dose levels",
+   ylab = "Number of embryos") +
```

```
+ as.layer(xyplot(pred ~ dose | sugar, type = "l",
+ col = 1, data = orange.pred))
```

Figure 5.4. Number of embryos produced by oranges of the *Caipira* variety treated with six sugars at five different concentrations and fitted dose-response curves from the negative binomial type-I model.

Exponential and Weibull models using package `survival`

In a clinical trial to assess the efficacy of maintenance chemotherapy for acute myelogenous leukemia (AML), after reaching a state of remission through treatment with chemotherapy, 23 patients were randomized into two groups. The first group continued receiving maintenance chemotherapy and the second did not. The observed outcome was the time (in weeks) until relapse and it was right-censored (Miller, 1997). The AML dataset is available from the object `leukemia` from the package `survival` (Therneau, 2014):

```
R> require("survival")
R> data("leukemia")
```

Two widely used models in survival analysis are the exponential and the Weibull models. They are both easily fitted to data using the function `survreg` from the package `survival`. We proceed by fitting the exponential and Weibull models to the AML data including the treatment factor in the linear predictor:

```
R> fit_exp <- survreg(Surv(time, status) ~ x - 1,
+ dist = "exponential", data = leukemia)
R> fit_weib <- survreg(Surv(time, status) ~ x - 1,
+ dist = "weibull", data = leukemia)
```

We now must provide the three helper functions for `hnp` in the same way as in the previous section, albeit with slight differences. The `resid` function for `survreg` objects can be used to compute the deviance residuals for survival analysis models. The three helper functions may be written as:

```
R> d.fun <- function(obj) resid(obj, type = "deviance")
R> s.fun.exp <- function(n, obj) {
```

```

+   mu <- predict(obj)
+   rexp(n, rate = 1/mu)
+ }
R> s.fun.weib <- function(n, obj) {
+   sc <- obj$scale
+   mu <- predict(obj)
+   rweibull(n, shape = 1/sc, scale = mu)
+ }
R> f.fun.exp <- function(y.) {
+   leukemia$censoring <- c(sample(leukemia$status[1:11]),
+                             sample(leukemia$status[12:23]))
+   return(survreg(Surv(y., censoring) ~ x - 1,
+                   dist = "exponential", data = leukemia))
+ }
R> f.fun.weib <- function(y.) {
+   leukemia$censoring <- c(sample(leukemia$status[1:11]),
+                             sample(leukemia$status[12:23]))
+   return(survreg(Surv(y., censoring) ~ x - 1,
+                   dist = "weibull", data = leukemia))
+ }
R> hnp(fit_exp, newclass = TRUE,
+     diagfun = d.fun, simfun = s.fun.exp, fitfun = f.fun.exp,
+     xlab = "Half-normal scores", ylab = "Deviance residuals",
+     main = "(a) Exponential model", pch = 4, cex = 1)
R> hnp(fit_weib, newclass = TRUE,
+     diagfun = d.fun, simfun = s.fun.weib, fitfun = f.fun.weib,
+     xlab = "Half-normal scores", ylab = "Deviance residuals",
+     main = "(b) Weibull model", pch = 4, cex = 1)

```

We observe that both models seem to fit the data reasonably well, see Figure 5.5. This is as we might expect as the Weibull scale parameter is not significantly different from 1, which corresponds to the exponential distribution.

Figure 5.5. Half-normal plot with simulated envelope for the deviance residuals from (a) the exponential model and (b) the Weibull model fitted to the AML data.

Weibull model using function `optim` to maximise the likelihood

Now suppose that we want to fit a model that has not yet been implemented in any R function. It is also possible to produce the half-normal plots with simulation envelopes if a fitting function is written for this new type of model. For sake of simplicity, here we show how this can be done by programming an estimation procedure for the Weibull model and using the AML dataset. For parametric survival analysis models, the log-likelihood function can be written as

$$l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^n \delta_i \log\{f(y_i; \boldsymbol{\theta})\} + (1 - \delta_i) \log\{S(y_i; \boldsymbol{\theta})\},$$

where $\boldsymbol{\theta}$ is the vector of parameters, y_i is the time for the i -th individual, $i = 1, \dots, n$, $f(\cdot)$ is the probability density function, $S(\cdot) = 1 - F(\cdot)$ is the survival function and δ_i is the censoring indicator (for observation i , $\delta_i = 1$ if not censored and $\delta_i = 0$ if censored). The log-likelihood function under non-informative censoring for the Weibull model may be written in R as below, with `param` the vector of parameters, `y` the data vector, `cens` the censoring indicator vector and `X` a generic model matrix:

```
R> loglik.weibull <- function(params, y, cens, X) {
+   shape <- exp(params[1])
+   mu <- exp(X %*% params[-1])
+   llik <- cens * dweibull(y, shape, mu, log = TRUE) +
+         (1 - cens) * pweibull(y, shape, mu, lower.tail = FALSE, log = TRUE)
+   return(-sum(llik))
+ }
```

Parameter estimates are easily obtained through `optim` and are comparable to the estimates obtained through `survreg`:

```
R> weibull.fit <- optim(log(c(1, 60, 24)), loglik.weibull,
+   y = leukemia$time, cens = leukemia$status,
+   X = model.matrix(~ x - 1, leukemia))
R> round(exp(weibull.fit$par), 2)

[1] 1.26 60.90 24.05

R> round(as.numeric(with(fit_weib, c(1/scale, exp(coefficients)))), 2)

[1] 1.26 60.89 24.04
```

To use the `hnp` function to obtain the half-normal plot with simulated envelope, we must first write a function for fitting the model. To make the later writing of `hnp`'s helper functions easier, we can include the calculation of a vector of residuals in the fitting function. In this case we will use the modified Cox-Snell residuals, defined as

$$\hat{r}_{cs_i} = -\log \hat{S}(y_i) + 1 - \delta_i,$$

where $\hat{S}(\cdot)$ is the estimated survival function and y_i is the observation or censoring time. The following code may be used to implement these functions and fit the model:

```
R> cox.snell <- function(params, y, X, cens) {
+   shape <- params[1]
+   mu <- X %*% params[-1]
+   cs <- -pweibull(y, shape, mu, lower.tail = FALSE, log = TRUE)
```

```

+   return(cs + 1 - cens)
+ }
R> fit.model <- function(fmla, cens, init, data) {
+   resp <- with(data, eval(fmla[[2]]))
+   X <- model.matrix(fmla, data)
+   fit <- optim(init, loglik.weibull, y = resp, cens = cens, X = X)
+   result <- list(pars = exp(fit$par),
+                 resid = cox.snell(params = exp(fit$par), y = resp,
+                 X = X, cens = cens))
+   return(result)
+ }
R> fit_weib.lik <- fit.model(fmla = time ~ x - 1, cens = leukemia$status,
+   init = log(c(1, 60, 24)), data = leukemia)
R> fit_weib.lik

```

\$pars

```
[1] 1.264268 60.899983 24.046215
```

\$resid

```

[1] 0.08916287 0.14193473 1.14193473 0.21417391 0.29198102 1.37442192
[7] 0.42584008 0.47859201 1.68213340 0.74012510 4.41810086 0.13730057
[13] 0.13730057 0.24873372 0.24873372 0.41529928 1.59747151 0.94531324
[19] 1.15774844 1.32270784 1.49209119 2.08510851 2.20846458

```

Now we may write the helper functions for `hnp` as

```

R> d.fun <- function(obj) obj$resid
R> s.fun <- function(n, obj) {
+   params <- obj$pars
+   rweibull(n, shape = params[1], scale = rep(params[2:3], each = 15))
+ }
R> f.fun <- function(y.) {
+   censoring <- c(sample(leukemia$status[1:11]),
+                 sample(leukemia$status[12:23]))
+   leukemia$new.response <- y.
+   return(fit.model(new.response ~ x - 1, cens = censoring,
+                 init = log(fit_weib.lik$pars), data = leukemia))
+ }

```

And produce the half-normal plot:

```

R> hnp(fit_weib.lik, newclass = TRUE,
+   diagfun = d.fun, simfun = s.fun, fitfun = f.fun,
+   xlab = "Half-normal scores", ylab = "Modified Cox-Snell residuals",
+   main = "", pch = 4, cex = 1, cex.lab = .8, cex.axis = .8)

```

As expected, the model fits the data, see Figure 5.6. As we are using a different type of residuals, the shape of the envelope is different than for the Weibull model fitted using `survreg`. Indeed, for a well-fitting model the Cox-Snell residuals should follow an exponential distribution with unit mean, rather than the normal distribution for the deviance residuals.

Figure 5.6. Half-normal plot with simulated envelope for the modified Cox-Snell residuals from the Weibull model fitted to the AML data.

5.5 Discussion

Half-normal plots with simulation envelopes are a useful tool to assess goodness-of-fit for a range of different models, such as classical linear models, generalized linear models, models for overdispersed and zero-inflated data, survival analysis models, as well as mixed models. The key point lies in the simulation procedures which sometimes may be problematic due to model complexity. The `hnp` package allows for implementation of any model for which it is possible to write diagnostic extraction, simulation and fitting codes. This may also be useful for didactic purposes and in simulation studies.

There are other packages that provide a (half-)normal plot with simulated envelope for a few model classes. For example, package `ssym`'s function `envelope` produces a normal plot with simulated envelope for semi-parametric log-symmetric regression models (Vanegas and Paula, 2015). Package `pgam` also includes an `envelope` function to produce normal plots with simulated envelopes for Poisson-gamma additive models fitted using the roughness penalty approach (Junger and Leon, 2012). Package `betareg` includes a `plot` method (argument `which = 5`) that produces a half-normal plot with simulated envelope for residuals of a beta regression model (Cribari-Neto and Zeileis, 2010). Also, package `car` provides a function, `qqPlot`, whose method for `lm` objects generates a normal plot with simulated envelope for studentized residuals of a linear regression (FOX; WEISBERG, 2011).

In the `hnp` function, by default the half-normal distribution is used to plot the diagnostic quantities against its expected order statistics (the normal distribution may be used by setting `halfnormal = FALSE`). However there is no reason to do so other than several types of residuals are shown to follow the normal distribution and so we would expect them to form a straight line in the plot. Again the aim in producing a simulation envelope is to reduce subjective bias as to the departure of points from a straight line (when the diagnostic quantity distribution is expected to be normal) and to provide the expected shape of the plot when the distribution is not expected to be normal. Depending upon model complexity the simulation and especially the fitting procedures may be time consuming. Therefore, considerable time may be spent to produce a simulation envelope with the default 99 simulations when one fitting

procedure already takes a long time. So, sometimes it may be wiser to use just 19 simulations (`sim = 19`) and form the envelope from the minimum and maximum values obtained in the simulations (`conf = 1`), as proposed by Atkinson (1985), so that there is a chance of approximately 1 in 20 that the observed value of the diagnostic quantity is the most extreme and lies outside of the envelope.

Throughout this paper we have used different types of residuals in the examples, but for generalized linear models mainly deviance residuals, which are shown to have important asymptotic properties and be suitable for diagnostic analyses (McCullagh and Nelder, 1989). For mixed models, the question of which type of residuals should be used for goodness-of-fit assessment is still an active area of research (Nobre and Singer, 2007). However, any type of model diagnostic may be used because they are simply being compared with what we might expect if the model were true, as given by the repeated sequences of “data simulation — model re-fitting — diagnostic extraction”. It is important to point out that every time this plot is produced the envelope bands change slightly, hence it sometimes may be useful to produce several half-normal plots and observe how many points lie outside of the bands as well as their position. Of course, not only are we not ruling out the use of other goodness-of-fit assessment techniques, but also we encourage that different tools are used to ensure that the model fits the data well so that no misleading inference is made. The `hnp` package merely provides a, hopefully, useful and flexible tool to help in this assessment.

References

- ATKINSON, A. C. (1985) **Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis**. Oxford: Clarendon Press. 282 p.
- BARTLETT, M. S. (1937) Properties of sufficiency and statistical tests. **Proceedings of the Royal Society of London – Series A**, v. 160, p. 268–282.
- BATES, D; MARCHLER, M.; BOLKER, B. (2015) lme4: Linear mixed-effects models using Eigen and Eigen. R package version 1.1-10. URL <<http://CRAN.R-project.org/package=lme4>>
- BLOM, G. (1958) **Statistical estimates and transformed beta-variables**. Stockholm: Almqvist & Wiksell.
- CRIBARI-NETO, F.; ZEILEIS, A. (2010) Beta regression in R. **Journal of Statistical Software**, v. 34, p. 1–24.
- DEMÉTRIO, C. G. B.; HINDE, J. (1997) Half-normal plots and overdispersion. **GLIM Newsletter**, v. 27, p. 19–26.
- DEMÉTRIO, C. G. B.; HINDE, J.; MORAL, R. A. (2014) Models for overdispersed data in entomology. In: FERREIRA, C. P.; GODOY, W. A. C. (Eds.). **Ecological modeling applied to entomology**. Springer. 262 p.
- FOX, J.; WEISBERG, S. (2011) **An R companion to applied regression**. Thousand Oaks: Sage. 472 p.
- HINDE, J.; DEMÉTRIO, C. G. B. (1998) Overdispersion: models and estimation. **Computational Statistics & Data Analysis**, v. 27, p. 151–170.
- JANSAKUL, N.; HINDE, J. (2004) Linear mean-variance negative binomial models for analysis of orange tissue-culture data. **Songklanakarin Journal of Science and Technology**, 26, 683–696.
- JØRGENSEN, B. (2002) Generalized linear models. In: EL-SHAARAWI, A. H.; PIEGORSCH, W. W. (Eds.). **Encyclopedia of Environmetrics**. Chichester: John Wiley. p. 873–880.

- JUNGER, W.; LEON, A.P. (2012) *pgam*: Poisson-Gamma Additive Models. R package version 0.4.12, URL <<https://CRAN.R-project.org/package=pgam>>
- LESNOFF, M.; LANCELOT, R. (2013) *aods3*: Analysis of Overdispersed Data Using S3 Methods. R package version 0.4-1. URL <<https://CRAN.R-project.org/package=aods3>>
- MCCULLAGH, P.; NELDER, J. A. (1989) **Generalized linear models**, London: Chapman and Hall. 511 p.
- MILLER, R. G. (1997) **Survival analysis**. New York: John Wiley & Sons. 238 p.
- NELDER, J. A.; WEDDERBURN, R. W. M. (1972) Generalized linear models. **Journal of the Royal Statistical Society A**, v. 135, p. 370–384.
- NOBRE, J. S.; SINGER, J. M. (2007) Residual analysis for linear mixed models. **Biometrical Journal**, v. 49, p. 863–875.
- RIBEIRO, L. P.; VENDRAMIN, J. D.; BICALHO, K. U.; ANDRADE, M. S.; FERNANDES, J. B.; MORAL, R. A.; DEMÉTRIO, C. G. B. (2013) *Annona mucosa* Jacq. (Annonaceae): A promising source of bioactive compounds against *Sitophilus zeamais* Mots. (Coleoptera: Curculionidae). **Journal of Stored Products Research**, v. 55, p. 6–614.
- RIGBY, R. A.; STASINOPOULOS, D. M. (2005) Generalized additive models for location, scale and shape (with discussion). **Journal of the Royal Statistical Society – Series C (Applied Statistics)**, v. 54, p. 507–554.
- ROYSTON, J. P. (1982) Algorithm AS 177: Expected normal order statistics (exact and approximate). **Journal of the Royal Statistical Society – Series C (Applied Statistics)**, 31, 161–165.
- SHAPIRO, S. S.; WILK, M. B. (1965) An analysis of variance test for normality (complete samples). **Biometrika**, v. 52, p. 591–611.
- SKAUG, H.; FOURNIER, D.; BOLKER, B.; MAGNUSSON, A.; NIELSEN, A. (2014) Generalized linear mixed models using AD Model Builder. R package version 0.8.1. URL <<http://glmmadmb.r-forge.r-project.org/>>
- THERNEAU, T. (2014) A package for survival analysis in S. R package version 2.38-3. URL <<http://CRAN.R-project.org/package=survival>>
- TOMAZ, M. L.; MENDES, B. M. J.; FILHO, F. A. M.; DEMÉTRIO, C. G. B.; JANSAKUL, N.; RODRIGUEZ, A. P. M. (1997) Somatic embryogenesis in *Citrus* spp.: carbohydrate stimulation and histodifferentiation. **In Vitro Cellular & Developmental Biology – Plant**, v. 37, p. 446–452.
- VANEGAS, L. H.; PAULA, G. A. (2015) *ssym*: Fitting semi-parametric log-symmetric regression models. R package version 1.5.4. <<https://CRAN.R-project.org/package=ssym>>
- VENABLES, W. N.; RIPLEY, B. D. (2002) **Modern applied statistics with S**. Fourth Edition. New York: Springer. 504 p.
- WOOD, S. N. (2006) **Generalized additive models: An introduction with R**. Boca Raton: Chapman and Hall. 384 p.
- YEE, T. W. (2010) The VGAM package for categorical data analysis. **Journal of Statistical Software**, v. 32, p. 1–34.

ZEILEIS, A.; KLEIBER, C.; JACKMAN, S. (2008) Regression models for count data in R. **Journal of Statistical Software**, v. 27, p. 1–25.

6 ASSESSING GOODNESS-OF-FIT FOR BIVARIATE MODELS USING RESIDUAL PLOTS WITH SIMULATION POLYGONS

Abstract

When using univariate models, goodness-of-fit can be assessed through many different methods, including graphical tools such as half-normal plots with a simulation envelope. This is straightforward due to the notion of ordering of a univariate sample, which can readily reveal possible outliers. In the bivariate case, however, it is often difficult to detect extreme points and verify whether a sample of residuals is a reasonable realisation from a fitted model. We propose a new framework, implemented as the `bivrp` R package, available on the Comprehensive R Archive Network. Our framework uses the same principles of the simulation envelope in a half-normal plot, but as a simulation polygon for each point in a bivariate sample. By using algorithms of convex hull construction and polygon area reduction, we describe how our method works and illustrate its functionality with three examples: with simulated bivariate normal data, real bivariate Poisson data, and simulated bivariate binomial-Poisson mixture data. We show how different model diagnostics can produce different results and pinpoint potential drawbacks of our approach, such as the limitations in terms of computational burden and convex hull bias.

Keywords: `bivrp` package; Goodness-of-fit; Graphical methods; PIT diagnostics

6.1 Introduction

It is common for researchers, when carrying out experiments or observational studies, to record more than one response variable. Frequently, the interest lies not only on quantifying the effect of covariates on the responses but also on the correlation between response variables. In those cases, there is a wide range of multivariate statistical techniques and modelling frameworks that may be of interest. When fitting any sort of statistical model, it is important to assess goodness-of-fit, so that inference made using a specific model to describe the data and make predictions is reliable. As a particular case of multivariate data analysis, the use of bivariate models has increased since 2010, through extensions of univariate models (searching for the keyword “bivariate” as a topic in *Web of Science* yielded 1643, 1836, 2007, 2166, 2409, and 2717 yearly publications from 2010 to 2015).

For both univariate and multivariate models, it is possible to obtain summary goodness-of-fit statistics to compare model fits, such as information criteria (e.g. Akaike and Bayesian) and adjusted R-squared values. However, while these statistics are useful for comparing models fitted to the same data, they cannot be used to inform on the suitability of a particular model to the data. In this case, diagnostic analyses play a significant role (Atkinson, 1985). For univariate models, a possible alternative is the use of half-normal plots with simulation envelopes. Under a suitable model, we expect the majority of any chosen model diagnostics to fall within a simulated envelope, computed using the fitted model.

To produce a half-normal plot with a simulated envelope, we (i) obtain any model diagnostics in absolute value and in order; (ii) simulate 99 response variables from the fitted model, using the same model matrix, error distribution and estimated parameters; (iii) refit the same model to each simulated response variable and obtain the same model diagnostics, again ordered absolute values; and (iv) compute the 2.5% and 97.5% percentiles for each ordering index to form the envelope (Moral et al., *in press*). However, extending it to the bivariate case may become problematic as the notion of ordering for multivariate samples is not so straightforward (Barnett and Lewis, 1994).

For example, one might consider the use of the distance of points to a center point as a notion of ordering, but this can bring problems depending on the dispersion of each group of points (Healy, 1968).

Different solutions to the ordering problem have been proposed, especially for the bivariate case. These include the use of generalised distances and techniques of sub-ordering. Then, it is possible to detect outliers and then (empirically) test whether these points are plausible realisations of the fitted statistical model (see Barnett and Lewis, 1994, ch. 7).

Here, we propose a different approach, which relies on the same strategy used for producing half-normal plots with simulation envelopes. Our proposal makes it possible for one to graphically examine whether the observed data is a plausible realisation of the fitted model, without the need for first detecting outliers. We begin by explaining our approach and then illustrate with three examples. The first example uses simulated bivariate normal data, the second a real bivariate count dataset, and the third uses simulated data from a binomial-Poisson mixture. Finally, we discuss alternative approaches and potential drawbacks.

6.2 A bivariate residual plot for assessing goodness-of-fit

Let $\mathbf{y}_i = (y_{1i}, y_{2i})^\top$ be a vector of bivariate responses, to which a model is fitted. Also, let $\mathbf{r}_i = (r_{1i}, r_{2i})^\top$ be the vector of bivariate model diagnostics (e.g. residuals, Cook's distances, leverages). Then, a very basic and straightforward way of ordering the bivariate diagnostics is by calculating the angle

$$\alpha_i = \tan^{-1} \left(\frac{r_{2i}}{r_{1i}} \right)$$

they form with the origin (see Figure 6.1(a)), to obtain the ordered diagnostics $\mathbf{r}_{(i)} = (r_{1(i)}, r_{2(i)})^\top$. Here, the ordering is done not to define extremes, but to aid in the simulation process.

We then simulate 99 bivariate response variables $\mathbf{y}_i^s = (y_{1i}^s, y_{2i}^s)^\top$, $s = 1, \dots, 99$ from the fitted model, using the same model matrices, error distribution and fitted parameters, and refit the same model to each simulated sample, obtaining the same type of model diagnostics $\mathbf{r}_{(i)}^s = (r_{1(i)}^s, r_{2(i)}^s)^\top$, ordered by the angles they form with the origin. We have, for each bivariate diagnostic $\mathbf{r}_{(i)}$, 99 simulated bivariate diagnostics $\mathbf{r}_{(i)}^s$ (see Figures 6.1(b) and (c)), forming the whole cloud of simulated diagnostics (see Figure 6.1(d)). We then obtain the convex hulls of each set of the s sets of points and obtain a reduced polygon whose area is 95% of the original convex hull's area, forming the simulated polygon (see Figure 6.1(e)). The points are then connected to the centroids of their respective simulated polygons and, if they lie outside the polygons, they are drawn in red (see Figure 6.1(e)). For the final display, the polygons are erased so as to ease visualization (see Figure 6.1(f)). For details on polygon operations, refer to Appendix D.

Throughout our illustrative examples we use raw residuals (for normal and mixture models) and deviance residuals (for Poisson models). Moreover, we produce the plot for the probability integral transform (PIT) diagnostics for the normal models, which are written, for observation i of variable j , as

$$r_{ji}^{pit} = F(y_{ji}; \hat{\boldsymbol{\theta}}_{ji}),$$

where $F(\cdot)$ is the cumulative distribution function of the statistical model being used and $\hat{\boldsymbol{\theta}}_{ji}$ is the vector of estimated model parameters. For a well-fitted model, the PIT diagnostics are expected to follow a uniform $(0, 1)$ distribution (Dawid, 1984). For discrete models, however, the PIT diagnostics are no longer uniformly distributed under the hypothesis of a well-fitted model (Czado et al., 2009). Then, it is recommended that the randomized PIT diagnostics are used, written as

$$r_{ji}^{rand.pit} = F(y_{ji} - 1; \hat{\boldsymbol{\theta}}_{ji}) + u_i \{F(y_{ji}; \hat{\boldsymbol{\theta}}_{ji}) - F(y_{ji} - 1; \hat{\boldsymbol{\theta}}_{ji})\}, \quad (6.1)$$

with $F(-1) = 0$, and u_i is a realization of $U_i \sim \text{Uniform}(0, 1)$. There is also a non-randomized, yet uniform version of these diagnostics, proposed by Czado et al. (2009), but in this work we will use the randomized PIT diagnostics for discrete models.

Figure 6.1. The process of constructing the bivariate residual plot with simulation polygons; (a) sorting the points by the angles they make with the origin, (b), (c), (d) obtaining the points from refitting the model to 99 simulated responses and sorting the diagnostics using their angles, (e) obtaining and shrinking the convex hulls of the points to form the 95% simulation polygons, and (f) removing them from the plot for better visualization. The open dots in (e) and (f) are the centroids of the simulation polygons, linked to the solid dots, which are the observed bivariate diagnostics. The points outside their simulated polygons are displayed in red.

It is usual to compare the empirical cumulative density function of the PIT diagnostics with the identity function or to examine the histogram (Smith, 1985). Here, we look at the kernel density plots for the marginal diagnostics, as described by Czado et al. (2009).

This graphical method is implemented as the R package `bivrp` (Moral et al., 2016).

6.3 Applications

We now apply our goodness-of-fit assessment method to three different bivariate modelling frameworks. Firstly, we simulate from a bivariate normal model, then we use a bivariate Poisson model to analyse data from an experiment in entomology, and, finally, we simulate data from a binomial-Poisson mixture applied in a longitudinal context.

6.3.1 Bivariate normal responses

We begin illustrating our approach with a simple simulated example with a moderate sample size $n = 80$. Let Y_{1i} and Y_{2i} , $i = 1, \dots, 80$, be two normally distributed correlated random variables. We may write

$$\mathbf{Y}_i = \begin{bmatrix} Y_{1i} \\ Y_{2i} \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \mu_{1i} \\ \mu_{2i} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right),$$

which yields, marginally,

$$Y_{ji} \sim N(\mu_{ji}, \sigma_j^2), \quad j = 1, 2,$$

with $\text{Cov}(Y_{1i}, Y_{2i}) = \sigma_{12}$. Hence, the marginal distributions remain unaltered regardless of σ_{12} , which when equal to zero implies independence between the random variables.

We supposed a simple linear regression of the form

$$\mu_{ji} = \beta_{j0} + \beta_{j1}x_i$$

between both variables and a covariate x_i , which ranges from 1 to 10 in equal steps of approximately 0.114, giving 80 distinct x_i values. We fixed the true parameter values as $\beta_{10} = 2, \beta_{11} = 0.4, \beta_{20} = \beta_{21} = 0.2, \sigma_1^2 = 2, \sigma_2^2 = 3$, and a negative covariance $\sigma_{12} = -1.7$. The likelihood function may be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{i=1}^{80} (2\pi)^{-1} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \right\},$$

with $\boldsymbol{\beta}^\top = [\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}]$, $\mathbf{y}_i = \begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix}$, $\boldsymbol{\mu}_i = \begin{bmatrix} \mu_{1i} \\ \mu_{2i} \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$. The log-likelihood function to be maximised can then be written as

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \sum_{i=1}^{80} \log\{L(\boldsymbol{\beta}, \boldsymbol{\Sigma})\} \\ &\propto -80 \log |\boldsymbol{\Sigma}| - \sum_{i=1}^{80} (\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \\ &= -80 \log(\sigma_1^2 \sigma_2^2 - \sigma_{12}^2) - \sum_{i=1}^{80} \frac{\sigma_2^2 (y_{1i} - \mu_{1i})^2 - 2\sigma_{12} (y_{1i} - \mu_{1i})(y_{2i} - \mu_{2i}) + \sigma_1^2 (y_{2i} - \mu_{2i})^2}{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2}. \end{aligned}$$

We simulated the data in R (R Core Team, 2017) using the function `rmvnorm` from the package `mvtnorm` (Genz and Bretz, 2009) with the seed fixed as 2016 to allow for reproducibility. We then fitted the bivariate model using the BFGS algorithm implemented in the function `optim` twice, first estimating the covariance σ_{12} and then the second time fixing $\sigma_{12} = 0$, which is the same as fitting two separate linear regressions. The results are given in Table 6.1. The parameter estimates were virtually the same for both approaches, with the obvious exception of the covariance estimate. The likelihood ratio was 43.33 on 1 degree of freedom, which is highly significant, giving evidence that σ_{12} is different from zero, as expected.

Table 6.1. Parameter estimates (standard errors) for both models fitted to the simulated correlated bivariate normal data and true values

Parameter	Assuming independence	Estimating covariance	True value
β_{10}	1.50 (0.35)	1.50 (0.35)	2.00
β_{11}	0.43 (0.06)	0.43 (0.06)	0.40
β_{20}	0.78 (0.48)	0.78 (0.48)	0.20
β_{21}	0.18 (0.08)	0.18 (0.08)	0.20
σ_1^2	1.79 (0.28)	1.79 (0.28)	2.00
σ_2^2	3.48 (0.55)	3.49 (0.55)	3.00
σ_{12}	0.00 (—)	−1.61 (0.33)	−1.70
−2×loglik	600.42	557.10	—

We then proceeded to produce the bivariate residual plot with simulation polygons for both model fits. The raw residuals may be defined as

$$r_{ji}^o = y_{ji} - \hat{\mu}_{ji},$$

where $\hat{\mu}_{ji} = \hat{\beta}_{j0} + \hat{\beta}_{j1}x_i$, and the PIT diagnostics are

$$\begin{aligned} r_{ji}^{pit} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y_{ji}} \frac{1}{\sqrt{\sigma_j^2}} \exp \left\{ -\frac{1}{2\sigma_j^2} (y_{ji} - \hat{\mu}_{ji})^2 \right\} dy_{ji} \\ &= \frac{1}{2} \left\{ 1 + \operatorname{erf} \left(\frac{y_{ji} - \hat{\mu}_{ji}}{\sqrt{2\sigma_j^2}} \right) \right\}, \end{aligned}$$

where $\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-t^2} dt$ is the error function, for which many numerical approximations are available, such as the one implemented in the R function `pnorm`, which is a C translation of the FORTRAN algorithm devised by Cody (1993).

We observe that fitting the models independently for each variable yields a poor model fit, with many points lying outside of the simulation polygons when compared with the joint model (see Figure 6.2). Looking at the plots for the ordinary residuals, we observe that because of the negative correlation between the variables the expected two-dimensional pattern of the scatterplot of residuals is actually an ellipse (Figure 6.2(b)) instead of a circle (Figure 6.2(a)). Hence, even though when inspecting each variable marginally it might seem that the behaviour of the residuals is as expected by the 99 simulations, this is not true when looking at the bivariate nature of these residuals. Here, our proposed approach worked well in identifying these bivariate patterns and was appropriate to aid in model selection and in identifying possible outliers.

6.3.2 Bivariate Poisson responses – attack data

Now we turn to an example with count data, obtained from an experiment in entomology. The objective was to study the aggressive behaviour of two species of predators, *Podisus nigrispinus*, a stinkbug, and *Euborellia annulipes*, an earwig, both natural enemies of *Alabama argillacea*, an important pest of cotton. Moral (2012, *unpublished data*) placed one female of each predator in Petri dishes together with larvae of the pest. Then the number of times one predator attacked the other in one hour of continuous observation was recorded. The insect that attacked its competitor first was also recorded. The joint modelling of these two response variables is a natural way of analysing these data, since when one predator is attacked, it may counterattack to defend itself and hence, the number of attacks of one predator is correlated to the number of attacks of the other. Here, we use a subset of this dataset where all observations were taken from when the earwig was the specimen that made the first attack, giving $n = 56$ bivariate observations.

To define the bivariate Poisson model we begin by considering independent random variables X_0 , X_1 , and X_2 , that are all Poisson distributed with parameters λ_0 , λ_1 , and λ_2 , respectively. Under these settings, the random variables

$$Y_1 = X_0 + X_1 \tag{6.2}$$

$$Y_2 = X_0 + X_2 \tag{6.3}$$

jointly follow a bivariate Poisson distribution (Johnson et al., 1997), i.e.

$$(Y_1, Y_2) \sim \text{BP}(\lambda_0, \lambda_1, \lambda_2)$$

with probability function

$$P(Y_1 = y_1, Y_2 = y_2) = e^{-(\lambda_0 + \lambda_1 + \lambda_2)} \frac{\lambda_1^{y_1} \lambda_2^{y_2}}{y_1! y_2!} \sum_{k=0}^{\min(y_1, y_2)} \binom{y_1}{k} \binom{y_2}{k} k! \left(\frac{\lambda_0}{\lambda_1 \lambda_2} \right)^k. \tag{6.4}$$

Figure 6.2. Diagnostic plots for the bivariate normal simulated data using ordinary residuals (a and b) and PIT diagnostics (c and d), while assuming there was no covariance between the two random variables (a and c) or estimating this covariance with a joint model (b and d). Filled red points are outside of their respective simulated polygon; there are (a) 45, (b) 6, (c) 53, and (d) 23 points outside of the simulated polygons out of 80 total points.

Marginally, we have that

$$\begin{aligned} Y_1 &\sim P(\lambda_0 + \lambda_1) \\ Y_2 &\sim P(\lambda_0 + \lambda_2) \end{aligned}$$

and also $\text{Cov}(Y_1, Y_2) = \lambda_0$, therefore λ_0 may be interpreted as a measure of dependence between the random variables. Clearly, if $\lambda_0 = 0$, (6.4) reduces to the product between two Poisson distributions and then Y_1 and Y_2 are independent.

Estimation for this model is problematic and may present identifiability issues when the sample size is small or when there are too many parameters to estimate. Several methods have been proposed, such as pseudo-likelihood maximization (Gourieroux et al., 1984), a Newton-Raphson algorithm (Jung

and Winkelmann, 1993; Kocherlakota and Kocherlakota, 2001), a generalized least squares method (Ho and Singer, 2001), and Bayesian methods (Tsonas, 2001). Here, we use the EM algorithm as devised by Karlis and Ntzoufras (2005). For our experiment, let y_{ji} be the observed number of attacks of species j , $j = 1, 2$, taken on replicate i , $i = 1, \dots, 56$. We begin by writing the complete data log-likelihood for the unobserved variables X_{0i} , X_{1i} , and X_{2i} , as

$$\begin{aligned}
 l(\lambda_0, \lambda_1, \lambda_2) &= \sum_{i=1}^{56} \log\{P(X_{0i} = x_{0i}, X_{1i} = x_{1i}, X_{2i} = x_{2i})\} \\
 &= \sum_{i=1}^{56} \log\{P(X_{0i} = x_{0i})P(X_{1i} = x_{1i})P(X_{2i} = x_{2i})\} \\
 &= \sum_{i=1}^{56} \log\left\{\prod_{j=0}^2 \frac{e^{-\lambda_{ji}} \lambda_{ji}^{x_{ji}}}{x_{ji}!}\right\} \\
 &= \sum_{i=1}^{56} \sum_{j=0}^2 (-\lambda_{ji} + x_{ji} \log \lambda_{ji} - \log x_{ji}!). \tag{6.5}
 \end{aligned}$$

In the E-step, for iteration m , we calculate the expected values $z_i^{(m)}$ of X_{0i} conditional on Y_{1i} and Y_{2i} :

$$\begin{aligned}
 z_i^{(m)} &= E(X_{0i} | Y_{1i}, Y_{2i}) \\
 &= \sum_{x_{0i}=0}^{\infty} x_{0i} P(X_{0i} = x_{0i} | Y_{1i} = y_{1i}, Y_{2i} = y_{2i}) \\
 &= \sum_{x_{0i}=0}^{\infty} x_{0i} \frac{P(X_{0i} = x_{0i}, Y_{1i} = y_{1i}, Y_{2i} = y_{2i})}{P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i})} \\
 &= \sum_{x_{0i}=0}^{\infty} x_{0i} \frac{P(X_{0i} = x_{0i})P(X_{1i} = y_{1i} - x_{0i})P(X_{2i} = y_{2i} - x_{0i})}{P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i})} \\
 &= \lambda_0 \sum_{x_{0i}=1}^{\infty} \frac{P(X_{0i} = x_{0i} - 1)P(X_{1i} = y_{1i} - x_{0i})P(X_{2i} = y_{2i} - x_{0i})}{P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i})}
 \end{aligned}$$

now let $r = x_{0i} - 1$:

$$= \lambda_0 \sum_{r=0}^{\infty} \frac{P(X_{0i} = r)P(X_{1i} = y_{1i} - r - 1)P(X_{2i} = y_{2i} - r - 1)}{P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i})}$$

but $Y_{ji} = X_{0i} + X_{ji} = r + y_{ji} - r - 1 = y_{ji} - 1$, and hence

$$= \lambda_0^{(m)} \frac{P(Y_{1i} = y_{1i} - 1, Y_{2i} = y_{2i} - 1)}{P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i})}. \tag{6.6}$$

Then, in the M-step, we replace the x_{0i} by their conditional expectations $z_i^{(m)}$ to maximize the complete data log-likelihood function given in (6.5), which is linear in x_{0i} . Because the X_{ji} are independent, this can be done by fitting three Poisson generalized linear models (GLMs) using $y_{1i} - z_i^{(m)}$, $y_{2i} - z_i^{(m)}$, and $z_i^{(m)}$ as response variables and, in our case, intercept-only linear predictors, giving $\lambda_1^{(m+1)}$, $\lambda_2^{(m+1)}$, and $\lambda_0^{(m+1)}$, respectively. This process is iterated until the difference between the complete data log-likelihood functions evaluated at two successive iterations is less than some chosen small number, say 10^{-8} . To obtain standard errors of the maximum likelihood estimates, we estimated the model for 1000 bootstrap samples obtained from the fitted model and calculated the standard errors from the samples of estimates, hence a parametric bootstrap approach.

We fitted the bivariate Poisson model and two independent Poisson GLMs to the data and the estimates are displayed in Table 6.2. The residual deviances for the independent Poisson models

were 88.17 for the stinkbug and 65.10 for the earwig, both on 55 degrees of freedom. There is evidence of overdispersion for the stinkbug data ($p < 0.01$) but not for the earwig ($p = 0.17$), and this may be due to the excessive number of zero counts in the stinkbug data. The likelihood-ratio between the two independent Poisson model fits and the bivariate Poisson fit is 4.49 on 1 degree of freedom, giving evidence of a correlation between both variables that is different from zero ($p = 0.03$), $\hat{\lambda}_0 = e^{-1.66} = 0.19$. The bivariate model provides a slightly better fit than the independence model.

Table 6.2. Parameter estimates (standard errors) for both models fitted to the attack data

Parameter	Assuming independence	Estimating covariance
$\log \lambda_1$ (stinkbug)	-0.85 (0.20)	-1.43 (0.66)
$\log \lambda_2$ (earwig)	1.00 (0.08)	0.93 (0.09)
$\log \lambda_0$	0.00 (-)	-1.66 (0.71)
$-2 \times \log \text{lik}$	332.32	327.83

To produce the bivariate residual plot with simulation polygons we used the deviance residuals, which may be defined as

$$r_{ji}^d = \text{sign}(y_{ji} - \hat{\mu}_j) \sqrt{2} \left\{ y_i \log \frac{y_i}{\hat{\mu}_j} + \hat{\mu}_j - y_{ji} \right\}^{\frac{1}{2}}$$

and the PIT diagnostics,

$$r_{ji}^{pit} = e^{-\hat{\mu}_j} \sum_{x=0}^{y_{ji}} \frac{\hat{\mu}_j^x}{x!},$$

with $\hat{\mu}_j = \hat{\lambda}_0 + \hat{\lambda}_j$. Since the distribution here is discrete, we used the randomized version of the PIT diagnostics, as described in equation (6.1). Looking at the bivariate residual plots with simulation polygons, it is clear that the inclusion of the covariance between the variables improves model goodness-of-fit (see Figure 6.3). In this case, the use of the deviance residuals was important to aid in the detection of possible outliers related to the stinkbug attack counts. Finally, we were able to conclude that not only are earwigs more aggressive than stinkbugs, but also the correlation between the attack counts gives evidence that the stinkbugs may, in some cases, opt to defend themselves by counterattacking as observed throughout the experiment.

6.3.3 Longitudinal counts of two species on different sites

As a final illustration, we simulate data from a mixture model used to analyse longitudinal counts of animals on different sites. Let Y_{1ij} and Y_{2ij} be the number of animals of two species recorded at site i on occasion j , with $i = 1, \dots, I$, and $j = 1, \dots, J$. Royle (2004) supposes that these observations represent a proportion of the true abundances N_{1i} and N_{2i} of these species at each site and constructs a hierarchical model of the form

$$\begin{aligned} Y_{sit} | N_{si} &\sim \text{Bin}(N_{si}, p_{sij}) \\ N_{si} &\sim \text{P}(\lambda_{si}), \end{aligned} \tag{6.7}$$

where $s = 1, 2$, p_{sij} are called detection probabilities and λ_{si} are abundance parameters, which can be modelled with covariates.

To extend the model and include a correlation between the species, which may correspond to a site-association between them, Moral et al. (*unpublished paper*¹) propose the joint hierarchical

¹MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. B.; REIGADA, C.; GODOY, W. A. C. Models for jointly estimating abundances of two unmarked site-associated species subject to imperfect detection. *Submitted to JABES*.

Figure 6.3. Diagnostic plots for the stinkbug (Residuals 1) and earwig (Residuals 2) attack data using deviance residuals (a and b) and randomized PIT diagnostics (c and d), while assuming there was no covariance between the two random variables (a and c) or estimating this covariance with a joint model (b and d). Filled red points are outside of their respective simulated polygon; there are (a) 16, (b) 7, (c) 20, and (d) 8 points outside of the simulated polygons out of 56 total points.

formulation

$$\begin{aligned}
 Y_{1it}|N_{1i} &\sim \text{Bin}(N_{1i}, p_{1ij}) \\
 N_{1i} &\sim P(\lambda_{1i}) \\
 Y_{2it}|N_{1i}, N_{2i} &\sim \text{Bin}(N_{2i}, p_{2ij}) \\
 N_{2i}|N_{1i} &\sim P(\lambda_{2i} + \psi_i N_{1i})
 \end{aligned} \tag{6.8}$$

which yields

$$\text{Corr}(Y_{1ij}, Y_{2ij}) = \psi_2 \sqrt{\frac{\lambda_{1i} p_{1ij} p_{2ij}}{\lambda_{2i} + \lambda_{1i} \psi_2 (1 + \psi_2 p_{2ij})}}.$$

Here, for simplification purposes, when simulating data from this joint model we suppose that the parameters do not vary across sites or time occasions and hence take $p_{sit} = p_s$, $\lambda_{si} = \lambda_s$ and $\psi_i = \psi$. The likelihood functions for each species are

$$L_1(p_1, \lambda_1) = \prod_{i=1}^I \left\{ \sum_{n_{1i}=\max_j \{y_{1ij}\}}^{\infty} \left(\prod_{j=1}^J P(Y_{1ij} = y_{1ij} | N_{1i} = n_{1i}) \right) P(N_{1i} = n_{1i}) \right\}$$

$$L_2(p_2, \psi, \lambda_2) = \prod_{i=1}^I \left\{ \sum_{n_{2i}=\max_j \{y_{2it}\}}^{\infty} \left(\prod_{j=1}^J P(Y_{2ij} = y_{2ij} | N_{1i} = n_{1i}, N_{2i} = n_{2i}) \right) \right. \\ \left. \times \sum_{n_{1i}=\max_j \{y_{1it}\}}^{\infty} P(N_{2i} = n_{2i} | N_{1i} = n_{1i}) P(N_{1i} = n_{1i}) \right\},$$

which result in the joint likelihood function

$$L(p_1, p_2, \psi, \lambda_1, \lambda_2) = L_1(p_1, \lambda_1) \times L_2(p_2, \psi, \lambda_2)$$

$$= \prod_{i=1}^I e^{-(2\lambda_1 + \lambda_2)} \left\{ \sum_{n_{1i}=\max_j \{y_{1ij}\}}^{\infty} \left(\lambda_1^{n_{1i}} \left[\prod_{j=1}^J \frac{p_1^{y_{1ij}} (1-p_1)^{n_{1i}-y_{1ij}}}{y_{1ij}! (n_{1i}-y_{1ij})!} \right] \right) \right. \\ \left. \times \sum_{n_{1i}=\max_j \{y_{1ij}\}}^{\infty} \left(\frac{e^{-\psi n_{1i}} \lambda_1^{n_{1i}}}{n_{1i}!} \sum_{n_{2i}=\max_j \{y_{2it}\}}^{\infty} (\psi + \lambda_2)^{n_{2i}} \left[\prod_{j=1}^J \frac{p_2^{y_{2it}} (1-p_2)^{n_{2i}-y_{2it}}}{y_{2it}! (n_{2i}-y_{2it})!} \right] \right) \right\}$$

to be maximised for maximum likelihood estimation. For optimization purposes, the infinite sums are truncated at a large number M , chosen so that the Poisson tail probabilities are of the order of magnitude 10^{-10} , as suggested by Dennis et al. (2015). The detection probabilities and abundance parameters were estimated on the logit and logarithmic scale to constrain the estimates during the optimization process, which used the BFGS algorithm as implemented in the function `optim` in R.

We simulated data for $I = 20$ sites and $J = 40$ time occasions in R, because this model behaves better under larger samples. We used seeds 1234 and 2345 for the true latent abundances N_{1i} and N_{2i} , respectively, and the observed data y_{1ij} and Y_{2ij} using seeds 3456 and 4567, respectively, to allow for reproducibility. We fixed the true parameter values as $p_1 = 0.6$, $p_2 = 0.3$, supposing a lower detection probability for the dependent species, $\lambda_1 = 3$, $\lambda_2 = 2$, and $\psi = 2.5$, which gives $\text{Corr}(Y_{1ij}, Y_{2ij}) = 0.47$. We then fitted the joint model both estimating ψ and setting $\psi = 0$, which is equivalent to fitting two independent models, as in equation (6.8). The parameter estimates for species 1 were practically the same for both approaches, but it was clear that the estimation of the correlation parameter λ_2 significantly contributed to the log-likelihood (13.60 on 1 d.f.), see Table 6.3.

Table 6.3. Parameter estimates (standard errors) for both models fitted to the simulated correlated longitudinal species count data and true values

Parameter	Assuming independence	Estimating covariance	True value
logit(p_1)	0.43 (0.00)	0.43 (0.00)	0.41
logit(p_2)	-0.74 (0.01)	-0.90 (0.02)	-0.85
log λ_1	1.31 (0.01)	1.31 (0.01)	1.39
log λ_2	2.33 (0.01)	0.53 (1.15)	0.69
log ψ	0.00 (-)	0.96 (0.06)	0.92
$-2 \times \text{loglik}$	4939.36	4925.70	-

Proceeding to the construction of the bivariate residual plot with simulation polygons for these

models, we may write the raw residuals as

$$\begin{aligned} r_{1ij}^o &= y_{1ij} - \hat{\mu}_{1ij}, & \hat{\mu}_{1ij} &= E(Y_{1ij}) = E_{N_{1i}}\{E_{Y_{1ij}}(Y_{1ij}|N_{1i})\} = \hat{\lambda}_1 \hat{p}_1 \\ r_{2ij}^o &= y_{2ij} - \hat{\mu}_{2ij}, & \hat{\mu}_{2ij} &= E(Y_{2ij}) = E_{N_{1i}, N_{2i}}\{E_{Y_{2ij}}(Y_{2ij}|N_{1i}, N_{2i})\} = (\hat{\lambda}_2 + \hat{\psi} \hat{\lambda}_1) \hat{p}_2 \end{aligned}$$

using theorems of conditional expectation. To obtain the PIT diagnostics, we need to derive the marginal distributions of Y_{1ij} and Y_{2ij} . We have, for Y_{1ij} , that

$$\begin{aligned} P(Y_{1ij} = y_{1ij}) &= \sum_{n_{1i}=0}^{\infty} P(Y_{1ij} = y_{1ij}|N_{1i} = n_{1i})P(N_{1i} = n_{1i}) \\ &= \sum_{n_{1i}=0}^{\infty} \binom{n_{1i}}{y_{1ij}} p_1^{y_{1ij}} (1-p_1)^{n_{1i}-y_{1ij}} \frac{e^{-\lambda_1} \lambda_1^{y_{1ij}}}{n_{1i}!} \\ &= \frac{(\lambda_1 p_1)^{y_{1ij}} e^{-\lambda_1}}{y_{1ij}!} \sum_{n_{1i}=1}^{\infty} \frac{[\lambda_1(1-p_1)]^{n_{1i}-y_{1ij}}}{(n_{1i}-y_{1ij})!} \\ &= \frac{(\lambda_1 p_1)^{y_{1ij}} e^{-\lambda_1} e^{\lambda_1(1-p_1)}}{y_{1ij}!} \\ &= \frac{e^{-\lambda_1 p_1} (\lambda_1 p_1)^{y_{1ij}}}{y_{1ij}!} \end{aligned}$$

and hence $Y_{1ij} \sim P(\lambda_1 p_1)$, marginally. For Y_{2ij} , we have

$$\begin{aligned} P(Y_{2ij} = y_{2ij}) &= \sum_{n_{1i}=0}^{\infty} \sum_{n_{2i}=0}^{\infty} P(Y_{2ij} = y_{2ij}|N_{1i} = n_{1i}|N_{2i} = n_{2i})P(N_{2i} = n_{2i}|N_{1i} = n_{1i})P(N_{1i} = n_{1i}) \\ &= \sum_{n_{1i}=0}^{\infty} \sum_{n_{2i}=0}^{\infty} \binom{n_{2i}}{y_{2ij}} p_2^{y_{2ij}} (1-p_2)^{n_{2i}-y_{2ij}} \frac{e^{-(\lambda_2 + \psi n_{1i})} (\lambda_2 + \psi n_{1i})^{n_{2i}}}{n_{2i}!} \frac{e^{-\lambda_1} \lambda_1^{n_{1i}}}{n_{1i}!} \\ &= \sum_{n_{1i}=0}^{\infty} \frac{e^{-(\lambda_2 + \psi n_{1i}) p_2} [(\lambda_2 + \psi n_{1i}) p_2]^{y_{2ij}}}{y_{2ij}!} \frac{e^{-\lambda_1} \lambda_1^{n_{1i}}}{n_{1i}!} \\ &= \frac{e^{-(\lambda_2 p_2 + \lambda_1)} p_2^{y_{2ij}}}{y_{2ij}!} \sum_{n_{1i}}^{\infty} \frac{(\lambda_1 e^{-\psi p_2})^{n_{1i}} (\lambda_2 + \psi n_{1i})^{y_{2ij}}}{n_{1i}!}, \end{aligned}$$

which is analogous to a Neyman type-A distribution, and hence an overdispersion model. Therefore, the PIT diagnostics may be written as

$$\begin{aligned} r_{1ij}^{pit} &= e^{-\hat{\lambda}_1 \hat{p}_1} \sum_{k=0}^{y_{1ij}} \frac{(\hat{\lambda}_1 \hat{p}_1)^k}{k!} \\ r_{2ij}^{pit} &= \sum_{k=0}^{y_{2ij}} \left\{ \frac{e^{-(\hat{\psi} \hat{p}_2 + \hat{\lambda}_1)} \hat{p}_2^k}{k!} \sum_{n_1=0}^{\infty} \frac{(\hat{\lambda}_1 e^{-\hat{\psi} \hat{p}_2})^{n_1} (\hat{\lambda}_2 + \hat{\psi} n_1)^k}{n_1!} \right\} \end{aligned}$$

and due to the discrete nature of these models we use the randomized PIT diagnostics, as described by equation (6.1).

Once again, fitting the models independently, without considering the correlation between the variables, yields a poor model fit when compared with the joint modelling framework (see Figure 6.4). Also, when looking univariately, it seems that both models adequately fit the data, which is clearly not the case, and looking at the bivariate residual plots the lack-of-fit became clearer and was suitable to aid in assessing goodness-of-fit.

6.4 Discussion

We have proposed a graphical tool to help in the goodness-of-fit assessment for bivariate models and have shown how it can be applied to different types of models, for continuous and discrete data.

Figure 6.4. Diagnostic plots for the bivariate longitudinal Poisson-binomial mixture simulated data using standardized residuals (a and b) and randomized PIT diagnostics (c and d), while assuming there was no covariance between the two random variables (a and c) or estimating this covariance with a joint model (b and d). Filled red points are outside of their respective simulated polygon; there are (a) 116, (b) 66, (c) 100, and (d) 11 points outside of the simulated polygons out of 200 total points.

This approach can be used for any type of joint model for which it is possible to obtain specific model diagnostics and simulate the bivariate response variables from the fitted model.

Different diagnostics may result in better ways to visualise possible problems with particular data points. For example, the use of randomised PIT diagnostics, as seen in Figures 6.3 and 6.4, can be useful for discrete models, whose residuals may present a clumping of observations. Other types of diagnostics not explored here, for example the multiple-imputation based residuals for joint models (see Rizopoulos et al., 2010), can perhaps provide a better display. Moreover, as this is a simulation-based procedure, every time the plot is produced the polygons change a little and hence not only do we advocate the use of different types of model diagnostics but we also suggest that several plots are produced.

It is important to note that for complex models this procedure can be very computationally

intensive. Another potential drawback is how outliers obtained in the simulations might affect the convex hulls and hence the reduced-area polygons. An alternative would be to perform a convex hull peeling, removing the first hull and obtaining a second one, which would form the simulation polygon, analogously to the proposal in Bebbington (1978), which is to remove the points that lie on the outer convex hull prior to estimating a correlation coefficient. However, we have experimented with this strategy and concluded that it reduces the polygon area too much and also the points lying on the outer convex hull were generated by the fitted model and therefore the small area reduction can be enough to correct for extremeness.

To help examining the behaviour of bivariate data and detect outliers, bivariate extensions to the boxplot have been proposed, such as the bagplot (Rousseeuw et al., 1999) and the relplot, half and full quelplot (Goldberg and Iglewicz, 1992). Moreover, density ellipsoids, convex hulls and two-dimensional kernel density estimation can also be used. For example, with these approaches, we can identify a point that deviates too much from the central mass of the data and fit a model with and without that particular observation. Then, we can use the difference in the log-likelihoods to assess whether that point was an outlier. However, the approach described in this paper is different in the sense that no assumption is made about the behaviour of the bivariate diagnostics other than, if the observed data can truly be generated from the fitted model, the diagnostics obtained via simulation would display a consistent behaviour.

A different strategy, and perhaps a way of extending this framework to a n -variate setting, can be devised using the ideas from Healy (1968) and Gnanadesikan and Kettenring (1972). By using generalised distances from the mean, weighted by the variances and covariances of the variables, we obtain a univariate set of points which can be used to produce a half-normal plot with a simulated envelope, for example. Another way of reducing the dimensionality of the problem is to first perform a principal component analysis on the n -variate model diagnostics and produce the bivariate residual plot with simulation polygons for the first two principal components. This has not been tested yet and is subject of ongoing work. We also do not rule out other forms of diagnostic checking, either analytical or graphical. Our approach represents a potentially helpful framework for assessing goodness-of-fit for bivariate models.

References

- ATKINSON, A. C. (1985) **Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis**. Oxford: Clarendon Press. 282 p.
- BARNETT, C.; LEWIS, T. (1994) **Outliers in statistical data**. Chichester: Wiley. 608 p.
- BEBBINGTON, A. C. (1978) A method of bivariate trimming for robust estimation of the correlation coefficient. **Journal of the Royal Statistical Society C – Applied Statistics**, v. 27, p. 221–226.
- CODY, W. D. (1993) Algorithm 715: SPECFUN – A portable FORTRAN package of special function routines and test drivers. **ACM Transactions on Mathematical Software**, v. 19, p. 22–32.
- CZADO, C.; GNEITING, T.; HELD, L. (2009) Predictive model assessment for count data. **Biometrics**, v. 65, p. 1254–1261.
- DAWID, A. P. (1984) Statistical theory: The prequential approach. **Journal of the Royal Statistical Society, Series A**, v. 147, p. 278–292.
- DENNIS, E. B.; MORGAN, B. T. J.; RIDOUT, M. S. (2015) Computational aspects of N-mixture models. **Biometrics**, v. 71, p. 237–246.

GENZ, A.; BRETZ, F. (2009) **Computation of multivariate normal and t probabilities**. Berlin: Springer. 126 p.

GNANADESIKAN, R.; KETTENRING, J. R. (1972) Robust estimates, residuals, and outlier detection with multiresponse data. **Biometrics**, v. 28, p. 81–124.

GOLDBERG, K. M.; IGLEWICZ, B. (1992) Bivariate extensions of the boxplot. **Technometrics**, v. 34, p. 307–320.

GOURIEROUX, C.; MONFORT, A.; TROGNON, A. (1984) Pseudo maximum likelihood methods: Applications to Poisson models. **Econometrica**, v. 52, p. 701–720.

HEALY, M. J. R. (1968) Multivariate normal plotting. **Journal of the Royal Statistical Society C**, v. 17, p. 157–161.

HO, L. L.; SINGER, J. M. (2001) Generalized least squares methods for bivariate Poisson regression. **Communications in Statistics – Theory and Methods**, v. 30, p. 263–277.

JOHNSON, N.; KOTZ, S.; BALAKRISHNAN, N. (1997) **Discrete multivariate distributions**. New York: Wiley. 328 p.

JUNG, R. C.; WINKELMANN, R. (1993) Two aspects of labor mobility: A bivariate Poisson regression approach. **Empirical Economics**, v. 18, p. 543–556.

KARLIS, D.; NTZOUFRAS, I. (2005) Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R. **Journal of Statistical Software**, v. 14, p. 1–36.

KOCHERLAKOTA, S.; KOCHERLAKOTA, K. (2001) Regression in the bivariate Poisson distribution. **Communications in Statistics – Theory and Methods**, v. 30, p. 815–825.

MORAL, R. A. (2012) Modelagem estatística e ecológica de pragas e inimigos naturais. Master's dissertation.

MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. B. bivrp: (2016) Bivariate Residual Plots with Simulation Polygons. R package version 1.0, Available at: <<http://CRAN.R-project.org/=bivrp>>

MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. B. (*in press*) Half-normal plots and overdispersed models in R: the hnp package. **Journal of Statistical Software**.

R CORE TEAM. (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, Disponvel em: <<http://www.R-project.org>>

RIZOPOULOS, D.; VERBEKE, G.; MOLENBERGHS, G. (2010) Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. **Biometrics**, v. 66, p. 20–29.

ROUSSEEUW, P. J.; RUTS, I.; TUKEY, J. W. (1999) The bagplot: A bivariate boxplot. **The American Statistician**, v. 53, p. 382–387.

ROYLE, J. A. (2004) N-mixture models for estimating population size from spatially replicated counts. **Biometrics**, v. 60, p. 108–115.

SMITH, J. Q. (1985) Diagnostic checks of non-standard time series models. **Journal of Forecasting** v. 4, p. 283–291.

TSIONAS, E. G. (2001) Bayesian multivariate Poisson regression. **Communications in Statistics – Theory and Methods**, v. 30, p. 243–255.

7 FINAL CONSIDERATIONS

In this thesis, we aimed to develop and extend statistical models and methods to help in the analysis of entomological and ecological data. Initially, we proposed to develop a single R package with statistical tools for analysing entomological data. However, we felt the need to split the implementation of the different developed tools into three separate packages. These packages provide functions for fitting bivariate mixture models to ecological data (as in Chapter 4), and assessing goodness-of-fit for univariate models (Chapter 5) and bivariate models (Chapter 6).

The first package submitted to the Comprehensive R Archive Network (CRAN), **hnp**, started being developed during the Master's degree and the first version was finished throughout the doctorate. Then, we submitted our paper to Journal of Statistical Software, and the reviewers asked for a complete reformulation of the functions within the package. This resulted in a series of updates to the first version and in the acceptance of our paper, which is in this thesis as Chapter 5. The second package submitted to CRAN, **jointNmix**, was developed to implement the joint modelling framework described on Chapter 4, which consists most of the paper submitted to Journal of Agricultural, Biological, and Environmental Sciences. The third and last package resulting from this thesis, **bivrp**, includes functions to produce the bivariate residual plot with simulation polygons, as described on Chapter 6, which consists of the paper that will be submitted to Journal of Computational and Graphical Statistics.

Given the flexibility of the frameworks we have developed in this thesis, we also illustrated them with examples from different areas, such as ecological, medical and tissue culture studies. A thorough tutorial aimed to the applied entomologist of the use of package **hnp** when working with generalized linear models and extensions to account for overdispersion is available in Demétrio et al. (2014)¹.

Further work includes extensions to the joint N-mixture modelling framework that relax the closure assumption, thus including parameters that reflect on the birth, death, and migration rates. Moreover, the development and application of multivariate models for ecological surveys which record occurrence and/or counts of several species in different habitats and time points is also of interest. Another extension would be the relaxation of the independence assumption, and the inclusion of spatial correlation in the model. Improvements on the packages developed here are still subject of ongoing research, and we appreciate any suggestions and comments. Finally, we would be more than happy to share code that is not already available on CRAN upon requests made to the author².

¹Demétrio, C.G.B., Hinde, J., Moral, R.A. Models for overdispersed data in entomology. In Ferreira, C.P., Godoy, W.A.C. Ecological modelling applied to entomology, Springer, 2014.

²rafael.moral@usp.br

APPENDIX

Appendix A: Optimization tools and techniques

Introduction

In statistical modelling and data analysis, likelihood-based inference is very important. Fitting a statistical model by the frequentist approach involves obtaining the set of parameter estimates that maximise the likelihood function. Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ be a vector of independent and identically distributed random variables with density $f(y_i|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the $p \times 1$ vector of parameters of the distribution. For a realization $\mathbf{y} = (y_1, \dots, y_n)$ of $\mathbf{Y}$, the likelihood function is given by the joint density

$$L(\boldsymbol{\theta}|\mathbf{y}) \equiv f(y_1, \dots, y_n|\boldsymbol{\theta}) = \prod_{i=1}^n f(y_i|\boldsymbol{\theta}),$$

but viewed as a function of $\boldsymbol{\theta}$. This definition is the same for a set of correlated or non-identically distributed random variables, that is, the likelihood function is their joint density, although in those cases it cannot be written as the product of the marginal densities. In practice, it is more convenient to work with the natural logarithm of $L(\cdot)$, as the maxima of $l(\cdot) = \log L(\cdot)$ coincide. Not only does this simplify the mathematical analysis (e.g., allows one to avoid computing exponentials and hence derivatives become much simpler), but also using logarithms, from a standpoint of computational complexity, makes the computation of likelihoods less expensive. The products become summations on the log-scale and as likelihoods become very small, computer software would run out of floating point precision quickly, yielding a so-called underflow, which is a condition where the result of a specific calculation is a number smaller than the computer can store in its memory.

One way to obtain the maximum likelihood estimator for $\boldsymbol{\theta}$ analytically is to find the roots of the system of equations

$$\frac{\partial l(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}} \equiv \nabla l(\boldsymbol{\theta}|\mathbf{y}), \quad (\text{A.1})$$

which is called gradient or score function. For the general linear model, for example, this is a set of linear equations in the parameters and hence it is particularly simple to obtain the solution. When the score equations are nonlinear in $\boldsymbol{\theta}$, we must rely on numerical methods. These may focus on solving the score equations or finding the maximum of $l(\cdot)$ or the minimum of $-l(\cdot)$ directly without the use of the derivatives of the log-likelihood function. We will briefly discuss these methods in the following sections of this Appendix.

Direct-search methods

One way of finding the maximum of the log-likelihood function is to evaluate the function at possible candidate points until the maximum is found. Full blind search assumes the evaluation of the function at all possible alternatives for $\boldsymbol{\theta}$. As the full search space is tested, the maximum is always found. However, this type of search is only applicable to discrete search spaces and is not feasible when the search space is continuous or too large. Another disadvantage is the computational cost of evaluating the log-likelihood function over a large number of alternatives for $\boldsymbol{\theta}$.

The search space may be reduced using techniques such as the grid search, which implements a regular hyper-dimensional search (a “grid”) with a given step size and hence fewer alternatives are tested. A variant called nested grid search uses several step sizes, the first one being large, with a second grid level applied over the best point and so on until the maximum is found. This can be a problem when evaluating a multimodal likelihood function, as only a local maximum may be found instead of the global one. Even though this method is less computationally intensive than pure blind search, grid search may

also be infeasible in many circumstances when there are many parameters to be estimated. This method is particularly used for hyperparameter optimization of machine learning algorithms (Cortez, 2010).

Another blind search method is the Monte Carlo search, which involves generating N points randomly from a probability distribution over the search space instead of using regular step sizes. An advantage is that it is less computationally intensive than the previous two methods, as the number of evaluations of the log-likelihood function is fixed to N . However, each time the method is run, a different result may be obtained. Again, for a high-dimensional problem and/or a complex log-likelihood function, this method may be infeasible.

These methods are all blind search methods, which do not use information on a previous evaluation of the log-likelihood function to guide the next one. In the next sections, the methods which will be described are guided, which means that previous searches (for the maximum/minimum or for a solution of the score equations) are used to direct the current search.

Simulated annealing

The simulated annealing method was inspired from the annealing process in metallurgy, which involves heating a metal structure and then performing a controlled cooling (Cortez, 2010). In the statistical method, we use a temperature variable to simulate the heating process. Initially, it is set as high and then gets smaller as the algorithm runs. This temperature variable is directly proportional to the probability of the algorithm accepting worse solutions to the optimization problem than the current one. Hence, the method has the ability to jump from local optima to find the global optimum. As the temperature gets lower, the probability of accepting a worse solution becomes smaller.

The gradual decrease in the temperature variable is what makes this method really effective when the log-likelihood function has many local maxima. As the candidate solutions are generated randomly, this method is similar to the Monte Carlo search, however the search here is proportional to the temperature. In R, this method is implemented in function `optim` by setting the argument `method="SANN"`. An advantage is that it can be used when the log-likelihood function is non-differentiable, however it is usually slower to run when compared to the other optimization methods which will be discussed below.

Nelder-Mead method

The Nelder-Mead simplex method for function minimization was proposed by Nelder and Mead (1965). It involves the comparison of the values of $-l(\boldsymbol{\theta}|\mathbf{y})$ at the $(p + 1)$ vertices of a simplex. The method then proceeds by replacing the worst vertex of the simplex by a new one using one of three operations, called *reflection*, *expansion*, and *contraction*. Hence, after a number of operations the simplex contracts to the function's minimum.

For sake of simplicity, we will describe the method for a bivariate function, $h(x, y)$. In this case, the simplex has $2 + 1 = 3$ vertices and hence is a triangle. We need three starting values for (x, y) to form the vertices $\mathbf{A} = (x_1, y_1)$, $\mathbf{B} = (x_2, y_2)$, and $\mathbf{C} = (x_3, y_3)$, such that $h(x_1, y_1) < h(x_2, y_2) < h(x_3, y_3)$ and hence the vertex $\mathbf{A}$ is the point which is closest to the minimum and vertex $\mathbf{C}$ is the furthest. Then, the midpoint of side $\overline{\mathbf{AB}}$, $\mathbf{M} = \frac{1}{2}(\mathbf{A} + \mathbf{B})$ is calculated. The next step is to perform a *reflection* of the triangle $\overline{\mathbf{ABC}}$ through the side $\overline{\mathbf{AB}}$, by calculating the point $\mathbf{R} = 2\mathbf{M} - \mathbf{C} = (x_R, y_R)$ and forming the new triangle $\overline{\mathbf{ABR}}$, see Figure A.1(a).

If $h(x_R, y_R) < h(x_3, y_3)$, then the triangle moved to the right direction and we perform an *extension* of the triangle $\overline{\mathbf{ABR}}$ by replacing $\mathbf{R}$ by $\mathbf{E} = 2\mathbf{R} - \mathbf{M} = (x_E, y_E)$. If $h(x_E, y_E) < h(x_R, y_R)$, then the vertex $\mathbf{R}$ is replaced by $\mathbf{E}$, see Figure A.1(b). However, if $h(x_R, y_R) \geq h(x_3, y_3)$, then we perform a *contraction* of the triangle $\overline{\mathbf{ABC}}$ by obtaining the two midpoints $\mathbf{S}_1$ and $\mathbf{S}_2$ of the segments

Figure A.1. The operations of (a) reflection, (b) expansion, and (c) contraction used in the Nelder-Mead method, considering a simple example function of two variables

$\overline{CM}$ and $\overline{RM}$, respectively, and the point with the smaller function value $S = (x_S, y_S)$ replaces C to form the triangle $\overline{ABS}$ (see Figure A.1(c)) and the process starts over with another *reflection* operation until the minimum is obtained. If $h(x_R, y_R) < h(x_3, y_3)$, the vertices B and C are replaced by M and S , respectively, before proceeding with the operations. To compute the Hessian matrix, Nelder and Mead (1965) proposed using a quadratic approximation of the surface at the minimum.

This is the default method of optimization when R's `optim` function is called. It has the advantage of not using derivatives and requires only one or two evaluations of the function per iteration (except when $h(x_R, y_R) < h(x_3, y_3)$, but this happens rarely in practice).

Newton-Raphson method

The Newton-Raphson method is a technique for solving equations numerically. In the optimization context, it may be used to solve the score equations (A.1) and hence obtain maximum likelihood estimates. Consider a first-order Taylor series expansion of a given function $g(x)$ at $(x - x_0)$:

$$g(x) \approx g(x_0) + g'(x_0)(x - x_0).$$

Setting the approximation equal to zero and rearranging the terms yields the solution

$$x = x_0 - \frac{g(x_0)}{g'(x_0)}. \quad (\text{A.2})$$

Geometrically, this is the solution to the equation of the tangent line of $g(x)$ at x_0 , see Figure A.2. The Newton-Raphson method uses (A.2) to proceed iteratively, that is,

$$x_{(m+1)} = x_{(m)} - \frac{g(x_{(m)})}{g'(x_{(m)})}, \quad (\text{A.3})$$

until $|x_{(m+1)} - x_{(m)}| < \epsilon$, a very small constant.

In circumstances when $g'(x)$ is too difficult or impossible to derive or when the computational cost to compute the derivative is very large, the derivative may be approximated by a finite difference,

Figure A.2. Geometrical interpretation of the Newton-Raphson method. Starting at x_0 , x_1 is the solution of the line equation which is tangent to $g(x)$ at $g(x_0)$ and so on until convergence and the root of $g(x)$ is found

i.e.

$$g'(x_{(m)}) \approx \frac{g(x_{(m)}) - g(x_{(m-1)})}{x_{(m)} - x_{(m-1)}}$$

and the resulting method is called the secant (or *regula falsi*, the “false position”) method (Thisted, 1988).

The Newton-Raphson method may also be used for solving the system of score equations (A.1) through a multivariate generalization of (A.3), which may be written as

$$\boldsymbol{\theta}_{(m+1)} = \boldsymbol{\theta}_{(m)} - \mathbf{H}_{\boldsymbol{\theta}=\boldsymbol{\theta}_{(m)}}^{-1} \nabla l(\boldsymbol{\theta}_{(m)}|\mathbf{y}),$$

where

$$\mathbf{H}_{\boldsymbol{\theta}} = \frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$$

is the symmetric matrix of second-order derivatives of the log-likelihood function, called the Hessian matrix.

A variant of the Newton-Raphson method, called the Fisher scoring method, replaces the Hessian matrix by its expected value:

$$\mathfrak{I}_{\boldsymbol{\theta}} = -\mathbf{E}_{\mathbf{Y}}\{\mathbf{H}_{\boldsymbol{\theta}}\},$$

called Fisher’s information matrix, hence resulting in the following iterative process:

$$\boldsymbol{\theta}_{(m+1)} = \boldsymbol{\theta}_{(m)} + \mathfrak{I}_{\boldsymbol{\theta}=\boldsymbol{\theta}_{(m)}}^{-1} \nabla l(\boldsymbol{\theta}_{(m)}|\mathbf{y}).$$

Usually, the Newton-Raphson method converges more quickly, but it may not converge when the function presents multiple inflection points. In contrast, the Fisher scoring method converges more slowly but is more robust and may converge in cases where the Newton-Raphson method will not (KNIGHT, 2000). Other advantages of using the expected instead of the observed information matrix include the fact

that $\mathfrak{S}_\theta$ is more precise to estimate standard errors for maximum likelihood estimates (by taking the diagonal elements of $\mathfrak{S}_\theta^{-1}$) and there are cases where $\mathfrak{S}_\theta$ is of a simpler form than $\mathbf{H}_\theta$. However, for some model classes, the structure of the log-likelihood makes the evaluation of the expectation unfeasible or impractical.

Gauss-Seidel iteration

Another method for solving a system of nonlinear equations uses the Gauss-Seidel iteration. The method consists in partitioning the system of p score equations into p subproblems. By fixing the solutions for the other $p - 1$ equations to an initial guess, each of the nonlinear equations is now solved, one at a time, using any root-finding subroutine, such as the Newton-Raphson method (yielding the so-called Gauss-Seidel-Newton iteration). Then, the process iterates until convergence.

It is an easy-to-program method and another advantage is that when the score equations are linear in some parameters, say p_1 , then they can be decomposed in a linear block which is solved directly for those p_1 parameters and then the method moves on to solving the remaining $p - p_1$ equations one at a time. Ortega and Rheinboldt (1970) show that the method converges globally to a unique solution, given that $\mathbf{H}_\theta$ is symmetric and positive definite for all θ and its smallest eigenvalue is bounded away from zero.

Quasi-Newton methods

In some circumstances it might be difficult to obtain and hence program the Hessian matrix. In these situations, it can be approximated in much the same way as in the secant method, described in the previous section. Quasi-Newton methods update an approximation of the Hessian matrix at each iteration, hence the information on the curvature of $l(\theta|\mathbf{y})$ is re-evaluated at each iteration and is accumulated as the iterations go on. We may write the iterative process as

$$\theta_{(m+1)} = \theta_{(m)} - \mathbf{H}_{(m)}^{-1} \nabla l(\theta_{(m)}|\mathbf{y}).$$

One important quasi-Newton method is the Davidon-Fletcher-Powell algorithm (Davidon, 1959; Fletcher and Powell, 1963), which is well-known in statistical computing. Let $\mathbf{d}_{(m)} = \theta_{(m+1)} - \theta_{(m)}$ and let $\mathbf{g}_{(m)} = \nabla l(\theta_{(m+1)}|\mathbf{y}) - \nabla l(\theta_{(m)}|\mathbf{y})$, then the Davidon-Fletcher-Powell update for the inverse of the Hessian is

$$\mathbf{H}_{(m+1)}^{-1} = \mathbf{H}_{(m)}^{-1} + \frac{\mathbf{d}_{(m)}\mathbf{d}_{(m)}^T}{\mathbf{d}_{(m)}^T\mathbf{g}_{(m)}} - \frac{\mathbf{H}_{(m)}^{-1}\mathbf{g}_{(m)}\mathbf{g}_{(m)}^T\mathbf{H}_{(m)}^{-1}}{\mathbf{g}_{(m)}^T\mathbf{H}_{(m)}^{-1}\mathbf{g}_{(m)}}.$$

Another update for the Hessian matrix was proposed independently by Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970) and may be written as

$$\mathbf{H}_{(m+1)} = \mathbf{H}_{(m)} + \frac{\mathbf{g}_{(m)}\mathbf{g}_{(m)}^T}{\mathbf{d}_{(m)}^T\mathbf{g}_{(m)}} - \frac{\mathbf{H}_{(m)}\mathbf{d}_{(m)}\mathbf{d}_{(m)}^T\mathbf{H}_{(m)}}{\mathbf{d}_{(m)}^T\mathbf{H}_{(m)}\mathbf{d}_{(m)}}.$$

This is the well known BFGS update, implemented in optimization functions for R software, such as `optim` and `optimx`. The BFGS algorithm is considered to be superior to the Davidon-Fletcher-Powell algorithm based on extensive studies of numerical analysis (Thisted, 1988).

There is a limited-memory modification of the BFGS algorithm, called L-BFGS, and a version with box constraints, called L-BFGS-B (BYRD et al., 1995), also implemented in R's function `optim`. The latter is used when the optimization task is subject to bounded domains for one or more parameters in the likelihood, e.g., the variance of a normal distribution, which is strictly positive. When this is the case, different techniques can be used, such as step-halving, which simply uses the midpoint of an optimization step and the previous one when the guided search moves away from the bounded region of the parametric space, thus forcing the algorithm to search within the bounds.

The Expectation-Maximization algorithm

The Expectation-Maximization (EM) algorithm is a family of algorithms firstly proposed by Dempster, Laird and Rubin (1977). It provides an iterative procedure to obtain maximum likelihood estimates and is specifically useful in the presence of missing data and when working with mixtures (McLachlan and Krishnan, 1997). Again, we want to maximize $l(\boldsymbol{\theta}|\mathbf{y})$, but suppose that we are now working with the conditional distribution $\mathbf{Y}|\mathbf{Z}$, $\mathbf{Z}$ being latent continuous random variables with realizations $\mathbf{z}$. The marginal density of $\mathbf{Y}$ is now written as

$$f(\mathbf{y}|\boldsymbol{\theta}) = \int_{\mathbb{R}} f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta}) d\mathbf{z}.$$

Let the current estimate at the m -th iteration be represented by $\boldsymbol{\theta}_{(m)}$. Hence, we may write

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{y}) - l(\boldsymbol{\theta}_{(m)}|\mathbf{y}) &= \log \int_{\mathbb{R}} f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} - \log f(\mathbf{y}|\boldsymbol{\theta}_{(m)}) \\ &= \log \int_{\mathbb{R}} f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta}) \frac{f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(m)})}{f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(m)})} d\mathbf{z} - \log f(\mathbf{y}|\boldsymbol{\theta}_{(m)}) \\ &= \log \int_{\mathbb{R}} f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(m)}) \left[\frac{f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta})}{f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(m)})} \right] d\mathbf{z} - \log f(\mathbf{y}|\boldsymbol{\theta}_{(m)}) \\ (\text{Jensen's inequality}) &\geq \int_{\mathbb{R}} f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(m)}) \log \left[\frac{f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta})}{f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(m)})} \right] d\mathbf{z} - \log f(\mathbf{y}|\boldsymbol{\theta}_{(m)}) \end{aligned}$$

and therefore

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{y}) &\geq l(\boldsymbol{\theta}_{(m)}|\mathbf{y}) + \int_{\mathbb{R}} f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(m)}) \log \left[\frac{f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta})}{f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(m)})} \right] d\mathbf{z} - \log f(\mathbf{y}|\boldsymbol{\theta}_{(m)}) \\ &\triangleq l(\boldsymbol{\theta}|\boldsymbol{\theta}_{(m)}). \end{aligned}$$

This shows that the function $l(\boldsymbol{\theta}|\boldsymbol{\theta}_{(m)})$ is bounded above by the log-likelihood function and hence any $\boldsymbol{\theta}$ that increases the value of $l(\boldsymbol{\theta}|\boldsymbol{\theta}_{(m)})$ also increases the value of $l(\boldsymbol{\theta}|\mathbf{y})$. To achieve the greatest possible increase at the m -th iteration, we must maximize $l(\boldsymbol{\theta}|\boldsymbol{\theta}_{(m)})$, that is,

$$\begin{aligned} \boldsymbol{\theta}_{(m+1)} &= \arg \max_{\boldsymbol{\theta}} \{l(\boldsymbol{\theta}|\boldsymbol{\theta}_{(m)})\} \\ &= \arg \max_{\boldsymbol{\theta}} \left\{ l(\boldsymbol{\theta}_{(m)}|\mathbf{y}) + \int_{\mathbb{R}} f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(m)}) \log \left[\frac{f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta})}{f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(m)})} \right] d\mathbf{z} - \log f(\mathbf{y}|\boldsymbol{\theta}_{(m)}) \right\} \\ &= \arg \max_{\boldsymbol{\theta}} \left\{ \int_{\mathbb{R}} f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(m)}) \log [f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta})] d\mathbf{z} \right\} \\ &= \arg \max_{\boldsymbol{\theta}} \left\{ \int_{\mathbb{R}} f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(m)}) \log [f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})] d\mathbf{z} \right\} \\ &= \arg \max_{\boldsymbol{\theta}} \left\{ \mathbf{E}_{\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}_{(m)}} \{\log f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})\} \right\}. \end{aligned}$$

Therefore, each iteration of the EM algorithm consists in

- (i.) determining the conditional expectation $\mathbf{E}_{\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}_{(m)}} \{\log f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})\}$ (*E-step*) and
- (ii.) maximizing this expression with respect to $\boldsymbol{\theta}$ (*M-step*).

In essence, EM algorithms are a particular case of a class of algorithms called MM, which stands for majorization-minimization (for minimization problems) or minorization-maximization (for maximization problems) (Hunter and Lange, 2004). These are algorithms that replace the log-likelihood by a surrogate function, say $g(\boldsymbol{\theta}|\boldsymbol{\theta}_{(m)})$, which is less than or equal to the log-likelihood function for all $\boldsymbol{\theta}$, the equality satisfied if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_{(m)}$. In that sense, when the log-likelihood function has no closed form or is too complicated or computationally expensive, it may be replaced by an easier-to-compute function, which when iteratively maximized leads to the maximum of the log-likelihood function.

Example 1: Finite mixture of normal distributions

Suppose that for a given set of observations $y_i, i = 1, \dots, n$, one wishes to fit a mixture of K univariate normal distributions $N(\mu_k, \sigma^2)$. Let $P(\text{component} = k) = \pi_k$ be the probability of a component of the mixture, with $\sum_{k=1}^K \pi_k = 1$. We have that $P(y_i | \text{component} = k) = f(y_i | \mu_k, \sigma^2)$. Let G_{ik} be a dummy random variable, equal to 1 if observation i belongs to the k -th component of the mixture and equal to 0 otherwise. Supposing that the g_{ik} are known, we then have

$$P(y_i, g_{i1}, \dots, g_{iK}) = \prod_{k=1}^K [\pi_k f(y_i | \mu_k, \sigma^2)]^{g_{ik}}.$$

Then, the complete data likelihood and log-likelihood functions take the forms

$$\begin{aligned} L(\boldsymbol{\mu}, \sigma^2, \boldsymbol{\pi} | \mathbf{y}, \{g_{ik}\}) &= \prod_{i=1}^n \prod_{k=1}^K [\pi_k f(y_i | \mu_k, \sigma^2)]^{g_{ik}} \text{ and} \\ l(\boldsymbol{\mu}, \sigma^2, \boldsymbol{\pi} | \mathbf{y}, \{g_{ik}\}) &= \sum_{i=1}^n \sum_{k=1}^K g_{ik} [\log \pi_k + \log f(y_i | \mu_k, \sigma^2)], \end{aligned}$$

respectively.

In the E-step, we replace the g_{ik} by their conditional expectations

$$\begin{aligned} w_{ik} &\equiv E(G_{ik} | Y_i) = \sum_{g_{ik}=0}^1 g_{ik} P(G_{ik} = g_{ik} | y_i) = P(G_{ik} = 1 | y_i) = P(\text{component} = k | y_i) \\ &= \frac{P(\text{component} = k) P(Y_i = y_i | \text{component} = k)}{\sum_{j=1}^K P(\text{component} = j) P(Y_i = y_i | \text{component} = j)} \\ &= \frac{\pi_k f(y_i | \mu_k, \sigma^2)}{\sum_{j=1}^K \pi_j f(y_i | \mu_j, \sigma^2)} = \frac{\pi_k \exp \left\{ -\frac{(y_i - \mu_k)^2}{2\sigma^2} \right\}}{\sum_{j=1}^K \pi_j \exp \left\{ -\frac{(y_i - \mu_j)^2}{2\sigma^2} \right\}} \end{aligned}$$

which is the probability that observation y_i belongs to the k -th component of the finite mixture.

Then, in the M-step, we maximize the complete data log-likelihood function $l(\boldsymbol{\mu}, \sigma^2, \boldsymbol{\pi} | \mathbf{y}, \{w_{ik}\})$ with respect to μ_k, σ^2 and π_k . For μ_k and σ^2 the solutions are obtained directly:

$$\begin{aligned} \frac{\partial l}{\partial \mu_k} &= \sum_{i=1}^n w_{ik} \frac{\partial \log f(y_i | \mu_k, \sigma^2)}{\partial \mu_k} \\ &\Rightarrow \sum_{i=1}^n w_{ik} \frac{y_i - \hat{\mu}_k}{\sigma^2} = 0 \\ &\Leftrightarrow \hat{\mu}_k = \frac{\sum_{i=1}^n w_{ik} y_i}{\sum_{i=1}^n w_{ik}} \text{ and} \\ \frac{\partial l}{\partial \sigma^2} &= \sum_{i=1}^n \sum_{k=1}^K w_{ik} \frac{\partial \log f(y_i | \mu_k, \sigma^2)}{\partial \sigma^2} = 0 \\ &\Rightarrow \sum_{i=1}^n \sum_{k=1}^K \frac{w_{ik}}{\hat{\sigma}^3} \{(y_i - \mu_k)^2 - \hat{\sigma}^2\} = 0 \\ &\Leftrightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K (y_i - \mu_k)^2 \end{aligned}$$

As we have the constraint $\sum_{k=1}^K \pi_k = 1$, we add a Lagrange multiplier to the log-likelihood so that we

have

$$\begin{aligned} l^*(\boldsymbol{\pi}, \lambda | \mathbf{y}, \{w_{ik}\}) &= l(\boldsymbol{\mu}, \sigma^2, \boldsymbol{\pi} | \mathbf{y}, \{w_{ik}\}) - \lambda \left(\sum_{k=1}^K \pi_k - 1 \right) \\ &\propto \sum_{i=1}^n \sum_{k=1}^K w_{ik} \log \pi_k - \lambda \left(\sum_{k=1}^K \pi_k - 1 \right). \end{aligned}$$

We then solve the system of $K + 1$ equations

$$\begin{cases} \frac{\partial l^*}{\partial \pi_k} = 0 \\ \frac{\partial l^*}{\partial \lambda} = 0 \end{cases}$$

to obtain

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^n w_{ik}.$$

Finally, we iterate between the E and M steps until convergence.

Example 2: Mixed models

In a mixed-effects model, the random effects may be viewed as missing data. Suppose we have a longitudinal study with $i = 1, \dots, N$ groups and $j = 1, \dots, n_i$ observations within each group. Let the distribution of the observed data y_{ij} , conditional on the random effects b_i , have the density $f(y_{ij} | b_i, \boldsymbol{\theta}_1)$ and the distribution of the random effects have the density $f(b_i | \boldsymbol{\theta}_2)$. Let $\boldsymbol{\Psi} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$. Then, an EM algorithm for this case iteratively maximizes, with respect to $\boldsymbol{\Psi}$,

$$\begin{aligned} &\mathbf{E}_{B|Y, \boldsymbol{\Psi}_{(m)}} \{ \log f(Y, B | \boldsymbol{\Psi}) \} \\ &= \int_{\mathbb{R}} \log[f(\mathbf{y}, \mathbf{b} | \boldsymbol{\Psi})] f(\mathbf{b} | \mathbf{y}, \boldsymbol{\Psi}_{(m)}) d\mathbf{b} \\ &= \int_{\mathbb{R}} \log[f(\mathbf{y} | \mathbf{b}, \boldsymbol{\theta}_1) f(\mathbf{b} | \boldsymbol{\theta}_2)] \frac{f(\mathbf{y} | \mathbf{b}, \boldsymbol{\theta}_{1(m)}) f(\mathbf{b} | \boldsymbol{\theta}_{2(m)})}{\int_{\mathbb{R}} f(\mathbf{y} | \mathbf{b}, \boldsymbol{\theta}_{1(m)}) f(\mathbf{b} | \boldsymbol{\theta}_{2(m)}) d\mathbf{b}} d\mathbf{b} \\ &= k^{-1} \int_{\mathbb{R}} \log[f(\mathbf{y} | \mathbf{b}, \boldsymbol{\theta}_1) f(\mathbf{b} | \boldsymbol{\theta}_2)] f(\mathbf{y} | \mathbf{b}, \boldsymbol{\theta}_{1(m)}) f(\mathbf{b} | \boldsymbol{\theta}_{2(m)}) d\mathbf{b}, \end{aligned}$$

where

$$k = \int_{\mathbb{R}} f(\mathbf{y} | \mathbf{b}, \boldsymbol{\theta}_{1(m)}) f(\mathbf{b} | \boldsymbol{\theta}_{2(m)}) d\mathbf{b}.$$

A straightforward way to maximize $\mathbf{E}_{B|Y, \boldsymbol{\Psi}_{(m)}} \{ \log f(Y, B | \boldsymbol{\Psi}) \}$ is by solving

$$\frac{d}{d\boldsymbol{\Psi}} \mathbf{E}_{B|Y, \boldsymbol{\Psi}_{(m)}} \{ \log f(Y, B | \boldsymbol{\Psi}) \} = \mathbf{0}.$$

If we approximate the integral with a q point Gaussian quadrature, we may write

$$\begin{aligned} \frac{d}{d\boldsymbol{\Psi}} \mathbf{E}_{B|Y, \boldsymbol{\Psi}_{(m)}} \{ \log f(Y, B | \boldsymbol{\Psi}) \} &= k^{-1} \int_{\mathbb{R}} \frac{d \log[f(\mathbf{y} | \mathbf{b}, \boldsymbol{\theta}_1) f(\mathbf{b} | \boldsymbol{\theta}_2)]}{d\boldsymbol{\Psi}} f(\mathbf{y} | \mathbf{b}, \boldsymbol{\theta}_{1(m)}) f(\mathbf{b} | \boldsymbol{\theta}_{2(m)}) d\mathbf{b} \\ &= \sum_{p=1}^q w_p \frac{d \log[f(\mathbf{y} | \mathbf{b}, \boldsymbol{\theta}_1) f(\mathbf{b} | \boldsymbol{\theta}_2)]}{d\boldsymbol{\Psi}}, \end{aligned}$$

where $\mathbf{b}_p$ replace $\mathbf{b}$ as the quadrature points with associated weights $w_p = k^{-1} f(\mathbf{y} | \mathbf{b}_p, \boldsymbol{\theta}_{1(m)}) f(\mathbf{b}_p | \boldsymbol{\theta}_{2(m)})$. Anderson and Hinde (1988) show how this can be used in a generalized linear mixed modelling framework and give a five-step algorithm for fitting these models using an EM algorithm based on the one described above.

Discussion

Currently there are many optimization algorithms implemented in R which are available in the `stats` package, one of the core packages of the software. A widely used function is `optim`, whose default method is the Nelder-Mead ("**Nelder-Mead**"). It also provides the Brent method ("**Brent**") for univariate optimization problems and the BFGS ("**BFGS**"), L-BFGS-B ("**L-BFGS-B**"), conjugate gradient ("**CG**") and simulated annealing ("**SANN**") methods for multivariate optimization. A Newton-like method is also available and implemented in the function `nlm`. A quasi-Newton method is implemented in the function `nlminb`, and a constrained version of `optim` is `constrOptim`, which uses an adaptive barrier algorithm, but only the Nelder-Mead and BFGS methods are available in this case. Also, functions `optimize` and `uniroot` offer a minimization and root-finding algorithm, respectively, for univariate optimization problems. There are many packages available on the Comprehensive R Archive Network that perform optimization tasks for a range of different specific problems. The webpage <https://cran.r-project.org/web/views/Optimization.html> (access in 10 April, 2016) displays an overview of these packages and their functions.

It is important to note that there is no stand-out optimization method for all possible problems, but the objective is to find the most sensible approach for each specific optimization problem. For example, MacDonald (2014) draws attention to the use of the EM algorithm as the one-and-only method for obtaining maximum likelihood estimates for a range of examples where the use of state-of-the-art numerical maximizers would solve the problems more easily. The advances of computational engineering make it possible for new and improved implementations of methods which may supersede older ones (e.g., the use of `optimx` and `optreplace` instead of `optim` is recommended by Nash (2014)). For a short review on optimization alternatives focussing on R software and suggestions on ways of moving forward towards best practice optimization approaches, see Nash (2014).

References

- ANDERSON, D. A.; HINDE, J. P. (1988) Random effects in generalized linear models and the EM algorithm. **Communications in Statistics - Theory and Methods**, v. 17, p. 3847–3856.
- CORTEZ, P. (2010) **Modern optimization with R**, New York: Springer. 182 p.
- BROYDEN, C. G. (1970) The convergence of a class of double-rank minimization algorithms 2. The new algorithm. **IMA Journal of Applied Mathematics**, v. 6, p. 222–231.
- DAVIDON, W. C. (1959) Variable metric algorithm for minimization. **Argonne National Laboratory Report**, ANL-2990, Illinois.
- DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society: Series B**, v. 39, p. 1–38.
- FLETCHER, R. (1970) A new approach to variable metric algorithms. **The Computer Journal**, v. 13, p. 317–322.
- FLETCHER, R.; POWELL, M. J. D. (1963) A rapidly convergent descent method for minimization. **The Computer Journal**, v. 6, p. 163–168.
- GOLDFARB, D. (1970) A family of variable-metric methods derived by variational means. **Mathematics of Computation**, v. 24, p. 23–26.
- HUNTER, D.; LANGE, K. (2004) A tutorial on MM algorithms. **The American Statistician**, v. 58, p. 30–37.

- KNIGHT, K. (2000) **Mathematical statistics**, Boca Raton: Chapman and Hall. 504 p.
- MACDONALD, I. L. (2014) Numerical maximisation of likelihood: A neglected alternative to EM? **International Statistical Review**, v. 82, p. 296–308.
- MCLACHLAN, G. J.; KRISHNAN, T. (1997) **The EM algorithm and extensions**, New York: John Wiley & Sons. 274 p.
- ORTEGA, J. M.; RHEINBOLDT, W. C. (1970) **Iterative solutions of nonlinear equations in several variables**, New York: Academic Press. 572 p.
- NASH, J. C. (2014) On best practice optimization methods in R. **Journal of Statistical Software**, Los Angeles, v. 60, p. 1–14.
- NELDER, J. A.; MEAD, R. (1965) A simplex algorithm for function minimization. **The Computer Journal**, v. 7, p. 308–313.
- SHANNO, D. F. (1970) Conditioning of quasi-Newton methods for function minimization. **Mathematics of Computation**, v. 24, p. 647–656.
- THISTED, R. A. (1988) **Elements of statistical computing: Numerical computation**, New York: Chapman and Hall. 448 p.

Appendix B: Derivations for Chapter 3

Location exponential model

Let X be a positive random variable following the exponential distribution with parameter $\lambda > 0$. The pdf of X is given by

$$f_X(x) = \lambda e^{-\lambda x} \quad (\text{A.4})$$

and its distribution function is

$$F_X(x) = \int_0^x f_X(x) dx = 1 - e^{-\lambda x}. \quad (\text{A.5})$$

Let $Y = \log X$. The distribution of Y can be found by the Jacobian method. We have that

$$y = \log x \Rightarrow x = e^y.$$

Therefore, the modulus of the Jacobian is $|J| = \frac{dx}{dy} = e^y$. Then,

$$f_Y(y) = f_X(e^y)|J| \quad (\text{A.6})$$

$$= \lambda e^{-\lambda e^y} e^y \quad (\text{A.7})$$

$$= \lambda e^{y - \lambda e^y}. \quad (\text{A.8})$$

Now let $\lambda = e^{-\mu}$, $-\infty < \mu < \infty$ is a location parameter, then

$$f_Y(y) = e^{y - \mu - e^{y - \mu}} \quad (\text{A.9})$$

is the pdf of Y . The survival function is

$$S_Y(y) = 1 - F_X(e^y) \quad (\text{A.10})$$

$$= e^{-e^{y - \mu}} \quad (\text{A.11})$$

Generating random samples from the location exponential model

Using the inverse transform theorem, let $F_Y(y) = U \sim U(0, 1)$. Then,

$$1 - e^{-e^{y - \mu}} = U \quad (\text{A.12})$$

$$e^{-e^{y - \mu}} = 1 - U \quad (\text{A.13})$$

$$e^{y - \mu} = -\log(1 - U) \quad (\text{A.14})$$

$$y = \mu + \log[-\log(1 - U)]. \quad (\text{A.15})$$

Therefore, from a random sample of a uniform $(0, 1)$ distribution, if μ is fixed, one can obtain a random sample for the distribution of Y through (A.15).

Location-scale Weibull model

Let X be a positive random variable following the Weibull distribution with parameters $\alpha, \gamma > 0$. The pdf of X is given by

$$f_X(x) = \frac{\gamma}{\alpha^\gamma} x^{\gamma-1} e^{-\left(\frac{x}{\alpha}\right)^\gamma} \quad (\text{A.16})$$

and its distribution function is

$$F_X(x) = \int_0^x f_X(x) dx = 1 - e^{-\left(\frac{x}{\alpha}\right)^\gamma}. \quad (\text{A.17})$$

Let $Y = \log X$. The distribution of Y can be found by the Jacobian method. We have that

$$y = \log x \Rightarrow x = e^y.$$

Therefore, the modulus of the Jacobian is $|J| = \frac{dx}{dy} = e^y$. Then,

$$f_Y(y) = f_X(e^y)|J| \quad (\text{A.18})$$

$$= \frac{\gamma}{\alpha^\gamma} (e^y)^{\gamma-1} e^{-\left(\frac{e^y}{\alpha}\right)^\gamma} e^y. \quad (\text{A.19})$$

Now let $\alpha = e^\mu$, $-\infty < \mu < \infty$ is a location parameter, and $\gamma = \frac{1}{\sigma}$, $\sigma > 0$ is a scale parameter, then

$$f_Y(y) = \frac{1}{\sigma} e^{\frac{y-\mu}{\sigma}} e^{-e^{\frac{y-\mu}{\sigma}}} \quad (\text{A.20})$$

is the pdf of Y . The survival function is

$$S_Y(y) = 1 - F_X(e^y) \quad (\text{A.21})$$

$$= e^{-e^{\frac{y-\mu}{\sigma}}} \quad (\text{A.22})$$

Generating random samples from the location-scale Weibull model

Using the inverse transform theorem, let $F_Y(y) = U \sim U(0, 1)$. Then,

$$1 - e^{-e^{\frac{y-\mu}{\sigma}}} = U \quad (\text{A.23})$$

$$e^{-e^{\frac{y-\mu}{\sigma}}} = 1 - U \quad (\text{A.24})$$

$$e^{\frac{y-\mu}{\sigma}} = -\log(1 - U) \quad (\text{A.25})$$

$$y = \mu + \sigma \log[-\log(1 - U)]. \quad (\text{A.26})$$

Therefore, from a random sample of a uniform $(0, 1)$ distribution, if μ and σ are fixed, one can obtain a random sample for the distribution of Y through (A.26).

Location-scale exponentiated-Weibull model

Let X be a positive random variable following the exponentiated-Weibull distribution with parameters $a, \alpha, \gamma > 0$. The pdf of X is given by

$$f_X(x) = a \left[1 - e^{-\left(\frac{x}{\alpha}\right)^\gamma} \right]^{a-1} \frac{\gamma}{\alpha^\gamma} x^{\gamma-1} e^{-\left(\frac{x}{\alpha}\right)^\gamma} \quad (\text{A.27})$$

and its distribution function is

$$F_X(x) = \int_0^x f_X(x) dx = \left[1 - e^{-\left(\frac{x}{\alpha}\right)^\gamma} \right]^a. \quad (\text{A.28})$$

Let $Y = \log X$. The distribution of Y can be found by the Jacobian method. We have that

$$y = \log x \Rightarrow x = e^y.$$

Therefore, the modulus of the Jacobian is $|J| = \frac{dx}{dy} = e^y$. Then,

$$f_Y(y) = f_X(e^y)|J| \quad (\text{A.29})$$

$$= a \left[1 - e^{-\left(\frac{e^y}{\alpha}\right)^\gamma} \right]^{a-1} \frac{\gamma}{\alpha^\gamma} (e^y)^{\gamma-1} e^{-\left(\frac{e^y}{\alpha}\right)^\gamma} e^y. \quad (\text{A.30})$$

Now let $\alpha = e^\mu$, $-\infty < \mu < \infty$ is a location parameter, and $\gamma = \frac{1}{\sigma}$, $\sigma > 0$ is a scale parameter, then

$$f_Y(y) = a \left[1 - e^{-e^{\frac{y-\mu}{\sigma}}} \right]^{a-1} \frac{1}{\sigma} e^{\frac{y-\mu}{\sigma}} e^{-e^{\frac{y-\mu}{\sigma}}} \quad (\text{A.31})$$

is the pdf of Y . The survival function is

$$S_Y(y) = 1 - F_X(e^y) \quad (\text{A.32})$$

$$= 1 - \left[1 - e^{-e^{\frac{y-\mu}{\sigma}}} \right]^a \quad (\text{A.33})$$

Generating random samples from the location-scale exponentiated-Weibull model

Using the inverse transform theorem, let $F_Y(y) = U \sim U(0, 1)$. Then,

$$\left[1 - e^{-e^{\frac{y-\mu}{\sigma}}} \right]^a = U \quad (\text{A.34})$$

$$1 - e^{-e^{\frac{y-\mu}{\sigma}}} = U^{\frac{1}{a}} \quad (\text{A.35})$$

$$e^{\frac{y-\mu}{\sigma}} = -\log(1 - U^{\frac{1}{a}}) \quad (\text{A.36})$$

$$y = \mu + \sigma \log[-\log(1 - U^{\frac{1}{a}})]. \quad (\text{A.37})$$

Therefore, from a random sample of a uniform $(0, 1)$ distribution, if a , μ and σ are fixed, one can obtain a random sample for the distribution of Y through (A.37).

Appendix C: Derivations for Chapter 4 and simulation studies

Dropping the subscripts i and t for easier reading, consider the joint model

$$\begin{aligned} Y_1|N_1 &\sim \text{Bin}(N_1, p_1) \\ N_1 &\sim \text{P}(\lambda_1) \text{ or } N_1 \sim \text{NB}(\lambda_1, \phi_1) \\ Y_2|N_1, N_2 &\sim \text{Bin}(N_2, p_2) \\ N_2|N_1 &\sim \text{P}(\psi + \lambda_2 N_1) \text{ or } N_2|N_1 \sim \text{NB}(\psi + \lambda_2 N_1, \phi_2) \end{aligned}$$

From the binomial, Poisson and negative binomial distributions, the results from the conditional expectations and variances follow directly, assuming observations from different sites are independent:

$$\begin{aligned} \text{E}(Y_1|N_1) &= N_1 p_1 \\ \text{Var}(Y_1|N_1) &= N_1 p_1 (1 - p_1) \\ \text{E}(Y_1^2|N_1) &= N_1 p_1 - N_1 p_1^2 + N_1^2 p_1^2 \\ \text{E}(Y_2|N_1, N_2) &= N_2 p_2 \\ \text{Var}(Y_2|N_1, N_2) &= N_2 p_2 (1 - p_2) \\ \text{E}(Y_2^2|N_1, N_2) &= N_2 p_2 - N_2 p_2^2 + N_2^2 p_2^2 \\ \text{E}(N_1) &= \lambda_1 \\ \text{E}(N_2|N_1) &= \psi + \lambda_2 N_1 \end{aligned}$$

For N_1 and $N_2|N_1$ distributed as Poisson, we have

$$\begin{aligned} \text{Var}(N_1) &= \lambda_1 \\ \text{E}(N_1^2) &= \lambda_1 + \lambda_1^2 \\ \text{Var}(N_2|N_1) &= \psi + \lambda_2 N_1 \\ \text{E}(N_2^2|N_1) &= \psi + \lambda_2 N_1 + (\psi + \lambda_2 N_1)^2 \end{aligned}$$

and for N_1 and $N_2|N_1$ distributed as negative binomial, we have

$$\begin{aligned} \text{Var}(N_1) &= \lambda_1 + \frac{\lambda_1^2}{\phi_1} \\ \text{E}(N_1^2) &= \lambda_1 + \lambda_1^2 + \frac{\lambda_1^2}{\phi_1} \\ \text{Var}(N_2|N_1) &= \psi + \lambda_2 N_1 + \frac{(\psi + \lambda_2 N_1)^2}{\phi_2} \\ \text{E}(N_2^2|N_1) &= \psi + \lambda_2 N_1 + (\psi + \lambda_2 N_1)^2 + \frac{(\psi + \lambda_2 N_1)^2}{\phi_2} \end{aligned}$$

The unconditional variance of N_2 changes for every case, yielding the following results:

$$\begin{aligned} \text{Var}(N_2) &= \text{E}_{N_1}[\text{Var}_{N_2}(N_2|N_1)] + \text{Var}_{N_1}[\text{E}_{N_2}(N_2|N_1)] \\ &= \begin{cases} \text{E}_{N_1}(\psi + \lambda_2 N_1) + \text{Var}_{N_1}(\psi + \lambda_2 N_1), & \text{for the P-P and NB-P cases} \\ \text{E}_{N_1}\left[\psi + \lambda_2 N_1 + \frac{(\psi + \lambda_2 N_1)^2}{\phi_2}\right] + \text{Var}_{N_1}(\psi + \lambda_2 N_1), & \text{for the P-NB and NB-NB cases} \end{cases} \\ &= \begin{cases} \psi + \lambda_1 \lambda_2 + \lambda_1 \lambda_2^2, & \text{for the P-P case} \\ \psi + \lambda_1 \lambda_2 + \lambda_1 \lambda_2^2 + \frac{\lambda_1^2 \lambda_2^2}{\phi_1}, & \text{for the NB-P case} \\ \psi + \lambda_1 \lambda_2 + \lambda_1 \lambda_2^2 + \frac{\psi^2 + 2\psi \lambda_1 \lambda_2 + \lambda_1 \lambda_2^2 + \lambda_1^2 \lambda_2^2}{\phi_2}, & \text{for the P-NB case} \\ \psi + \lambda_1 \lambda_2 + \lambda_1 \lambda_2^2 + \frac{\psi^2 + 2\psi \lambda_1 \lambda_2 + \lambda_1 \lambda_2^2 + \lambda_1^2 \lambda_2^2}{\phi_2} + \frac{\lambda_1^2 \lambda_2^2}{\phi_1} + \frac{\lambda_1^2 \lambda_2^2}{\phi_1 \phi_2}, & \text{for the NB-NB case} \end{cases} \end{aligned}$$

From theorems of conditional expectation and variance, we may derive the following unconditional expectations of Y_1 and Y_2 :

$$\begin{aligned}
 E(Y_1) &= E_{N_1}[E_{Y_1}(Y_1|N_1)] \\
 &= E_{N_1}(N_1 p_1) = \lambda_1 p_1 \\
 E(Y_2) &= E_{N_1, N_2}[E_{Y_2}(Y_2|N_1, N_2)] \\
 &= E_{N_1, N_2}(N_2 p_2) \\
 &= p_2 E_{N_1}(\psi + \lambda_2 N_1) = p_2(\psi + \lambda_1 \lambda_2)
 \end{aligned}$$

To calculate the covariance between Y_1 and Y_2 , we obtain

$$\begin{aligned}
 E(Y_1 Y_2) &= E_{N_1, N_2}[E_{Y_1, Y_2}(Y_1 Y_2|N_1, N_2)] \\
 &= E_{N_1, N_2}[N_1 p_1 N_2 p_2] \\
 &= E_{N_1}[N_1 p_1 p_2(\psi + \lambda_2 N_1)] \\
 &= \psi \lambda_1 p_1 p_2 + \lambda_2 p_1 p_2 E_{N_1}(N_1^2)
 \end{aligned}$$

and then derive

$$\begin{aligned}
 \text{Cov}(Y_1, Y_2) &= E(Y_1 Y_2) - E(Y_1)E(Y_2) \\
 &= \begin{cases} \lambda_1 \lambda_2 p_1 p_2, & \text{for the P-P and P-NB cases} \\ \lambda_1 \lambda_2 p_1 p_2 \left(1 + \frac{\lambda_1}{\phi_1}\right), & \text{for the NB-P and NB-NB cases} \end{cases}
 \end{aligned}$$

Now we obtain the unconditional variance of Y_1 using theorems of conditional variance:

$$\begin{aligned}
 \text{Var}(Y_1) &= E_{N_1}[\text{Var}_{Y_1}(Y_1|N_1)] + \text{Var}_{N_1}[E_{Y_1}(Y_1|N_1)] \\
 &= E_{N_1}[N_1 p_1(1 - p_1)] + \text{Var}_{N_1}(N_1 p_1) \\
 &= \begin{cases} \lambda_1 p_1, & \text{for the P-P and P-NB cases} \\ \lambda_1 p_1 \left(1 + \frac{\lambda_1 p_1}{\phi_1}\right), & \text{for the NB-P and NB-NB cases} \end{cases}
 \end{aligned}$$

To derive the unconditional variance of Y_2 it is easier to work from

$$\text{Var}(Y_2) = E(Y_2^2) - E(Y_2)^2.$$

We have

$$\begin{aligned}
 E(Y_2^2) &= E_{N_1, N_2}[E_{Y_2}(Y_2^2|N_1, N_2)] \\
 &= E_{N_1, N_2}[N_2 p_2 - N_2 p_2^2 + N_2^2 p_2^2] \\
 &= E_{N_1}[p_2(\psi + \lambda_2 N_1) - p_2^2(\psi + \lambda_2 N_1) + p_2^2 E_{N_2}(N_2^2|N_1)] \\
 &= p_2(\psi + \lambda_1 \lambda_2) - p_2^2(\psi + \lambda_1 \lambda_2) + p_2^2 E_{N_1}[E_{N_2}(N_2^2|N_1)] \\
 &= \begin{cases} p_2(\psi + \lambda_1 \lambda_2) + p_2^2[\lambda_1 \lambda_2^2 + (\psi + \lambda_1 \lambda_2)^2], & \text{for the P-P case} \\ p_2(\psi + \lambda_1 \lambda_2) + p_2^2 \left[\lambda_1 \lambda_2^2 + (\psi + \lambda_1 \lambda_2)^2 + \frac{\lambda_1^2 \lambda_2^2}{\phi_1} \right], & \text{for the NB-P case} \\ p_2(\psi + \lambda_1 \lambda_2) + p_2^2 \left[\lambda_1 \lambda_2^2 + (\psi + \lambda_1 \lambda_2)^2 + \frac{\lambda_1 \lambda_2^2 + (\psi + \lambda_1 \lambda_2)^2}{\phi_2} \right], & \text{for the P-NB case} \\ p_2(\psi + \lambda_1 \lambda_2) + p_2^2 \left[\lambda_1 \lambda_2^2 + (\psi + \lambda_1 \lambda_2)^2 + \frac{\lambda_1 \lambda_2^2 + (\psi + \lambda_1 \lambda_2)^2}{\phi_2} + \frac{\lambda_1^2 \lambda_2^2}{\phi_1} + \frac{\lambda_1^2 \lambda_2^2}{\phi_1 \phi_2} \right], & \text{for the NB-NB case} \end{cases}
 \end{aligned}$$

Then, we have

$$\text{Var}(Y_2) = \begin{cases} p_2[\psi + \lambda_1\lambda_2(1 + \lambda_2p_2)], & \text{for the P-P case} \\ p_2 \left[\psi + \lambda_1\lambda_2 \left(1 + \lambda_2p_2 + \frac{\lambda_1\lambda_2p_2}{\phi_1} \right) \right], & \text{for the NB-P case} \\ p_2 \left\{ \psi + \lambda_1\lambda_2(1 + \lambda_2p_2) + \frac{p_2}{\phi_2} [\lambda_1\lambda_2^2 + (\psi + \lambda_1\lambda_2)^2] \right\}, & \text{for the P-NB case} \\ p_2 \left\{ \psi + \lambda_1\lambda_2 \left(1 + \lambda_2p_2 + \frac{\lambda_1\lambda_2p_2}{\phi_1} + \frac{\lambda_1\lambda_2p_2}{\phi_1\phi_2} \right) + \frac{p_2}{\phi_2} [\lambda_1\lambda_2^2 + (\psi + \lambda_1\lambda_2)^2] \right\}, & \text{for the NB-NB case} \end{cases}$$

Finally, the correlation between Y_1 and Y_2 is derived using

$$\rho(Y_1, Y_2) = \frac{\text{Cov}(Y_1, Y_2)}{\sqrt{\text{Var}(Y_1)\text{Var}(Y_2)}}$$

and the results are presented in Table 4.1.

Now we present all simulation studies carried out with the proposed model. We investigated model performance for the P-P case under 14 different scenarios, varying parameter values and sample size, from 1000 simulations. We obtained the root mean square error (RMSE) as a measure of estimator quality (see Tables A.1 and A.2). We also studied the behaviour of the NB-NB case under three scenarios and because of computational burden we only used 200 simulations. For this case, some datasets seemed less overdispersed and hence the maximum likelihood estimate (MLE) of the dispersion parameter was infinite. So, we also checked the mean, median and RMSE of the estimates excluding the cases when $\hat{\phi}_1$ was larger than 10 (Table A.2).

It is important to note that when the true parameter values were on the boundary of the parametric space, bias was larger but when looking at the median MLEs it was reduced. Further investigation provided evidence that the bias in the mean MLEs is due to a group of really large ψ estimates associated with low p_2 estimates. Moreover, as expected, when sample size increased the mean and median MLEs were closer to the true value and the RMSE was smaller, and hence our proposed model is able to recover the true parameter values of simulated datasets.

Table A.1. Estimator mean, median and RMSE for the proposed joint model under 14 scenarios, which varied true parameter value and number of sites (R) and sampling occasions (T), considering the latent abundances distributed as Poisson for both species (P–P case). Results are based on 1000 simulations

P–P with ψ , $(R, T) = (10, 10)$												
Parameter	Scenario 1				Scenario 2				Scenario 3			
	True	Mean	Median	RMSE	True	Mean	Median	RMSE	True	Mean	Median	RMSE
p_1	0.60	0.60	0.60	0.05	0.60	0.60	0.60	0.05	0.60	0.60	0.60	0.05
p_2	0.20	0.20	0.20	0.09	0.20	0.18	0.19	0.09	0.20	0.18	0.18	0.10
λ_1	8.00	8.02	7.94	1.15	8.00	8.12	8.02	1.17	8.00	8.06	7.98	1.19
ψ	0.00	3.22	1.21	13.54	2.00	4.85	1.44	17.81	2.00	6.02	2.39	17.35
λ_2	0.25	0.17	0.04	0.37	0.00	0.23	0.03	0.78	0.25	0.55	0.14	1.34
P–P with ψ , $(R, T) = (100, 2)$												
Parameter	Scenario 4				Scenario 5				Scenario 6			
	True	Mean	Median	RMSE	True	Mean	Median	RMSE	True	Mean	Median	RMSE
p_1	0.60	0.60	0.60	0.06	0.60	0.60	0.60	0.06	0.60	0.60	0.60	0.06
p_2	0.20	0.21	0.21	0.10	0.20	0.18	0.18	0.11	0.20	0.18	0.18	0.10
λ_1	8.00	8.09	8.01	0.85	8.00	8.08	7.96	0.86	8.00	8.07	7.95	0.90
ψ	0.00	2.84	0.74	12.48	2.00	7.17	1.29	21.84	2.00	5.73	2.31	15.90
λ_2	0.25	0.26	0.13	0.67	0.00	0.24	0.10	0.66	0.25	0.62	0.23	1.30
P–P with ψ , $(R, T) = (100, 10)$												
Parameter	Scenario 7				Scenario 8				Scenario 9			
	True	Mean	Median	RMSE	True	Mean	Median	RMSE	True	Mean	Median	RMSE
p_1	0.60	0.60	0.60	0.02	0.60	0.60	0.60	0.02	0.60	0.60	0.60	0.02
p_2	0.20	0.21	0.21	0.03	0.20	0.19	0.19	0.03	0.20	0.20	0.20	0.03
λ_1	8.00	7.99	8.00	0.34	8.00	8.01	8.01	0.34	8.00	8.00	7.99	0.35
ψ	0.00	0.61	0.26	0.90	2.00	1.26	1.54	1.02	2.00	2.15	2.23	1.28
λ_2	0.25	0.17	0.20	0.12	0.00	0.10	0.06	0.14	0.25	0.25	0.22	0.20
P–P with ψ , $(R, T) = (100, 10)$												
Parameter	Scenario 10				Scenario 11							
	True	Mean	Median	RMSE	True	Mean	Median	RMSE				
p_1	0.60	0.60	0.60	0.02	0.60	0.60	0.60	0.02				
p_2	0.20	0.21	0.21	0.03	0.20	0.20	0.20	0.03				
λ_1	8.00	7.99	7.98	0.35	8.00	8.01	7.99	0.33				
ψ	0.00	0.63	0.21	1.08	2.00	2.34	2.15	1.62				
λ_2	0.80	0.71	0.73	0.21	0.80	0.79	0.77	0.33				
P–P without ψ												
Parameter	Scenario 12 – $(R, T) = (10, 10)$				Scenario 13 – $(R, T) = (100, 2)$				Scenario 14 – $(R, T) = (100, 10)$			
	True	Mean	Median	RMSE	True	Mean	Median	RMSE	True	Mean	Median	RMSE
p_1	0.60	0.60	0.60	0.05	0.60	0.60	0.61	0.06	0.60	0.60	0.60	0.02
p_2	0.20	0.19	0.19	0.06	0.20	0.19	0.19	0.11	0.20	0.20	0.20	0.03
λ_1	8.00	8.11	8.02	1.15	8.00	8.09	7.95	0.87	8.00	7.99	7.99	0.36
λ_2	0.25	0.65	0.26	1.53	0.25	0.66	0.26	1.37	0.25	0.25	0.25	0.04

P, Poisson; NB, negative binomial; RMSE, root mean square error.

Table A.2. Estimator mean, median and RMSE for the proposed joint model under 3 scenarios, which varied true parameter value and number of sites (R) and sampling occasions (T), considering the latent abundances distributed as negative binomial for both species (NB–NB case). Results are based on 200 simulations, except where indicated otherwise

NB–NB with ψ , $(R, T) = (100, 10)$												
Parameter	Scenario 1				Scenario 2				Scenario 3			
	True	Mean	Median	RMSE	True	Mean	Median	RMSE	True	Mean	Median	RMSE
p_1	0.60	0.60	0.60	0.01	0.60	0.60	0.60	0.01	0.60	0.60	0.60	0.01
p_2	0.20	0.20	0.20	0.02	0.20	0.20	0.20	0.01	0.20	0.20	0.20	0.01
λ_1	8.00	7.94	7.87	0.32	8.00	8.05	8.05	0.31	8.00	8.01	8.01	0.28
ψ	0.00	2.11	2.04	0.96	2.00	2.08	2.06	0.19	2.00	2.11	2.05	0.19
λ_2	0.25	0.17	0.20	0.06	0.00	0.10	0.06	0.07	0.25	0.25	0.22	0.11
ϕ_1	2.00	3.24	0.75	15.19	2.00	49.56	1.48	207.49	2.00	173.40	1.70	569.34
ϕ_2	1.00	0.84	0.36	0.40	1.00	1.04	0.70	0.36	1.00	0.68	0.14	0.60
NB–NB with ψ , $(R, T) = (100, 10)$, excluding cases where $\hat{\phi}_1 > 10$												
Parameter	Scenario 1 (199 simulations)				Scenario 2 (188 simulations)				Scenario 3 (192 simulations)			
	True	Mean	Median	RMSE	True	Mean	Median	RMSE	True	Mean	Median	RMSE
p_1	0.60	0.60	0.60	0.01	0.60	0.60	0.60	0.02	0.60	0.60	0.60	0.02
p_2	0.20	0.20	0.20	0.04	0.20	0.20	0.20	0.03	0.20	0.20	0.20	0.03
λ_1	8.00	7.94	7.88	0.72	8.00	8.06	8.03	0.68	8.00	7.99	7.95	0.62
ψ	0.00	2.11	2.04	2.15	2.00	2.11	2.09	0.42	2.00	2.12	2.08	0.44
λ_2	0.25	0.16	0.19	0.14	0.00	0.11	0.10	0.15	0.25	0.44	0.47	0.25
ϕ_1	2.00	0.83	0.75	1.22	2.00	1.72	1.43	1.17	2.00	2.03	1.66	1.18
ϕ_2	1.00	0.85	0.37	0.89	1.00	1.10	0.87	0.80	1.00	0.71	0.17	1.36

P, Poisson; NB, negative binomial; RMSE, root mean square error.

Appendix D: Review on techniques used on Chapter 6

One and two-dimensional kernel density estimation

A useful tool to estimate the empirical density function of a group of observations is the kernel density estimator. Considering that $x_1, \dots, x_n$ are a random sample from an unknown probability model, we can estimate the shape of its density function through

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right),$$

where $K(\cdot)$ is called the kernel function, with $K(x) \geq 0, \forall x \in \mathbb{R}$, and $\int_{\mathbb{R}} K(x)dx = 1$, and h is the smoothing parameter, called bandwidth, which scales the estimated function. Common choices of $K(x)$ include the Gaussian probability density function, uniform, triangular, biweight, Epachenikov, among other functions.

In this work, we used the Gaussian kernel, resulting in

$$\hat{f}_h(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^n \exp\left\{-\frac{(x - x_i)^2}{2h^2}\right\}.$$

We used the rule of thumb for choosing the bandwidth as proposed by Scott (1992), which is

$$h = \left(\frac{4}{3n}\right)^{1/5} \times s, \quad s = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}}$$

is the standard deviation of the sample. This is implemented as the **density** function in R, which includes many other types of kernel functions and rules for choosing h .

This can be easily extended to bivariate samples $(x_{11}, x_{12}), \dots, (x_{n1}, x_{n2})$. Still considering a Gaussian kernel, the estimated density of a bivariate sample can be written as

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{2\pi n \sqrt{|\mathbf{H}|}} \sum_{i=1}^n \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{x}_i)^\top \mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_i)\right\}.$$

Again, using Scott's (1992) rule of thumb, the bandwidth matrix $\mathbf{H} = \text{diag}\{h_1, h_2\}$, with

$$h_j = \frac{s_j^2}{n^{1/3}}, \quad s_j^2 = \sum_{i=1}^n \frac{(x_{ij} - \bar{x}_j)^2}{n-1}$$

is the estimated variance of the j -th variable, $j = 1, 2$. This is implemented as the **kde2d** function in package **MASS** (Venables and Ripley, 2002), which also includes other rules for choosing the bandwidth matrix $\mathbf{H}$.

Convex polygon geometry

Let $P_1, P_2, \dots, P_n$ be the n vertices of the convex polygon $\mathbf{P} = \overline{P_1 P_2 \dots P_n}$, with $P_i = (x_i, y_i)$. By the definition of a convex polygon, the angles in each vertex do not exceed 180° . Consider the quantity

$$A = \frac{1}{2} \left\{ \sum_{i=1}^{n-1} (x_i y_{i+1} - x_{i+1} y_i) + (x_n y_1 - x_1 y_n) \right\}.$$

The area of polygon $\mathbf{P}$ may be written as $A_{\mathbf{P}} = |A|$ and its centre of mass is $C_{\mathbf{P}} = (x_C, y_C)$, with

$$\begin{aligned} x_C &= \frac{1}{6A} \left\{ \sum_{i=1}^{n-1} (x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i) + (x_n + x_1)(x_n y_1 - x_1 y_n) \right\} \\ y_C &= \frac{1}{6A} \left\{ \sum_{i=1}^{n-1} (y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i) + (y_n + y_1)(x_n y_1 - x_1 y_n) \right\} \end{aligned}$$

To determine whether a point $P_0 = (x_0, y_0)$ is inside the area defined by polygon $\mathbf{P}$, one strategy may be devised as follows:

- 1 if $P_0 = P_i$ for some i , P_0 coincides with one of the vertices of polygon $\mathbf{P}$ and is considered to be outside;
- 2 if $P_0 \in \overline{P_i P_{i+1}}$, $x_i < x_{i+1}$, then P_0 lies in the border of polygon $\mathbf{P}$ and is considered to be outside. A simple way to test this is to compare the angular coefficients of the segments $\overline{P_0 P_i}$ and $\overline{P_0 P_{i+1}}$, i.e., test whether

$$\frac{y_0 - y_i}{x_0 - x_i} = \frac{y_{i+1} - y_0}{x_{i+1} - x_0}.$$

If this equality holds, and $x_i < x_0 < x_{i+1}$, then the point P_0 belongs to a side of polygon $\mathbf{P}$;

- 3 if P_0 is not a vertex of the polygon nor does it belong to a side of $\mathbf{P}$, then a simple test may be applied, which will be explained below.

Firstly, if $x_0 < \min_i \{x_i\}$ or $x_0 > \max_i \{x_i\}$, the point P_0 is outside polygon $\mathbf{P}$. Analogously, the same conclusion applies if $y_0 < \min_i \{y_i\}$ or $y_0 > \max_i \{y_i\}$. When $\min_i \{x_i\} \leq x_0 \leq \max_i \{x_i\}$ and $\min_i \{y_i\} \leq y_0 \leq \max_i \{y_i\}$, we use the fact that for any convex polygon, if the point P_0 is inside its area, the ray $\vec{P_0}$, parallel to the x -axis, intersects a side of the polygon only once. If, however, the point is outside the polygon, the ray $\vec{P_0}$ does so twice or does not intersect any side (see Figure A.3). Using the line equation, it is then possible to find the x -coordinate x_i^+ of the intersection points of the ray $\vec{P_0}$ with all segments $\overline{P_i P_{i+1}}$ of the polygon through

$$x^+ = x_i + \frac{(x_{i+1} - x_i)(y_0 - y_i)}{y_{i+1} - y_i}.$$

Then, if the condition $x_0 > x_i^+$ is satisfied, an intersection is counted if $x_i \leq x_i^+ \leq x_{i+1}$ (if $x_{i+1} > x_i$) or $x_{i+1} \leq x_i^+ \leq x_i$ (if $x_i > x_{i+1}$). This method can fail if the ray $\vec{P_0}$ intersects a vertex, as the intersection would be counted twice. To overcome this, after the number of intersections has been counted, if $y_0 = y_i$, with $x_0 < x_i$, then 1 is subtracted from the total number of computed intersections.

There are other strategies to find out whether a point is inside a convex or concave polygon. This is referred to as the “point in polygon” problem (Haines, 1994).

We will now describe a way to obtain a second convex polygon, say $\mathbf{P}' = \overline{P'_1 \dots P'_n}$, with $P'_i = (x'_i, y'_i)$ such that it is inside $\mathbf{P}$ (see Figure A.4(a)) and its area $A_{\mathbf{P}'}$ is equal to a known proportion α of the area $A_{\mathbf{P}}$ of polygon $\mathbf{P}$, i.e.

$$A_{\mathbf{P}'} = \alpha A_{\mathbf{P}}, \quad 0 < \alpha < 1. \quad (\text{A.38})$$

As described above, let $C_{\mathbf{P}} = (x_C, y_C)$ be the centre of mass of polygon $\mathbf{P}$. Let $d_i = \sqrt{(x_i - x_C)^2 - (y_i - y_C)^2}$ be the length of segment $\overline{P_i C_{\mathbf{P}}}$. We now want to find a constant k to be subtracted from d_i to form the new lengths $d'_i = d_i - k$ of the segments $\overline{P'_i C_{\mathbf{P}'}}$ in order to obtain the coordinates of $\mathbf{P}'$ (see Figure A.4(b)).

Using rules of congruent triangles (see Figure A.4(b)), the coordinates of polygon $\mathbf{P}'$ may be derived from

$$\frac{x'_i - x_C}{x_i - x_C} = \frac{y'_i - y_C}{y_i - y_C} = \frac{d'_i}{d_i}$$

and hence written as

$$\begin{aligned} x'_i &= \frac{d'_i}{d_i}(x_i - x_C) + x_C = \frac{d_i x_i - k \tilde{x}_i}{d_i} \\ y'_i &= \frac{d'_i}{d_i}(y_i - y_C) + y_C = \frac{d_i y_i - k \tilde{y}_i}{d_i} \end{aligned}$$

Figure A.3. A five-sided convex polygon, delimited areas where any point would be outside its area (gray expressions), and three points with their respective rays parallel to the x -axis, out of which one intersects only one side of the polygon (in blue) and is hence inside its area and the other two present zero or two intersections (in red), thus lying outside the polygon, with crosses indicating the intersections

where $\tilde{x}_i = x_i - x_C$ and $\tilde{y}_i = y_i - y_C$. Now we have to solve equation (A.38) for k , which after algebraic simplification results in a quadratic equation of the form

$$ak^2 + bk + c = 0,$$

with

$$\begin{aligned} a &= \sum_{i=1}^{n-1} \frac{\tilde{x}_i \tilde{y}_{i+1} - \tilde{x}_{i+1} \tilde{y}_i}{d_i d_{i+1}} + \frac{\tilde{x}_n \tilde{y}_1 - \tilde{x}_1 \tilde{y}_n}{d_n d_1} \\ b &= \sum_{i=1}^{n-1} \frac{d_i (\tilde{x}_{i+1} \tilde{y}_i - \tilde{x}_i \tilde{y}_{i+1}) + d_{i+1} (\tilde{x}_{i+1} \tilde{y}_i - \tilde{x}_i \tilde{y}_{i+1})}{d_i d_{i+1}} + \frac{d_n (\tilde{x}_1 \tilde{y}_n - \tilde{x}_n \tilde{y}_1) + d_1 (\tilde{x}_1 \tilde{y}_n - \tilde{x}_n \tilde{y}_1)}{d_n d_1} \\ c &= 2(A \pm \alpha A_P) \end{aligned}$$

Then, the smallest real solution of this equation is the value to be subtracted from each centroid-vertex segment to form a polygon of reduced area. It is important to note that if α is set too low and if there are vertices in $\mathbf{P}$ with interior angles close to 180° , this operation may result in a concave $\mathbf{P}'$. Usually, for $\alpha = 0.95$ and $\alpha = 0.99$, the most common values used, the resulting polygon is convex.

Convex hull peeling

Let $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be a set of points in the Euclidean plane (see Figure A.5(a)). The convex hull of S , denoted by $\text{Conv}(S)$ is defined as the intersection of all convex sets that contain all

Figure A.4. (a) The reduced polygon $P' = \overline{P_1' \dots P_5'}$ inside polygon $P = \overline{P_1 \dots P_5}$ in such a way that the area of P' is equal to a proportion α of the area of P ; (b) relationships of the two congruent triangles used to obtain the coordinates of P_i'

the elements of S . There are many algorithms for obtaining this. Here we use the one devised by Eddy (1977), which works as follows:

1. find the points with the minimum and maximum x values of the set, which will be part of the convex hull and use the line formed by these two points to split the set (see Figure A.5(b));
2. for each subset find the point with the largest distance from the line segment and form triangles (see Figure A.5(c)), removing the points that lie inside the triangles, as they will not form the convex hull (see Figure A.5(d));
3. for each line segment, repeat this process until there are no more points in the set other than those forming the convex hull (see Figures A.5(d) and A.5(e)).

By successively eliminating the points that belong to $\text{Conv}(S)$ from set S and then obtaining a new convex hull, we are then performing what is referred to as convex hull peeling (Eddy, 1982).

References

- EDDY, W. F. (1977) A new convex hull algorithm for planar sets. **ACM Transactions on Mathematical Software**, v. 3, p. 398–403.
- EDDY, W. F. (1982) Convex hull peeling. **Proceedings of the COMPSTAT 1982 5th Symposium held at Toulouse**, p. 42–47.
- HAINES, E. (1994) Point in polygon strategies. In: HEHCKBERT, P. S. (Ed.) **Graphics gems IV**. AP Professional. 585 p.
- SCOTT, D. W. (1992) **Multivariate Density Estimation: Theory, Practice, and Visualization**. Wiley. 317 p.

Figure A.5. The “quickhull” algorithm as devised by Eddy (1977) to obtain the convex hull for (a) a set of points, beginning by (b) obtaining the points with the minimum and maximum x values, then (c) obtaining a first polygon with the points with minimum and maximum y values and (d) eliminating the points that lie inside that initial polygon. After that, for each line segment, (e) triangles are formed with the point with the largest distance from the segment and eliminating points that lie within the polygon recursively until (f) the convex hull is obtained

VENABLES, W. N.; RIPLEY, B. D. (2002) **Modern applied statistics with S**. Fourth Edition. New York: Springer. 504 p.

Appendix E: R packages from the thesis

The hnp package

The `hnp` package, available on <https://cran.r-project.org/web/packages/hnp/index.html>, has two main functions, called `hnp` and `plot.hnp`. The first identifies the class of the model object and passes to the specific internal function that will simulate the datasets and refit the model, extracting the diagnostics. Prior to producing the plot, another internal function, `.makehnp` is called to prepare the object to be passed to `plot.hnp`, which may also be used directly for a saved `hnp` object. For model classes not implemented in the package, the function `newhnp` is called to produce the envelope, given the user provides three helper functions, to extract model diagnostics, simulate new datasets and refit the model. The package also includes example datasets:

- `cbb`: counts of *Hypothenemus hampei*, a major pest of commercial coffee, obtained using different traps (longitudinal count data)
- `chryso`: mortality of the predator *Chrysoperla externa* on different doses of lime sulphur, a substance used to control pests on trees (proportion data)
- `corn`: corn grain damage by *Sitophilus zeamais*, a major pest of stored maize worldwide (proportion data)
- `fungi`: mortality of the Citrus psyllid, *Diaphorina citri*, a major pest of Citrus worldwide, when exposed to different concentrations of two fungi species, *Beauveria bassiana* and *Isaria fumosorosea* (proportion data)
- `oil`: effects of three agricultural oils on *Diaphorina citri* oviposition (count data)
- `orange`: number of embryos of orange variety *Caipira* produced with different sugars (count data)
- `progeny`: progeny of *Sitophilus zeamais* when treated with different organic extracts of *Annona mucosa*
- `wolbachia`: viability of *Trichogramma galloi*, a parasitoid wasp, when infected with *Wolbachia*, a bacteria known to change its reproductive aspects (proportion data)

As of version 1.2-3, the following model classes are automatically supported:

- **Continuous data:** normal, using `lm`, `aov` and `glm`; gamma and inverse gaussian, using `glm`
- **Proportion data:** binomial and quasi-binomial, using `glm`; beta-binomial, using `VGAM::vglm`, `aods3::aodml`, `gamlss::gamlss` and `glmmadmb::glmmADMB`; zero-inflated binomial, using `VGAM::vglm`, `gamlss::gamlss` and `glmmadmb::glmmADMB`; zero-inflated beta-binomial, using `gamlss::gamlss` and `glmmadmb::glmmADMB`; multinomial, using `nnet::multinom`
- **Count data:** Poisson and quasi-Poisson, using `glm`; negative binomial, using `glm` and `aods3::aodml`; zero-inflated and hurdle Poisson and negative binomial, using `pscl::zeroinfl` and `pscl::hurdle`
- **Mixed models:** normal, Poisson and binomial, using `lme4::lmer` and `lme4::glmer`

The jointNmix package

The `jointNmix` package, available on <https://cran.r-project.org/web/packages/jointNmix/index.html> includes functions to fit univariate N-mixture models (`Nmix`) with Poisson, negative binomial and Neyman-A distributions for the latent abundances, bivariate N-mixture models (`jointNmix`), as described on Chapter 4. Several methods are available for fitted model objects:

- calculating the Akaike Information Criterion for a fitted model (`AIC.Nmix`)
- likelihood-ratio test between two fitted model objects (`anova.Nmix`)
- extracting model coefficients (`coef.Nmix`)
- extracting the log-likelihood for a fitted model (`logLik.Nmix`)
- calculating fitted values (`fitted.uniNmix` and `fitted.jointNmix`)
- computing the correlation and covariance for the joint N-mixture model (`getcorr.jointNmix` and `getcov.jointNmix`)
- plotting the posterior distributions of the latent abundances (`plot.uniNmix` and `plot.jointNmix`)
- computing ordinary and standardised residuals (`residuals.uniNmix` and `residuals.jointNmix`)
- simulating new samples from a fitted model (`simulate.uniNmix` and `simulate.jointNmix`)

The bivrp package

The `bivrp` package, available on <https://cran.r-project.org/web/packages/bivrp/index.html> has two main functions, `bivrp` and `plot.bivrp`, used to produce the bivariate residual plot with simulation polygons. The first function performs all simulations and model refitting procedures to compute the information needed to produce the plot, while the second uses this information to obtain the reduced-area polygons and plot them. It calls eight internal functions, used to add the density plots in the margins (`add.dplots.prep` and `add.dplots.plot`), obtain the convex hulls (`chp.perpoint`), sort the points (`sorttheta`), and do polygon-related operations:

- `polygon.area`: returns the area and coordinates for the centre of mass of a convex polygon
- `is.point.inside`: returns whether a point is inside or outside a convex polygon
- `get.k`: calculates k to reduce the area of a convex polygon using the method described in Appendix D (Convex polygon geometry)
- `get.newpolygon`: obtain the coordinates of the reduced-area polygon, given k

SUPPLEMENTARY MATERIAL

Supplement S1: Dataset for Chapter 3

Table S.1. Dataset used in Chapter 3. “E” represents *E. annulipes*, “P” represents *P. nigrispinus*, “Par” represents the parasitised larva and “NPar” represents the non-parasitised larva

(continues)							
Day	Block	Species	Gender	Time until attack (seconds)	First attack	Effective consumption	
1	1	E	M	75	NPar	NPar	
1	1	E	F	58	NPar	NPar	
1	1	P	M	392	NPar	NPar	
1	1	P	F	664	Par	NPar	
1	2	E	M	220	NPar	NPar	
1	2	E	F	175	NPar	NPar	
1	2	P	M	42	Par	Par	
1	2	P	F	1098	NPar	NPar	
1	3	E	M	250	Par	NPar	
1	3	E	F	100	NPar	NPar	
1	3	P	M	910	NPar	NPar	
1	3	P	F	200	Par	Par	
1	4	E	M	64	NPar	NPar	
1	4	E	F	60	NPar	Par	
1	4	P	M	263	NPar	NPar	
1	4	P	F	80	NPar	NPar	
1	5	E	M	834	Par		
1	5	E	F	80	Par	Par	
1	5	P	M	470	NPar	NPar	
1	5	P	F				
1	6	E	M	140	NPar	NPar	
1	6	E	F	30	Par	Par	
1	6	P	M	345	NPar	NPar	
1	6	P	F	300	NPar	NPar	
1	7	E	M	370	NPar	NPar	
1	7	E	F	55	Par	NPar	
1	7	P	M	365	NPar		
1	7	P	F	140	NPar	Par	
1	8	E	M	230	Par	Par	
1	8	E	F	50	Par	Par	
1	8	P	M	156	Par	Par	
1	8	P	F	1202	NPar	NPar	
1	9	E	M	100	Par	Par	
1	9	E	F	95	NPar	NPar	
1	9	P	M	200	Par	NPar	
1	9	P	F	20	Par	Par	
1	10	E	M	60	Par	NPar	
1	10	E	F	10	NPar	NPar	
1	10	P	M	1650	NPar	NPar	
1	10	P	F				
1	11	E	M	366	NPar		
1	11	E	F	120	NPar	NPar	
1	11	P	M	140	NPar	NPar	
1	11	P	F	1625	NPar	NPar	

Table S.1. Dataset used in Chapter 3. “E” represents *E. annulipes*, “P” represents *P. nigrispinus*, “Par” represents the parasitised larva and “NPar” represents the non-parasitised larva

(continues)						
Day	Block	Species	Gender	Time until attack (seconds)	First attack	Effective consumption
2	1	E	M	38	Par	
2	1	E	F	15	NPar	Par
2	1	P	M	102	NPar	NPar
2	1	P	F	2192	NPar	NPar
2	2	E	M	61	NPar	NPar
2	2	E	F	79	Par	Par
2	2	P	M	46	Par	NPar
2	2	P	F	86	NPar	NPar
2	3	E	M	97	NPar	NPar
2	3	E	F	89	Par	Par
2	3	P	M	602	NPar	Par
2	3	P	F	826	NPar	NPar
2	4	E	M	560	NPar	NPar
2	4	E	F	77	NPar	NPar
2	4	P	M	142	NPar	NPar
2	4	P	F	295	Par	Par
2	5	E	M	126	NPar	NPar
2	5	E	F	171	Par	NPar
2	5	P	M	194	NPar	NPar
2	5	P	F	403	NPar	NPar
2	6	E	M	994	Par	
2	6	E	F	116	NPar	Par
2	6	P	M	500	NPar	NPar
2	6	P	F	300	NPar	NPar
2	7	E	M	96	Par	Par
2	7	E	F	146	Par	Par
2	7	P	M	634	NPar	NPar
2	7	P	F	355	Par	Par
2	8	E	M	150	Par	Par
2	8	E	F	85	Par	Par
2	8	P	M	110	NPar	NPar
2	8	P	F	275	NPar	NPar
2	9	E	M	15	Par	NPar
2	9	E	F	81	NPar	NPar
2	9	P	M	280	NPar	NPar
2	9	P	F	998	NPar	NPar
2	10	E	M	110	Par	NPar
2	10	E	F	80	NPar	NPar
2	10	P	M			
2	10	P	F	100	Par	Par
2	11	E	M	31	NPar	NPar
2	11	E	F	479	NPar	NPar
2	11	P	M	22	NPar	NPar
2	11	P	F	25	Par	Par

Table S.1. Dataset used in Chapter 2. “E” represents *E. annulipes*, “P” represents *P. nigrispinus*, “Par” represents the parasitised larva and “NPar” represents the non-parasitised larva

							(continued)
Day	Block	Species	Gender	Time until attack (seconds)	First attack	Effective consumption	
3	1	E	M	33	Par	NPar	
3	1	E	F	411	Par	Par	
3	1	P	M	166	NPar	NPar	
3	1	P	F	365	Par	Par	
3	2	E	M	522	NPar		
3	2	E	F	72	NPar	NPar	
3	2	P	M	200	NPar		
3	2	P	F	923	NPar		
3	3	E	M	1240	NPar		
3	3	E	F	220	NPar		
3	3	P	M	215	NPar	NPar	
3	3	P	F	132	Par	NPar	
3	4	E	M	29	Par	NPar	
3	4	E	F	90	NPar	NPar	
3	4	P	M	50	Par	Par	
3	4	P	F	127	NPar	NPar	
3	5	E	M	38	Par	Par	
3	5	E	F	108	NPar	NPar	
3	5	P	M	156	NPar	NPar	
3	5	P	F	79	Par	Par	
3	6	E	M	430	NPar		
3	6	E	F	152	Par	NPar	
3	6	P	M	719	NPar	NPar	
3	6	P	F	991	Par	Par	
3	7	E	M	92	NPar	Par	
3	7	E	F	86	NPar		
3	7	P	M	282	NPar	NPar	
3	7	P	F	166	Par	Par	
3	8	E	M	101	Par	NPar	
3	8	E	F	21	NPar	Par	
3	8	P	M	478	Par	NPar	
3	8	P	F	288	NPar	NPar	
3	9	E	M	114	NPar		
3	9	E	F	72	NPar	NPar	
3	9	P	M	929	Par	Par	
3	9	P	F	46	Par	Par	
3	10	E	M	108	NPar	NPar	
3	10	E	F	125	Par	Par	
3	10	P	M	148	NPar	NPar	
3	10	P	F	2410	NPar	NPar	
3	11	E	M	13	NPar		
3	11	E	F	14	NPar		
3	11	P	M	1215	Par	Par	
3	11	P	F	910	Par	Par	

Supplement S2: Datasets for Chapter 4**Table S.2.** Mite and springtail dataset used in Chapter 4. Cell numbers represent counts over six time occasions and at 36 sites

Site	Ecosystem	Area	Springtail						Mite					
			1	2	3	4	5	6	1	2	3	4	5	6
1	Patch	Exterior	2	1	1	0	1	11	12	1	1	3	0	5
2	Patch	Interior	3	0	77	2	17	0	24	0	27	2	1	1
3	Pasture	Exterior	0	5	1	0	0	0	0	1	0	0	1	0
4	Pasture	Interior	1	4	4	0	0	2	0	4	6	1	1	1
5	Patch	Exterior	0	1	0	0	1	0	1	1	2	4	11	3
6	Patch	Interior	0	1	4	2	0	27	6	7	3	5	0	5
7	Pasture	Exterior	0	2	0	0	0	0	0	3	4	0	0	0
8	Pasture	Interior	0	3	1	1	0	0	2	2	0	3	2	1
9	Patch	Exterior	1	0	0	2	2	16	9	2	1	3	0	1
10	Patch	Interior	0	0	0	1	0	0	3	0	0	5	1	6
11	Pasture	Exterior	0	0	0	0	0	1	0	2	1	0	0	0
12	Pasture	Interior	1	0	0	0	0	0	1	0	0	0	0	1
13	Patch	Exterior	1	0	0	0	5	0	4	0	1	0	3	1
14	Patch	Interior	3	5	0	0	0	0	19	18	0	7	0	4
15	Pasture	Exterior	8	2	7	1	0	0	0	0	0	5	0	0
16	Pasture	Interior	1	0	3	0	0	0	2	0	3	0	3	0
17	Patch	Exterior	0	0	5	0	0	2	10	1	3	0	1	2
18	Patch	Interior	0	0	2	0	0	0	21	2	15	2	1	21
19	Pasture	Exterior	0	1	2	1	0	0	0	4	1	1	0	2
20	Pasture	Interior	0	0	3	0	0	0	2	0	3	7	0	2
21	Patch	Exterior	1	1	2	0	0	0	21	5	7	3	9	14
22	Patch	Interior	18	8	0	11	0	0	91	1	2	2	0	10
23	Pasture	Exterior	0	0	0	0	0	0	0	0	1	0	2	0
24	Pasture	Interior	0	0	0	0	0	0	0	2	1	1	0	0
25	Patch	Exterior	0	1	0	0	0	0	8	4	1	1	0	0
26	Patch	Interior	3	1	0	2	20	0	16	3	0	10	28	2
27	Pasture	Exterior	0	2	2	0	0	0	1	1	3	1	3	0
28	Pasture	Interior	1	2	145	0	0	0	1	1	3	2	0	0
29	Patch	Exterior	7	0	19	20	0	3	4	9	1	9	11	6
30	Patch	Interior	3	0	6	1	3	2	6	0	4	0	0	3
31	Pasture	Exterior	1	1	1	1	0	0	1	0	1	0	1	0
32	Pasture	Interior	0	1	16	8	2	1	0	0	3	1	0	1
33	Patch	Exterior	1	2	3	0	2	7	4	3	7	3	6	0
34	Patch	Interior	2	2	3	2	0	1	12	0	5	3	1	9
35	Pasture	Exterior	2	1	0	0	0	1	0	0	0	0	0	0
36	Pasture	Interior	0	0	0	0	0	0	1	0	1	0	0	0

Table S.3. Mite and springtail dataset used in Chapter 4. Cell numbers represent relative humidity (%) and temperature (°C) over six time occasions and at 36 sites

Site	Ecosystem	Area	Relative humidity (%)						Temperature (°C)					
			1	2	3	4	5	6	1	2	3	4	5	6
1	Patch	Exterior	30.36	76.48	75.10	59.30	47.00	50.30	12.00	12.00	10.00	9.50	10.00	9.00
2	Patch	Interior	35.14	75.48	76.60	51.70	57.00	51.60	12.00	11.00	10.00	9.50	11.00	9.00
3	Pasture	Exterior	37.21	73.48	72.90	52.70	51.40	50.50	16.00	15.00	10.00	13.00	14.00	12.00
4	Pasture	Interior	32.74	72.87	71.20	52.50	50.00	51.30	16.00	16.00	10.00	13.00	16.00	13.00
5	Patch	Exterior	11.89	88.95	86.50	57.70	73.40	60.30	10.00	11.00	10.00	9.00	10.00	10.00
6	Patch	Interior	22.89	87.13	75.00	49.90	52.80	50.00	11.00	11.00	10.00	9.00	11.00	10.00
7	Pasture	Exterior	40.54	74.00	74.80	53.00	52.70	52.30	16.00	16.00	12.00	12.00	15.50	13.00
8	Pasture	Interior	42.41	73.69	74.80	52.30	52.90	55.40	16.00	16.00	12.00	12.00	15.00	12.00
9	Patch	Exterior	57.39	72.61	75.60	70.30	55.00	70.50	11.00	11.00	10.00	9.00	11.00	9.00
10	Patch	Interior	54.84	75.66	79.70	50.70	59.20	56.00	12.00	12.00	11.00	10.00	10.00	9.00
11	Pasture	Exterior	42.61	73.58	75.60	51.80	50.30	56.50	16.00	16.00	12.00	12.00	15.00	13.00
12	Pasture	Interior	36.91	74.40	67.50	50.20	50.60	51.50	16.00	16.00	12.00	12.00	15.00	13.00
13	Patch	Exterior	45.56	75.22	74.90	53.20	72.00	50.50	10.00	12.00	10.00	9.50	10.00	10.00
14	Patch	Interior	79.26	73.29	73.50	48.10	46.90	46.10	12.00	12.00	11.00	9.50	12.00	9.00
15	Pasture	Exterior	40.52	75.25	87.90	56.60	59.00	55.50	16.00	16.00	12.00	13.00	16.00	12.00
16	Pasture	Interior	42.83	75.65	77.20	56.30	57.70	57.10	16.00	16.00	12.00	12.00	16.00	13.00
17	Patch	Exterior	49.78	72.94	74.40	49.00	50.30	47.60	11.00	12.00	10.00	9.00	10.00	9.00
18	Patch	Interior	52.94	69.67	72.90	54.90	45.90	53.10	12.00	12.00	10.00	10.00	11.00	9.00
19	Pasture	Exterior	36.02	73.93	74.30	50.80	46.70	45.60	17.00	16.00	15.00	13.00	17.80	12.00
20	Pasture	Interior	31.81	71.64	73.30	49.50	40.40	47.70	17.00	17.00	15.00	13.00	18.00	12.00
21	Patch	Exterior	43.38	82.14	82.40	69.50	54.40	61.50	11.00	11.00	10.00	10.00	11.00	10.00
22	Patch	Interior	69.87	74.58	84.40	48.20	53.70	58.00	12.00	12.00	10.00	10.50	11.00	10.00
23	Pasture	Exterior	32.00	68.59	66.80	40.30	43.00	40.00	16.00	16.00	14.00	11.00	14.00	13.00
24	Pasture	Interior	25.51	75.00	67.50	41.70	43.80	51.79	18.00	16.00	13.00	12.00	15.00	12.50
25	Patch	Exterior	73.68	84.41	74.20	54.00	46.60	51.90	10.00	12.00	10.00	10.00	11.00	9.00
26	Patch	Interior	69.57	78.32	84.70	63.70	50.50	60.20	12.00	12.00	10.00	9.50	10.00	9.00
27	Pasture	Exterior	40.08	72.95	72.50	57.70	56.60	47.20	17.00	17.00	14.00	13.00	15.00	14.00
28	Pasture	Interior	29.88	74.03	74.90	55.40	46.20	45.10	17.00	17.00	14.00	12.00	15.00	13.00
29	Patch	Exterior	45.96	71.19	87.00	79.00	72.80	76.90	11.00	12.00	10.00	10.00	11.00	9.00
30	Patch	Interior	78.76	75.39	88.00	77.10	72.70	78.30	12.00	12.00	10.00	10.00	10.00	9.00
31	Pasture	Exterior	43.16	69.71	71.40	51.20	48.40	55.70	16.00	16.00	14.00	12.00	16.00	12.00
32	Pasture	Interior	40.57	74.86	75.90	52.80	57.50	49.90	16.00	16.00	14.00	12.00	17.00	11.00
33	Patch	Exterior	71.54	81.63	79.10	60.80	59.30	62.70	11.00	12.00	10.00	9.00	11.00	9.00
34	Patch	Interior	48.25	75.97	76.70	56.60	54.30	56.90	11.00	12.00	11.00	9.00	10.00	9.00
35	Pasture	Exterior	31.68	73.13	76.00	55.20	48.00	51.50	16.00	16.00	14.00	12.00	16.00	12.00
36	Pasture	Interior	37.76	73.92	73.10	52.10	47.80	56.30	17.00	16.00	14.00	12.00	17.00	13.00

Table S.4. Bald eagle and mallard dataset used in Chapter 4. Cell numbers represent mallard and eagle counts over six time occasions and at 36 sites and environmental covariates mean temperature ($^{\circ}\text{C}$) and maximum wind speed (Beaufort scale)

(continues)															
Site	State	Latitude	Longitude	Mallard					Bald eagle					Temperature	Wind speed
				1	2	3	4	5	1	2	3	4	5		
1	British Columbia	49.46	−115.52	0	0	1	0	4	0	0	0	1	4	16.00	1
2	British Columbia	49.15	−117.26	0	0	0	0	0	0	0	0	1	0	12.00	2
3	British Columbia	50.24	−115.87	0	0	0	0	0	0	0	0	1	0	9.00	0
4	British Columbia	50.92	−116.48	0	2	4	0	0	0	0	0	0	0	16.50	0
5	British Columbia	50.44	−118.74	0	0	0	2	0	0	0	0	0	0	18.50	2
6	British Columbia	50.02	−121.10	0	0	2	0	0	0	0	0	0	0	15.50	2
7	British Columbia	51.32	−116.99	2	0	2	0	0	0	0	0	0	0	10.00	4
8	British Columbia	52.15	−122.07	0	1	7	0	0	0	0	0	0	0	17.00	2
9	British Columbia	53.28	−120.08	2	0	0	0	0	0	0	0	0	0	7.50	1
10	British Columbia	53.04	−121.98	0	0	0	5	0	0	0	0	0	0	13.00	3
11	British Columbia	53.08	−119.67	0	0	0	0	0	0	0	0	0	1	12.00	5
12	British Columbia	54.97	−120.11	0	0	0	0	4	0	0	0	0	0	8.00	2
13	British Columbia	54.33	−129.30	1	0	0	0	0	0	0	0	0	0	13.50	1
14	British Columbia	55.45	−120.16	3	0	0	0	0	0	0	0	0	0	7.50	0
15	British Columbia	56.10	−129.28	0	0	1	0	0	0	0	0	0	0	15.50	0
16	British Columbia	49.04	−118.48	71	3	0	0	0	1	0	0	0	0	18.61	0
17	British Columbia	49.26	−123.26	0	0	1	0	0	0	0	0	0	0	16.00	2
18	British Columbia	50.88	−121.83	2	0	0	0	0	0	0	0	0	0	14.17	1
19	British Columbia	50.14	−122.96	0	0	0	0	0	0	0	0	0	1	11.50	0
20	British Columbia	50.18	−125.49	0	0	0	0	0	0	0	0	1	0	14.50	2
21	British Columbia	51.98	−124.37	0	3	0	0	0	0	0	0	0	0	10.00	1
22	British Columbia	53.94	−122.82	0	0	1	0	2	0	0	0	0	0	5.00	0
23	British Columbia	55.00	−127.32	0	0	1	0	0	0	0	0	0	0	11.50	1
24	British Columbia	56.59	−120.64	1	0	0	0	0	0	0	0	0	0	6.50	1
25	British Columbia	49.72	−115.79	0	0	0	0	0	0	0	0	2	0	10.00	1
26	British Columbia	49.78	−124.18	0	0	0	0	0	0	0	0	1	10	14.00	0
27	British Columbia	49.71	−118.78	0	0	0	1	0	0	0	0	1	5	10.50	2
28	British Columbia	49.78	−120.85	1	0	0	0	7	0	0	0	0	3	14.50	2
29	British Columbia	50.79	−121.61	0	0	1	2	0	0	0	0	0	0	18.89	0
30	British Columbia	49.52	−124.15	0	0	0	0	0	0	1	0	4	3	19.00	3
31	British Columbia	49.26	−119.53	3	0	0	0	0	0	0	0	0	0	22.50	3
32	British Columbia	49.25	−124.36	0	0	0	0	2	0	4	0	0	2	14.50	1
33	British Columbia	50.24	−121.58	0	0	1	1	0	0	0	0	0	0	19.00	1
34	British Columbia	49.37	−117.94	0	4	0	0	0	0	2	0	0	1	11.00	0

Table S.4. Bald eagle and mallard dataset used in Chapter 4. Cell numbers represent mallard and eagle counts over six time occasions and at 36 sites and environmental covariates mean temperature ($^{\circ}\text{C}$) and maximum wind speed (Beaufort scale)

(continues)

Site	State	Latitude	Longitude	Mallard					Bald eagle					Temperature	Wind speed
				1	2	3	4	5	1	2	3	4	5		
35	Oregon	45.38	-122.79	4	4	0	0	0	0	0	0	0	0	18.61	1
36	Oregon	45.08	-121.92	0	0	0	0	0	0	0	0	0	1	18.06	1
37	Oregon	45.22	-118.91	0	0	1	0	0	0	0	0	0	0	11.39	2
38	Oregon	44.55	-123.80	0	0	0	0	0	0	0	0	0	3	13.89	1
39	Oregon	44.16	-121.27	0	0	0	0	0	0	0	0	0	12	21.39	0
40	Oregon	44.13	-118.59	0	0	0	0	0	0	0	0	0	1	11.67	1
41	Oregon	43.37	-123.50	0	0	0	0	0	1	0	3	3	2	14.44	1
42	Oregon	43.91	-122.56	0	1	0	0	0	0	0	0	2	6	11.39	3
43	Oregon	43.34	-119.70	0	0	0	0	0	0	0	0	0	5	19.72	1
44	Oregon	45.28	-117.21	0	2	1	1	0	0	0	0	0	0	12.22	1
45	Oregon	45.23	-123.02	0	0	0	0	0	0	0	0	2	0	17.22	0
46	Oregon	45.26	-122.26	0	0	0	0	0	0	0	0	1	10	21.11	1
47	Oregon	45.54	-121.48	0	1	0	0	0	0	0	0	0	1	14.50	2
48	Oregon	44.42	-119.11	2	0	0	0	0	0	0	0	0	0	13.33	1
49	Oregon	43.99	-121.80	0	0	0	0	0	0	0	0	0	1	11.67	2
50	Oregon	42.96	-118.87	9	0	0	0	0	0	60	0	5	3	14.44	2
51	Oregon	43.52	-124.20	0	0	0	1	0	0	0	0	6	7	14.44	3
52	Oregon	42.95	-119.01	0	0	0	0	0	0	0	0	0	1	16.94	1
53	Oregon	44.34	-120.56	0	0	0	0	0	0	0	0	1	0	15.56	3
54	Oregon	46.02	-123.37	0	0	0	0	0	0	0	0	1	1	10.83	1
55	Oregon	42.11	-122.54	0	2	0	0	0	0	0	0	0	1	16.39	1
56	Oregon	42.39	-121.02	2	0	0	0	0	0	0	1	0	0	11.94	0
57	Oregon	43.22	-124.08	0	12	0	0	0	0	0	1	1	0	10.28	3
58	Oregon	43.64	-122.31	0	0	0	0	0	0	0	0	0	1	14.72	0
59	Oregon	42.35	-120.75	0	0	0	0	0	0	0	1	0	0	21.67	0
60	Oregon	42.62	-119.67	0	0	0	0	0	0	0	0	0	1	23.06	0
61	Oregon	42.66	-118.99	3	0	10	5	0	0	0	0	0	1	13.89	2
62	Oregon	45.01	-123.40	3	0	0	0	0	0	0	0	0	0	16.39	1
63	Oregon	44.97	-122.96	2	0	0	0	0	0	0	2	0	2	19.72	2
64	Oregon	43.84	-123.24	0	0	0	0	0	0	0	1	2	5	14.72	3
65	Oregon	43.33	-117.08	0	0	0	0	4	0	0	0	0	1	16.94	2
66	Oregon	42.87	-121.83	0	2	0	0	0	0	0	0	1	0	10.56	0
67	Oregon	42.58	-119.59	0	0	0	0	0	0	0	0	7	2	20.83	1
68	Oregon	42.53	-117.18	0	0	0	0	2	0	0	0	0	1	18.06	2
69	Oregon	44.67	-120.12	0	0	0	0	0	0	2	0	0	0	12.78	2
70	Oregon	44.60	-122.46	0	0	0	0	0	0	0	0	0	3	13.61	0
71	Oregon	45.79	-122.87	0	0	0	2	2	0	0	0	0	0	19.72	2

Table S.4. Bald eagle and mallard dataset used in Chapter 4. Cell numbers represent mallard and eagle counts over six time occasions and at 36 sites and environmental covariates mean temperature ($^{\circ}\text{C}$) and maximum wind speed (Beaufort scale)

															(continued)	
Site	State	Latitude	Longitude	Mallard					Bald eagle					Temperature	Wind speed	
				1	2	3	4	5	1	2	3	4	5			
72	Washington	48.11	−123.39	0	1	19	0	0	0	0	0	1	0	11.67	1	
73	Washington	48.16	−122.34	0	0	0	0	4	0	0	0	0	0	13.61	0	
74	Washington	48.36	−120.21	0	0	0	0	0	0	1	0	0	0	14.44	1	
75	Washington	48.09	−119.81	0	1	0	0	0	0	0	0	0	0	15.56	4	
76	Washington	47.09	−119.26	1	1	0	0	0	0	0	0	0	0	27.78	2	
77	Washington	46.26	−123.18	0	0	0	0	0	0	0	0	0	2	12.22	2	
78	Washington	46.64	−121.39	0	1	0	0	0	0	0	0	1	22	16.94	1	
79	Washington	45.83	−121.35	0	0	0	0	0	0	1	0	0	5	14.00	2	
80	Washington	48.97	−122.14	1	0	0	0	0	0	0	0	0	0	13.33	0	
81	Washington	48.39	−117.27	0	0	0	0	0	0	0	0	1	3	12.22	1	
82	Washington	47.98	−122.17	1	0	0	0	0	0	0	0	0	0	16.39	2	
83	Washington	47.23	−121.01	0	2	0	0	4	0	1	0	0	0	12.22	1	
84	Washington	47.55	−120.68	0	0	0	0	0	0	1	0	0	0	12.22	0	
85	Washington	47.77	−118.89	0	1	0	0	0	0	0	0	0	0	11.94	4	
86	Washington	46.27	−124.08	0	0	0	0	0	0	0	1	0	0	10.83	3	
87	Washington	46.40	−121.56	0	0	0	0	0	0	0	0	17	0	14.17	0	
88	Washington	46.96	−120.57	0	0	0	0	9	0	0	0	0	0	17.50	2	
89	Washington	47.24	−120.70	0	0	0	0	0	0	0	0	0	1	11.11	1	
90	Washington	47.13	−118.16	5	0	0	0	0	0	0	0	0	0	16.39	2	
91	Washington	46.97	−119.35	0	4	0	0	0	0	0	0	0	0	18.06	2	
92	Washington	46.47	−117.59	0	0	0	0	0	0	0	1	0	0	19.44	1	
93	Washington	48.54	−122.47	0	0	2	1	0	0	0	0	0	0	13.06	1	
94	Washington	48.53	−120.30	0	3	0	0	0	0	0	0	0	0	16.39	0	
95	Washington	48.22	−119.05	1	0	0	0	0	0	0	0	0	0	18.33	1	
96	Washington	48.68	−118.12	0	4	0	8	0	0	0	0	0	0	17.22	1	
97	Washington	47.66	−119.63	1	0	0	0	0	0	0	0	0	0	16.94	0	
98	Washington	46.54	−123.47	0	0	0	0	0	0	0	0	1	2	23.61	1	
99	Washington	47.74	−121.98	0	0	1	0	0	0	0	0	3	0	11.39	0	
100	Washington	46.77	−119.05	1	3	0	1	0	0	0	0	0	0	18.61	1	
101	Washington	47.41	−117.52	0	0	0	5	0	0	0	0	0	0	17.22	0	
102	Washington	48.46	−117.56	0	0	0	0	1	0	0	0	0	0	15.00	1	
103	Washington	46.39	−119.41	0	0	0	0	2	0	0	0	0	0	23.06	2	

Supplement S3: Dataset for Chapter 6**Table S.5.** Inscet attack dataset used in Chapter 6. Cell numbers represent number of times *Podisus nigrispinus* (stinkbug) or *Euborellia annulipes* (earwig) individuals attacked their competitor

Stinkbug	Earwig
0 0 0 1 2 0 1 0 0 0	1 2 4 1 7 2 4 5 1 1
0 0 0 6 0 0 0 0 0 0	5 1 1 2 2 2 1 4 1 3
1 3 2 0 0 0 0 0 0 0	4 4 4 4 1 1 1 1 4 1
0 0 0 0 0 0 0 0 1 0	1 5 7 4 2 2 4 1 1 8
0 0 0 1 0 0 0 0 0 0	4 1 1 5 2 3 2 3 3 1
0 0 0 4 0 2	2 3 1 7 1 3