Bayesian latent factor approaches for modeling ecological species communities

Gleb Tikhonov

LUOVA
Doctoral Programme in Wildlife Biology

Faculty of Biological and Environmental Sciences
University of Helsinki

Academic dissertation

To be presented for public examination with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki in the Auditorium K110, Latokartanonkaari 5, Helsinki, on 14th of December 2018 at 12 o'clock noon.

Helsinki 2018

Supervisors: Professor Otso Ovaskainen
University of Helsinki, Finland

Doctor Maria del Mar Delgado
University of Oviedo, Spain

Pre-examiners: Professor Sudipto Banerjee
University of California, Los Angeles, USA

Doctor James Thorson
National Marine Fisheries Service, USA

Opponent: Professor Alan Gelfand
Duke University, USA

Custos: Professor Otso Ovaskainen
University of Helsinki, Finland

Expert members of the thesis advisory committee:

Professor Anna Kuparinen
University of Jyväskylä, Finland

Doctor Aleksi Lehikoinen
University of Helsinki, Finland

ISBN 978-951-51-4664-9 (paperback)
ISBN 978-951-51-4665-6 (PDF)
https://ethesis.helsinki.fi
Unigrafia Helsinki 2018

Contents

List of articles included in the dissertation

The dissertation thesis is based on four articles that correspond to Chapters of the dissertation and are referred in the text by Roman numbers I-IV:

I. Ovaskainen, O., **Tikhonov, G.**, Norberg, A., Blanchet, F.G., Duan, L., Dunson, D., Roslin, T. & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters, 20, 561-576.

II. Minard, G.†, **Tikhonov, G.**†, Ovaskainen, O. & Saastamoinen, M. (submitted). The microbiota of a specialist herbivore in natural populations exhibits strong bacterial co-occurrence pattern that is only mildly affected by trophic interactions and individual background.

III. **Tikhonov, G.**, Abrego, N., Dunson, D. & Ovaskainen, O. (2017). Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. Methods in Ecology and Evolution, 443-452.

IV. **Tikhonov, G.**, Duan, L., Abrego, N., Newell, G., White, M., Dunson, D. & Ovaskainen, O. (submitted) Computationally efficient joint species distribution modelling of big spatial data.

† - these authors contributed equally and share first authorship

Table of contributions

	I	II	III	IV
Original idea	DD, NA, OO	GM	GT, OO	GT, OO
Data collection and preparation		GM, MS		GN, MW
Methods development	DD, GT, LD, OO	GT, OO	DD, GT, OO	DD, GT, LD
Software development	GB, GT		GT	GT
Data analysis	OO	GT, GM	GT	GT
Manuscript preparation	AN, DD, GB, GT, NA, OO, TR	GM, GT, MS, OO	DD, GT, NA, OO	DD, GN, GT, LD, MW, NA, OO
Supplement material preparation	AN, GB, GT, NA, OO	GM, GT, OO	GT, OO	GT, MW, OO

All authors are listed in the alphabetic order of their acronyms.

AN – Anna Norberg, DD – David Dunson, GB – Guillaume Blanchet, GM – Guillaume Minard, GN – Graeme Newell, GT – Gleb Tikhonov, LD – Leo Duan, MS – Marjo Saastamoinen, MW – Matt White, NA – Nerea Abrego, OO – Otso Ovaskainen, TR – Tomas Roslin.

Copyrights

Summary: © Gleb Tikhonov.
Chapter I: © 2017 The Authors. Ecology Letters.
Chapter II: © The Authors.
Chapter III: © 2017 The Authors. Methods in Ecology and Evolution.
Chapter IV: © The Authors.

Abstract

In the last decades, the aims of research in community ecology have been shifting from the mere description of observed patterns towards a mechanistic perspective that seeks to understand the processes shaping observed species communities. Simultaneously, the technical advances in data collection techniques dramatically raised the amount and quality of ecological data annually obtained and provided opportunities to address more comprehensive research questions. The combination of these novel aims and data increased the interest in the statistical ecology, seeking analytical methods capable to harness the full potential of the emerging data. A special interest has been focused on the development of approaches capable to combine multiple types of existing data and jointly model the dynamics and distributions of entire species communities or ecosystems.

This doctoral thesis contributes to the ongoing methodological development of analytical tools for the joint species modeling. In the presented research I combine both perspectives of the statistical ecology: the ecologist's practical point of view and the statistician's methodological/theoretical vision. The thesis consists of four Chapters that are arranged to form a coherent narrative. I start with a synthesis of the recent advances in joint species modeling and propose a unifying statistical framework that enables scientists to easily address many common questions in community ecology simultaneously. This framework, called Hierarchical Model of Species Communities (HMSC), is capable to incorporate information on species occurrences, environmental covariates, species traits and phylogenic relationships, as well as the structure of study design. Next, I devise and present two important extensions to this framework. The first extension enables HMSC to neatly assess the variation in species associations and relate it to environmental factors. My second extension aims to achieve better numerical properties for the HMSC-based analysis of numerous spatial observations. I carry out a set of simulated data experiments to assess the performances of the proposed extensions in comparison to existing methods. To demonstrate how the proposed methods can be used in practice, I accompany these methodological developments with real-data examples and additionally present one detailed applied ecological study.

My results demonstrate that the unifying HMSC framework can be robustly used to address a wide set of fundamental and applied ecological questions for various natural systems and contexts. Conducted simulation experiments verify that the proposed extensions considerably expand the framework's potential. The developed software implementation of the HMSC and detailed user manual provide a practical guidance for ecologists on how to apply this framework for analysis of their own data on species communities.

Although this thesis is a completed research item, it should be seen as a solid foundation for further developments in the field of joint species modeling. Some of these potential developments are related to how more comprehensive ecological questions could be answered with statistical models, while other correspond to the numerical challenges posed by emerging types and amounts of ecological data. I believe that advances and results of my study will enable future research to tackle these challenges and that the joint species modeling framework will become generally applicable and insightful for a wide array of real-world problems.

List of abbreviations

AQ – applied question

DAG – directed acyclic graph

FQ – fundamental question

GLM – generalized linear model

GMRF – Gaussian Markov Random Field

GP – Gaussian process

GPP – Gaussian predictive process

INLA – integrated nested Laplace approximation

JSDM – joint species distribution model(ing)

HMSC – Hierarchical Model(ing) of Species Communities

MCMC – Markov chain Monte Carlo

NNGP – Nearest Neighbor Gaussian process

OTU – operational taxonomic unit

SDM – (single) species distribution model(ing)

SGH – stress-gradient hypothesis

SSDM – stacked species distribution model(ing)

Introduction

Theoretical foundations of community ecology

Ecology has been described as the branch of biology that studies organisms and their interactions with surrounding environment and between each other, seeking the scientific understanding of factors determining their distributions (Smith 1966, Begon et al. 1996). This understanding can hardly be achieved by studying species separately one by one, since their abundances and distributions depend not only on their individual responses to the abiotic environment, but also on their interactions (Wisz et al. 2013). Consequently, the branch of community ecology studies the interactions between different species and aims to gain an integrative understanding of how biotic and abiotic factors shape observed local species pools at different spatiotemporal scales.

Community ecology began as a descriptive science in which communities were classified based on the identities and sizes of local species pools (Clements 1936). Modern community ecology is shifting beyond the mere description of observed patterns towards a mechanistic perspective, which seeks to understand the processes determining the identities and abundances of the species at different scales (Agrawal et al. 2007, Logue et al. 2011, Ovaskainen et al. 2016b). During the last few decades, experimental ecologists have used observations and experiments to assess the relative influences of stochasticity, competition and niche differentiation (Logue et al. 2011, Weiher et al. 2011), theoretical ecologists have developed models for predicting community dynamics (Pickett and McDonnell 1989, Bolker et al. 2003, Holyoak et al. 2005), and statistical ecologists have developed metrics for assessing compositional changes among local communities (Legendre and Legendre 1998). While a general theory to explain how communities are assembled across space and time is still lacking, community ecologists have converged towards a synthesis acknowledging that local species communities are shaped though both stochastic and deterministic processes, henceforth called assembly processes (Gravel et al. 2006, Leibold and McPeek 2006, Gotzenberger et al. 2012). These encompass abiotic or environmental filtering, biotic filtering, as well as neutral and contiguous processes such as speciation or dispersal (Vellend 2010).

Environmental filtering

Environmental filters correspond to those abiotic factors, such as temperature, moisture and soil nutrients, which enable or prevent the establishment or persistence of species in local communities, and thus outline the fundamental niche of a species (Kraft et al. 2014). One of the most intuitive and illustrative examples of environmental filtering effects comes from plant communities, for example the very distinctive and pronounced turnover of vegetation along the elevation gradient of a mountain slope. Although the scope of applicability of this term is currently debated by community ecologists, since in practice essentially no living being could be found influenced by abiotic environment only (Cadotte and Tucker 2017), historically, the idea of establishing the link between species distributions and abiotic factors was among the very foundations of ecological research (von Liebig and Playfair 1847).

Biotic filtering

Biotic filtering refers to interspecific and intraspecific interactions that determine the set of species in local communities, and thus define their realized niches (Wisz et al. 2013, Garnier et al. 2016). These interactions vary greatly in their outcomes for involved organisms and species, with at least the following categories being recognized: mutualism, commensalism, parasitism, neutralism, amensalism, competition, predation and pollination (Begon et al. 1996).

Nowadays, ecological theory proposes that different types of filtering processes may interact – the abiotic factors may modify the biotic interactions (Callaway and Walker 1997, Tylianakis et al. 2008). For instance, when resources become scarce, competition among species might be intensified (Goldberg and Barton 1992), whereas under abiotically stressful environmental conditions, facilitation might become particularly important (Brooker 2006, Maestre et al. 2009, He et al. 2013). Changes in the outcomes of interspecific interactions in relation to changing environmental conditions have been empirically found for a wide array of taxonomical groups (Erland and Finlay 1992, Brooker 2006, MacDougall et al. 2018).

Neutral and stochastic processes

Beyond the deterministic processes that drive the selection of realized species subsets, when zooming-in from regional to local species pools, stochastic processes create additional variation in the local communities. These processes – generally related to colonization, extinction, ecological drift, and environmental stochasticity – generate divergence among communities occupying identical environments (Chase and Myers 2011). For example, a long-term-stable population of a species in certain area could persist even despite of the location's characteristics being outside the fundamental niche, 'fueled' by an ongoing migration flow of individuals to this location. Another example would be the anthropogenic ecological barriers, such as highways, that for some species prevent the flow of individuals between the neighboring separated areas, which in turn could lead to very different ecological dynamics in these areas.

Statistical analysis of species communities

While faced with a variety of data types, community ecologists have so far been armed with rather disparate statistical tools for connecting them with theories on community assembly. In particular, we lack statistical frameworks that would enable us to robustly infer actual assembly processes from community samples (Logue et al. 2011), especially in observational studies. This leads to conceptual gap between predictions of theoretical models and available empirical data. Up to date, the most popular tools used to study community structure are distance-based ordinations (Braak and Oct 1986, Legendre and Legendre 1998) and diversity measures (Magurran 2004). While such approaches provide insights into patterns of diversity and community composition at different spatiotemporal scales (Legendre and Gauthier 2014), they offer little quantitative insight into the relative contributions of different assembly processes. To overcome these limitations, community ecologists are showing increasing interest in model-based approaches (Warton et al. 2015b).

Species distribution models

Species distribution models (SDM) have been widely used to explain and predict how different taxa respond to environmental variation (Guisan and Thuiller 2005). Most of generic regression or classification analytical tools have found their applications in species distribution modeling, ranging from simplest linear regression models to state-of-the-art artificial neural networks, boosted regression trees and non-parametric statistical methods, as well as ensembles of those (Elith and Leathwick 2009, Golding and Purse 2016). Inspired by this success, there is a growing interest in extending SDMs to community-level models (Guisan and Rahbek 2011). The most straightforward way for predicting community-level properties is to combine predictions of single-species models into 'stacked' species distribution models (SSDM), possibly with some post-hoc correction applied (Guisan and Rahbek 2011, Calabrese et al. 2014).

Ordination methods

Community ecologists have traditionally inferred the presence and strength of interspecific interactions from observational species occurrence data by examining species' co-occurrence patterns. Statistical methods for assessing species' co-occurrences include distance-based ordination approaches (Legendre and Legendre 1998), pairwise co-occurrence approaches (Veech 2014), metrics measuring species' aggregation and segregation patterns (Stone and Roberts 1990), and null model approaches (Gotelli 2000). A caveat with these methods is that they confound co-occurrence patterns generated by ecological interactions with those generated by co-variation in the species responses to abiotic variation, although more recent developments enable to examine whether the co-occurrences depend on environmental covariates (Williams et al. 2014). However, this approach does not necessarily clarify whether the environmental covariates influence the occurrences or co-occurrences of the species.

Joint species modeling

Another approach to community data analysis is the use of recently emerged joint species distribution models (JSDM), which explicitly acknowledge the multivariate nature of species assemblages, allowing one to gather more mechanistic and predictive insights into assembly processes (Warton et al. 2015a). JSDMs consider as the response variable the vector of occurrences or abundances of all species, and thus provide a model-based approach for inferring simultaneously species associations as well as species relationships to the abiotic environment (Ovaskainen and Soininen 2011, Pollock et al. 2012, Clark et al. 2014, Pollock et al. 2014, Ovaskainen et al. 2016a).

As JSDMs allow to control for the effects of measured environmental covariates on single species distributions, their estimates of species associations are more representative of true interactions than raw co-occurrence indices, especially if such inference on interactions is derived from partial correlations or from time-series data (Ives et al. 2003, Ovaskainen et al. 2016a, Thorson et al. 2016, Ovaskainen et al. 2017). Recently, community ecologists have adopted the Granger predictive causality principles and several studies exploited the vector autoregressive models for analysis of longitudinal observations of species communities, aiming to understand the community dynamics (Ovaskainen et al. 2017, Thorson et al. 2017). However, the potential of existing confounding factors makes the non-manipulative data on species occurrence insufficient for a conclusive causal inference on ecological interactions,

and therefore, species associations estimated by JSDMs should be treated with caution for claims regarding species interactions and preferably serve only as informed hypotheses, the validity of which should be verified in controlled experiments (Ovaskainen et al. 2010).

In the early phase of the JSDM development, these approaches suffered from the curse of dimensionality, limiting the estimation of species association matrices to only few tens of species (Latimer et al. 2009, Ovaskainen et al. 2010). Thanks to recently introduced statistical techniques based on latent factor modeling, e.g. Bhattacharya and Dunson (2011), current JSDMs are able to estimate species association matrices for hundreds of species, including study designs with multiple hierarchical levels (Ovaskainen et al. 2016a). Other recent method developments have made it possible to apply JSDMs to various types of ecological data, including presence-absence, counts and biomass (Hui 2016, Clark et al. 2017, Hui et al. 2017, Niku et al. 2017). Several studies have developed approaches to incorporate study designs of spatial, temporal or spatio-temporal nature (Sebastián-González et al. 2010, Thorson et al. 2015a, Ovaskainen et al. 2016c, Thorson et al. 2016). Increasing predictive performance of JSDMs by introducing potential non-linearity of included covariates' effects has been just one more topic of active research in last years (Harris 2015, Chen et al. 2016, Vanhatalo et al. 2018).

Conceptually or from non-statistician's point of view, JSDMs could be seen as a single-shot equivalent of consecutively applying SSDMs and additional post-hoc analysis of the resulted niches or residuals. For example, SSDM could be used to assess how species traits affect the structure of the community with respect to environmental covariates – first the SDMs are fitted separately and then the SDMs' parameters (e.g. linear regression coefficients in GLM) used as outcome variables in a second regression, which seeks to link the estimated niches to species traits. However, the joint probabilistic formulation of JSDMs allows to propagate the probabilistic uncertainty through all model components in a fairer way than could be achieved with analogous sequential analysis.

Research aims

In this doctoral thesis I make methodological contributions to the actively developing field of JSDM with the ambition to increase flexibility, robustness, computational efficiency, and practical usability of these models. My research is designed to provide practical resolve for existing challenges in modern community ecology, and my aims are split into two major categories: 1) development of novel statistical models and associated model fitting algorithms, and 2) demonstration of their utility for answering challenging questions in community ecology.

In **Chapter I**, my co-authors and I synthesize the results of several methodological enhancements for modeling species distributions jointly, which were introduced separately in recent years, and unite them within a single Bayesian statistical modeling framework. This statistical framework, named Hierarchical Model of Species Communities (HMSC) for its heavy dependence on hierarchical modeling techniques, provides a modular tool for statistical model-based analysis that is capable to incorporate and utilize multiple types of data common for community ecology: abundance/occurrences of species, quantified environmental factors, structure of sampling design, as well as species traits, attributes and phylogenic relationships. I present a full specification of the model structure alongside with a tailored block-Gibbs sampling scheme for efficient model fitting and its numerical implementation in Matlab and R. Together with co-authors, I exemplify how this generic model could be used to answer multiple ecological research questions by reproducing the analysis of three scientific research papers, which all focus on studying species communities, but differ greatly in ecological context, research questions and aims. I follow up with a more comprehensive and detailed example of a HMSC application in **Chapter II**, where my research is focused on studying the variation patterns of gut microbial communities in well-studied metapopulation of Glanville fritillary butterfly (*Melitaea cinxia*) in Åland Islands, Finland. This application poses a special interest for the development of JSDMs, as it deals with an extremely high number of taxa, which were sampled with high-throughput sequencing techniques that are becoming increasingly available and popular for obtaining community data on micro-organisms. Additionally, this study displays how the JSDM-based approach, originally designed for studies of macro-organism communities, can be successfully utilized in the analysis of micro-organisms, in which area statistical methods have been historically developing separately from those of macro-organisms.

Chapter III and **Chapter IV** introduce major augmentations to the baseline HMSC model, motivated by conceptual, methodological and numerical challenges actual for modern community ecology analysis. In **Chapter III**, I tackle the potential variation of species interactions and associations with respect to the environmental factors and propose an appropriate modification to HMSC model structure that enables a model-based approach to assess such variation. I demonstrate the technical validity and ecological relevance of this extension by comparing its performance to previously-published methods for both simulated data and a case study of arctic plant communities. In **Chapter IV**, I address the practical computational challenges of applying JSDMs in analyzing many (e.g. tens of thousands) spatially-structured observations. I investigate two techniques from recent spatial statistics methodological advances that were demonstrated to efficiently deal with modeling numerous

spatial observations, namely Gaussian predictive process (GPP) and Nearest Neighbor Gaussian process (NNGP). I devise modifications to incorporate them to HMSC structure and model fitting algorithm to appropriately harness their computational benefits. I study the properties of these solutions in terms of their computational burden and predictive performance, also comparing them to currently available approaches, and highlight the differences between them in context of modeling ecological communities. The relevance of the method is demonstrated by applying it to a large database on Australian plant communities.

Materials and methods

Typical data in community ecology include observations on the occurrence of species in a set of temporal and/or spatial replicates, henceforth called occurrence data and referred to as the Y matrix (Figure 1). Depending on the study/experimental design, research objectives and the subject organisms, the occurrence of the species can be recorded in various ways, and the occurrence matrix may thus describe e.g. presences-absences of species, species counts, percentage covered by each species or estimates of their biomass. The occurrence data are usually accompanied by environmental data consisting of a set of measured covariates that the ecologist hypothesized to be important in explaining community composition (X matrix, Figure 1). Beyond the effects of these environmental covariates, the spatiotemporal context may generate a structure to the data. In studies where the data have been collected in a hierarchical way (e.g. plots within sites), I call the finest scale (a single row of the data matrices X and Y) the 'sampling unit'. In studies treating space and/or time as continuous, the study design may be described by spatial or temporal coordinates. To relate community-level responses to environmental variation to response traits, one may wish to include data on species-specific traits (T matrix, Figure 1). These data may range from morphological traits such as body size, or physiological traits such as tolerance to salinity, to functional traits such as feeding type, or to the actual position of the species within the surrounding food web. Apart of trait data, an ecologist may also have information on phylogenic relationships (C matrix, Figure 1). The availability of phylogenetic data is rapidly increasing, allowing the construction of quantitative matrices of phylogenetic correlations for many organism groups. Where quantitative phylogenies are lacking, data on taxonomic identity (at the level of genus, family, order, class, phylum...) can be used as a proxy of phylogenetic relatedness.

Chapters I, III and IV of this thesis are primarily focused on methodological development of statistical modeling for analyzing data on species communities, and these chapters introduces new model structures and associated Bayesian model fitting algorithms.

Structure and formulation of Hierarchical Model of Species Communities

The statistical HMSC framework is illustrated graphically in Figure 2, and it is described in more detail below. I start by modeling the occurrence (e.g. presence–absence, count or biomass) of each species (denoted as j, where $j = 1 \dots n_s$) in each sampling unit (denoted as i, where $i = 1 \dots n_y$), i.e. the data summarized by occurrence matrix Y in Figure 1. For this, use a latent variable model, which technique is well-known to ecologists from generalized linear modeling (GLM) framework:

$$y_{ij} \sim D_j\left(L_{ij}, \sigma_j^2\right) \quad (1)$$

Here, D_j is a statistical distribution compatible with the particular type of measured data for observations in column j of matrix Y, L_{ij} is the latent variable, which corresponds to the location parameter of the distribution D_j, and σ_j^2 is the variance parameter that is omitted for certain distributions, e.g. Bernoulli with probit link function. The latent variable L_{ij} is modeled as a sum of fixed- (F) and random- (R) effects parts as $L_{ij} = L_{ij}^F + L_{ij}^R$. The fixed effects are modeled as a linear regression

$$L_{ij}^{F} = \sum_{k=1}^{n_c} x_{ik}\beta_{kj} \quad (2)$$

where, the index k runs over a set of n_c covariates, x_{ik} is the covariate k for sampling unit i, and β_{kj} is the response of species j to this covariate. The intercept is included by setting $x_{i1} = 1$ for all sampling units i, so that the number of included environmental covariates is $n_c - 1$. To allow the statistical framework to generate a community-level synthesis of how species respond to their environment, I assume that their responses to the environment (i.e. their regression parameters) adhere to a multivariate Gaussian distribution,

$$\beta_{\cdot j} \sim N(\mu_{\cdot j}, V) \quad (3)$$

I use the dot notation to single out a row or a column in a matrix, so that $\beta_{\cdot j}$ denotes the column-vector of regression coefficients for species j. As $\beta_{\cdot j}$ describes how species j responds to environmental covariates, it characterizes its environmental niche. The expected environmental niche of species j is denoted by column-vector $\mu_{\cdot j}$, and variation around this expectation is captured by the variance-covariance matrix V (Ovaskainen and Soininen 2011). The expected niche $\mu_{\cdot j}$ can either be assumed to be the same for all species, or alternatively it can model the influence of species-specific traits on species' responses. In the latter case, I assume another linear model

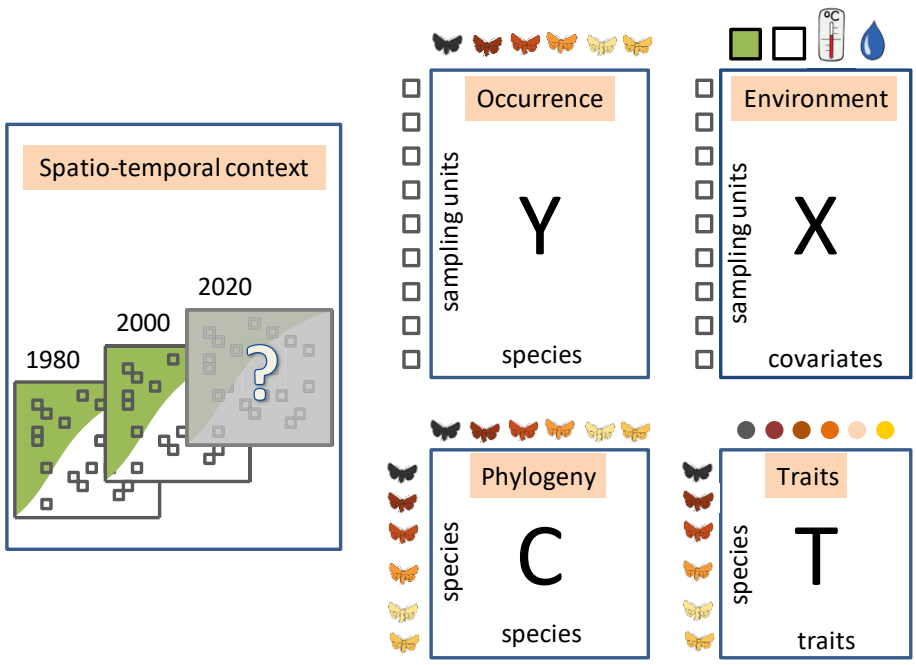


Figure 1. Data typically collected in community ecology. The occurrence data (denoted as the Y matrix) includes the occurrences of the species recorded in a set of temporal and/or spatial sampling units. The environmental data (denoted as the X matrix) consists of the environmental covariates measured over the sampling units. The traits data (denoted as the T matrix) consists of a set of traits measured for the species present in the Y matrix. To account for the phylogenetic dependencies among the species, we can include a fourth matrix consisting of the phylogenetic correlations among the species (denoted as the C matrix). The spatiotemporal context includes location and time information about the samples.

$$\mu_{kj} = \sum_{l=1}^{n_t} t_{jl}\gamma_{kl} \quad (4)$$

where t_{jl} is the value of trait l for species j (matrix T, Figure 1; with $t_{j1} = 1$ modeling the intercept) and the parameter γ_{kl} measures the effect of trait l on response to covariate k (Abrego et al. 2016). The equations (3) and (4) can be also used to ask what percentage of variation in species' environmental niches can be attributed to species' traits.

To account for phylogenetic relationships (summarized by matrix C, Figure 1), I add the joint structure for the multivariate Gaussian distributions of $\beta_{\cdot j}, j = 1 \ldots n_s$ as

$$\boldsymbol{\beta}^* = \left[\beta_{1\cdot}, \ldots, \beta_{n_c\cdot}\right]^T \sim N(\boldsymbol{\mu}^*, \Theta) \quad (5)$$

where $\boldsymbol{\mu}^* = \left[\mu_{1\cdot}, \ldots, \mu_{n_c\cdot}\right]^T$, and matrix Θ models the variation of responses among individual species around the trait-based expectation as

$$\Theta = V \otimes \left[\rho \mathrm{C} + \left(1 - \rho I_{n_s}\right)\right] \quad (6)$$

where $\otimes$ denotes Kronecker's product, and the parameter $0 \leq \rho \leq 1$ determines the strength of phylogenetic relationships on species responses to the covariates. The model can be applied without trait data by including the intercept as the only species trait, and it can be applied without phylogenetic data by fixing $\rho = 0$. From equation (6) it follows that for $\rho = 0$ the residual variance is independent among the species, implying that closely related species do not have more similar environmental niches than do distantly related ones. When ρ approaches $\rho = 1$, species' residual environmental niches (after accounting for the influences of the measured traits) are fully aligned according to their phylogeny, with related species having more similar niches than expected by random, implying niche conservatism.

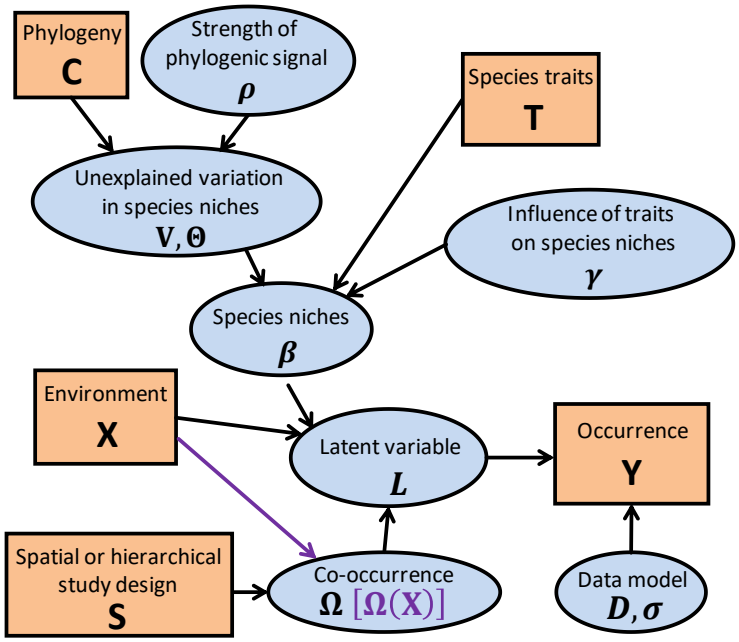
Next, I turn to the random terms L_{ij}^R, which model the variation in species occurrences and co-occurrences that cannot be attributed to the responses of the species to the measured covariates. If the study design consists of sampling units without any hierarchical, spatial or temporal structure, L_{ij}^R will simply be $L_{ij}^R = \varepsilon_{ij}$, referring to a random effect ε that operates at the level of the sampling unit. These random effects are modeled as

$$\varepsilon_{i\cdot} \sim N(0, \Omega), \quad (7)$$

where Ω is a residual species covariance matrix. Here, the word 'residual' refers to the fact that I have removed the influences of environmental covariates by the fixed effect part of the model. The diagonal element Ω_{jj} describes the amount of random variation that species j shows at the level of the sampling unit, whereas the off-diagonal element $\Omega_{jj'}$ describes the amount of covariation among the two species j and j'. For hierarchical study designs, the random terms L_{ij}^R are modeled as a sum of the random effects over all levels of design, each of which may additionally have a spatial or temporal structure:

$$L_{ij}^R = \sum_{r=1}^{n_r} L_{p_r(i)j}^r \quad (8)$$

Here n_r is the total number of study design levels, $p_r(\cdot)$ is the projection for study design level r, which maps the sampling units $i = 1 \ldots n_y$ to the corresponding units $q = 1 \ldots n_p^r$, and



$$y_{ij} \sim D_j\left(l_{ij},\ \sigma_j^2\right),\ \ L = XB + \sum_{r=1}^{n_r} P^r H^r \Lambda^r,$$

$$vec(B) \sim N(\boldsymbol{\mu}, \Theta),\ \ \boldsymbol{\mu} = vec(\Gamma T^T) = \left(T \otimes I_{n_c}\right) vec(\Gamma),\ \ \Theta = \left[\rho C + \left(1 - \rho I_{n_s}\right)\right] \otimes V,$$

$$V \sim W^{-1}(V_0, f_0),\ \ vec(\Gamma) \sim N(m_\Gamma, U_\Gamma),\ \ \sigma_j^{-2} \sim Ga\left(a_j, b_j\right),$$

$$\rho = 0, \text{ if no phylogeny},\ \ Pr\left(\rho = \rho_g \geq 0\right) = w_g,\ \sum_g w_g = 1, \text{ if phylogeny is included},$$

$$\lambda_{jh}^r | \phi_{jh}^r,\ \tau_h^r \sim N\left(0, 1/\left(\tau_h^r \phi_{jh}^r\right)\right),\ \ \phi_{jh}^r \sim Ga\left(\frac{v^r}{2}, \frac{v^r}{2}\right),\ \ \tau_h^r = \prod_{l=1}^{h} \delta_l^r,$$

$$\delta_1^r \sim Ga\left(a_1^r, b_1^r\right),\ \ \delta_l^r \sim Ga\left(a_2^r, b_2^r\right) for\ l = 2 \ldots n_f^r,$$

$$vec(H^r) = \boldsymbol{\eta}^r \sim N\left(0, I_{n_f^r} \otimes I_{n_p^r}\right), \text{ if latent factors on level r are not structured},$$

$$vec(H^r) = \boldsymbol{\eta}^r \sim N(0, K_H^r),\ \ K_H^r = diag\left(K^{r1}, \ldots, K^{rn_f^r}\right),\ \ K_{ii'}^{rh} = exp\left(-\frac{\|s_i - s_{i'}\|_2}{\alpha^{rh}}\right),$$

$$Pr\left(\alpha^{rh} = \alpha_g^{rh} \geq 0\right) = w_g^{rh},\ \sum_g w_g^{rh} = 1, \text{if latent factors on level r are structured}$$

Figure 2. Simplified directed acyclic graph (DAG) representation of Hierarchical Model of Species Communities (top) and the compact specification of the baseline model, introduced in Chapter I, using vector-matrix notation (bottom). Orange boxes stand for the data objects and blue ellipses denote model parameters. The purple color depicts the extension of Chapter III that enables association matrices to depend on environmental factors.

n_p^r is the total number of units at the design level r. Similar to equation (7), the random effects for each design level marginally follow a multivariate Gaussian distribution:

$$L_{q\cdot}^r \sim N(0, \Omega^r) \tag{9}$$

Therefore, the model with hierarchical study designs includes multiple species covariance matrices Ω^r, which correspond to potentially different patterns of species associations at different levels of the study design, see Ovaskainen et al. (2016a).

With n_s species, each covariance matrix Ω^r has bijective mapping to a space of $\frac{n_s(n_s+1)}{2}$ unrestricted parameters (Lewandowski et al. 2009), making their estimation numerically challenging. To facilitate the estimation of such matrices, I use a latent factor approach, which assumes a product representation of the matrix of random effects L^r through *latent factors* η_{qh}^r and *latent loadings* λ_{hj}^r

$$L_{qj}^r = \sum_{h=1}^{n_f^r} \eta_{qh}^r \lambda_{hj}^r \tag{10}$$

The factor loadings λ_{hj}^r themselves typically do not have a straightforward interpretation in terms of ecological interactions. They could be considered similar to linear regression parameters β_{kj} if the latent factors η_{qh}^r are interpreted to model some 'missing' covariates, which have an impact on the species occurrences and are not represented in the matrix X. For more detailed treatment of this interpretation see Warton et al. (2015a). Independently of their interpretation, given the classic assumption made in factor models that latent factors marginally follow multivariate Gaussian distribution $\eta_{q\cdot}^r \sim N\left(0, I_{n_f^r}\right)$, the latent loadings λ_{hj}^r provide a parametrization of Ω^r as

$$\Omega^r = (\Lambda^r)^T \Lambda^r. \tag{11}$$

The utility of the latent factor approach comes from the dimension-reduced parametrization of Ω^r in cases where $n_f^r \ll n_s$. I denote the species association network at the design level r by the correlation matrix R^r defined by $R_{jj'}^r = \frac{\Omega_{jj'}^r}{\sqrt{\Omega_{jj}^r \Omega_{j'j'}^r}}$. The correlation $R_{jj'}^r$ measures to what extent species j and j' are found together more or less often than expected by chance at the level r of the study design, after controlling for the environmental covariates and random effects of other design levels.

The number of latent factors n_f^r can be fixed *a priori* or treated as an unknown parameter through the shrinkage approach of Bhattacharya and Dunson (2011). From the practical point of view the shrinkage prior approach is beneficial compared to traditional methods of latent factor analysis for several reasons. First, in contrast to the typical strategy of restricting upper-triangular part of $(\Lambda^r)^T$ to zero, it imposes interchangeability in the prior for all modelled species. Next, this shrinkage approach enables an adaptive evaluation of the relevant latent factors number during the model fitting process, guided by the level of statistical support due to the complexity of observed associational patterns in the data. Finally, the estimated factors are naturally probabilistically ordered according to their relative importance, which step is often conducted during the interpretation of latent factor modelling with traditional priors (Thorson 2019). However, the prior of Bhattacharya and Dunson (2011) leads to the lack of identifiability for the parameters H^r and Λ^r, since those are invariant for simultaneous zero-flip of any latent factor and its loadings. However, the original authors stress that typically only the elements of Ω^r are of primary importance in applications and this derived matrix is identifiable. Alternatively, these latent factor approaches can be considered from the perspective of marginal prior on the covariance matrix Ω^r – with fixed number of factors all prior probability mass is assigned to covariance matrices which rank is no greater than n_f^r,

while the shrinkage approach is more flexible and assigns higher probability mass to matrices that are close to low-rank with respect to special matrix norm (see the original publication of Bhattacharya and Dunson (2011) for details).

I note that alternatively to the above-presented quantification of associations through covariance matrixes, a quantification based on the precision matrix (inverse of Ω) or related to it partial correlation matrix is possible, which are more likely to identify direct links among species than the correlation matrix, as the latter one is also influenced by indirect links (Ovaskainen et al. 2016a). As a further alternative, I note that instead of the latent variable approach, the random effects structure could be parameterized through a mixture modeling approach (Pledger and Arnold 2014).

The latent factor approach to modeling random effects enables convenient inclusion on spatial/temporal dependence at some levels of study design (Thorson et al. 2015a, Ovaskainen et al. 2016c). I denote the spatial/temporal coordinates of unit q at the design level r by $\boldsymbol{s}_q^r = [s_{id}^r]_{d=1\dots n_d^r} \in \mathbb{R}^{n_d^r}$, with typically $n_d^r = 2$ for spatial structure and $n_d^r = 1$ for temporal structure. To incorporate spatial/temporal structure in random effects L_{qj}^r, I assume that the latent factors $\eta_{\cdot h}^r$ come from realizations of a Gaussian process (GP) $w_h^r(\boldsymbol{s}) \sim \mathrm{GP}\left(0, k_h^r(\boldsymbol{s}_1^r, \boldsymbol{s}_2^r)\right)$ with a zero mean and a covariance function $k_h^r(\boldsymbol{s}_1^r, \boldsymbol{s}_2^r)$ (Rasmussen and Williams 2006). This implies that the h-th latent factor follows the multivariate Gaussian distribution

$$\eta_{\cdot h}^r \sim N\left(0, \mathrm{K}_{S^r S^r}^{(rh)}\right) \tag{12}$$

where the $\mathrm{K}_{S^r S^r}^{(rh)}$ is the covariance matrix for spatial/temporal units $S^r = \left\{\boldsymbol{s}_1^r, \dots, \boldsymbol{s}_{n_f^r}^r\right\}$ of the design level r included in the data, with pairwise covariance for the pair of units q and q' defined as $\left[\mathrm{K}_{S^r S^r}^{(h)}\right]_{qq'} = k_h^r\left(\boldsymbol{s}_q^r, \boldsymbol{s}_{q'}^r\right)$. This GP structure also implies spatial cross-covariance structure for the matrix of random effects L^r

$$\mathrm{vec}(L^r) \sim N\left(0, \sum_{h=1}^{n_f^r} \left((\Lambda_{h\cdot}^r)^T \Lambda_{h\cdot} \otimes K_{S^r S^r}^{(rh)}\right)\right) \tag{13}$$

Here the operation $\mathrm{vec}(A) \in \mathbb{R}^{nm}$ denotes the vector, produced by stacking the columns of a matrix $A \in \mathbb{R}^{n \times m}$: $\mathrm{vec}(A) = [A_{\cdot 1}^T, \dots, A_{\cdot m}^T]^T$.While this approach is valid with general classes of covariance functions that are parametrized by few scalar values, in the implemented HMSC framework it is restricted to the exponential covariance function $k_h^r(\boldsymbol{s}_1, \boldsymbol{s}_2) = \exp\left(-\frac{1}{\alpha_h^r}\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|_2\right)$ with unit variance and a single spatial range parameter α_h^r. The exponential covariance function implies stationarity and isotropy (Rasmussen and Williams 2006), and has been applied in previous work on spatial JSDMs (Ovaskainen et al. 2016c).

I design the implementation of the HMSC framework in Bayesian paradigm, as uncertainty quantification is essential for a reliable ecological inference, especially on association matrices. Hence the full model specification requires defining the prior distributions for all primary model parameters, which correspond to the source vertices in its DAG, see Figure 2.

Aiming to facilitate efficient model fitting schemes, whenever possible the priors are chosen to be conditionally conjugate to the associated parameters. In cases of parameters ρ and α^{rh}, the conditionally conjugate priors are not known, and I propose to assign a discrete point-mass prior over the ecologically relevant range of values. More details on prior distributions and recommended values of hyperparameters are provided in the Supplement part of the Chapter IV.

In Chapter III I extend the baseline HMSC model to allow the covariance matrix Ω^r to vary at different units of study design level r as a function of covariates $\Omega^r \mapsto \Omega^r(x_q^r)$, thus aiming to enable assessment of how species associations depend on environmental factors encoded with x_q^r.. I denote the matrix of covariates that is included for the r-th level of study design by X^r. These covariates may include some that duplicate the columns of the X matrix, but they may also include those that are not represented in the X matrix. To enable such variation of covariance matrices, I borrow from recent developments in the statistical literature (Hoff and Niu 2012, Fox and Dunson 2015) and, instead of assuming that latent loadings are constant, model them as a linear function of x_q^r.

$$\lambda_{hj}^r(x_{q\cdot}^r) = const \quad \mapsto \quad \lambda_{hj}^r(x_{q\cdot}^r) = \sum_{k=1}^{n_c^r} x_{qk}^r \lambda_{hjk}^r. \tag{14}$$

With this modification the equation (11) transforms to

$$\Omega^r(x_{q\cdot}^r) = \left(\Lambda^r(x_{q\cdot}^r)\right)^T \Lambda^r(x_{q\cdot}^r) \tag{15}$$

and correspondingly the correlation matrix also becomes dependent on the covariates $\mathrm{R}^r \mapsto \mathrm{R}^r(x_{q\cdot}^r)$ such that $R_{jj'}^r(x_{q\cdot}^r) = \frac{\Omega_{jj'}^r(x_{q\cdot}^r)}{\sqrt{\Omega_{jj}^r(x_{q\cdot}^r)\Omega_{j'j'}^r(x_{q\cdot}^r)}}$. Since apart of predicting the species associations for specific environmental conditions, ecologists are typically interested in quantifying the level of statistical evidence on whether the associations vary with different environmental conditions. To do so, I define the species-to-species matrix of posterior probabilities $G^r(x_{1\cdot}^r, x_{2\cdot}^r) = \Pr\left(R^r(x_{1\cdot}^r) > R^r(x_{2\cdot}^r)\right)$. A value of $G_{jj'}^r(x_{1\cdot}^r, x_{2\cdot}^r)$ close to 1 indicates that there is a high level of statistical support that the co-occurrence pattern among species j and j' is more positive under the environment $x_{1\cdot}^r$ than under the environment $x_{2\cdot}^r$, whereas a value of $G_{jj'}^r(x_{1\cdot}^r, x_{2\cdot}^r)$ close to 0 indicates that the opposite is true.

In Chapter IV I reconsider how the spatial dependence could be included to the HMSC framework. This extension is motivated by the computational complexity of a single Gibbs MCMC step in posterior sampling for HMSC model with spatial random effect defined at the sampling unit level – the complexity scales as $O(n_y^3)$ in processing time and $O(n_y^2)$ in memory storage. This means that models are practically infeasible to apply to datasets even with moderately large numbers of sites, such as n_y being in the order of thousands. Nowadays, the most popular method for dealing with spatial data in ecological is to apply integrated nested Laplace approximation (INLA) to a latent Gaussian model, approximated with Gaussian Markov Random Field (GMRF), which approach allows for both precise and fast Bayesian approximate inference on a wide class of models (Rue et al. 2017). However, as is

highlighted by the INLA authors, it is ill suited for models with high number of parameters, which property is inherent to multivariate statistical methods. Alternatively, other methods recently adapted the INLA utilization of GMRF to integrate over the random effects and combined it with automatic differentiation to perform maximal likelihood inference on the fixed effects (Thorson 2019). Contrasting to those, in this study I specifically aim to achieve comparable computational gains but to keep the attractive benefits of Bayesian approach. I tackle this problem with two methods from spatial statistics that were shown been capable to efficiently model big spatial datasets: Gaussian Predictive Process (Banerjee et al. 2008, Finley et al. 2015) and Nearest Neighbor Gaussian process (Datta et al. 2016). Both these methods replace the original covariance matrix in the equation (12) with its approximation of a special structure, which provides numerically feasible adjustments to the HMSC sampling algorithm. For simplicity, in the two following paragraphs I assume that the model contains only single spatially structured random effect at the level of sampling units and drop out the index r from the formulas.

The GPP denoted by $\widetilde{w}(\boldsymbol{s})$, is constructed from the values of the original GP $w(\boldsymbol{s})$ defined at m 'knot' locations $S^* = \{\boldsymbol{s}_1^*, \dots, \boldsymbol{s}_m^*\}$. Therefore, the value of the GPP at any site $\boldsymbol{s}_0$ is given by $\widetilde{w}(\boldsymbol{s}_0) = E(w(\boldsymbol{s}_0)|\boldsymbol{w}^*) = K_{\boldsymbol{s}_0 S^*} K_{S^* S^*}^{-1} \boldsymbol{w}^*$, where $\boldsymbol{w}^* = [w(\boldsymbol{s}_1^*), \dots, w(\boldsymbol{s}_m^*)]^T$ denotes the vector of the original GP values at the knot locations S^*, $K_{S^* S^*} = \left[k\left(\boldsymbol{s}_{i_1}^*, \boldsymbol{s}_{i_2}^*\right)\right]_{i_2=1\dots m}^{i_1=1\dots m}$ and $K_{\boldsymbol{s}_0 S^*} = [k(\boldsymbol{s}_0, \boldsymbol{s}_1^*), \dots, k(\boldsymbol{s}_0, \boldsymbol{s}_m^*)]$. With this definition, it follows that $\widetilde{w}$ is itself a GP: $\widetilde{w}(\boldsymbol{s}) \sim \mathrm{GP}\left(0, \tilde{k}(\boldsymbol{s}_1, \boldsymbol{s}_2)\right)$, where the covariance function $\tilde{k}(\boldsymbol{s}_1, \boldsymbol{s}_2) = K_{\boldsymbol{s}_1 S^*} K_{S^* S^*}^{-1} K_{\boldsymbol{s}_2 S^*}^T$ is non-stationary but factorizable (Banerjee et al. 2008). This key property of GPP greatly decreases the computational complexity of the model when $m \ll n_y$, as sampling the posterior distribution takes $O\left(n_y m^2\right)$ in processing time and $O\left(n_y m\right)$ in memory storage (Banerjee et al. 2008). In equation (11), my definition of the covariance matrix Ω assumes that the marginal prior distribution each latent factor η_{ih} is standard normal. However, the GPP fails to fulfill this requirement since its marginal variance generally decreases with increasing distance from the knot set S^*. To circumvent this misbehavior, I apply a correction to the marginal prior variance of the GPP, so that it always equals that of the original GP (Finley et al. 2009).

A more recent advance in spatial statistics, Nearest Neighbor Gaussian Process (Datta et al. 2016), builds upon the conditional representation of the original GP. Given a specified ordering over the set of sites $S = \left[\boldsymbol{s}_1, \dots, \boldsymbol{s}_{n_y}\right]$ the process $w(\boldsymbol{s}) \sim \mathrm{GP}\left(0, k(\boldsymbol{s}_1, \boldsymbol{s}_2)\right)$ over this set corresponds to a multivariate Gaussian distribution $\boldsymbol{w} = \left[w(\boldsymbol{s}_1), \dots w\left(\boldsymbol{s}_{n_y}\right)\right]^T \sim N(0, K_{SS})$ that can be specified in conditional manner:

$$
\begin{gathered}
w_1 \sim N(0, K_{11}) \\
\left(w_i | w_j, j < i\right) \sim N(\mu_i, d_i) \; \forall i \in 2 \dots n_y, \\
\text{where } \mu_i = \sum_{j=1}^{i-1} a_{ij} w_j \\
\left[a_{i,1}, \dots, a_{i,i-1}\right]^T = \left(\left[K_{j_1 j_2}\right]_{j_2=1\dots i-1}^{j_1=1\dots i-1}\right)^{-1} \left[K_{1i}, \dots, K_{i-1,i}\right]^T
\end{gathered}
\tag{16}
$$

$$d_i = K_{ii} - \left[K_{1i}, \ldots, K_{i-1,i}\right]\left[a_{i,1}, \ldots, a_{i,i-1}\right]^T$$

This representation corresponds to a factorization of the covariance matrix $K = \left(I_{n_y} - \mathrm{A}\right)^{-1} \mathrm{D} \left(\mathrm{I}_{n_y} - \mathrm{A}\right)^{-T}$, where A is the strictly lower triangular matrix with elements a_{ij} and D is the diagonal matrix with elements d_i. The Nearest Neighbor approach approximates the conditional distribution $\left(w_i | w_j, j < i\right) \sim N(\mu_i, d_i)$ by conditioning only on the m preceding closest neighbors of $\boldsymbol{s}_i$: $\left(w_i | w_j, j < i\right) \approx \left(w_i | w_j, j \in N(i)\right)$. This results in an approximate factorization of covariance matrix $K \approx \widehat{\mathrm{K}} = \left(\mathrm{I} - \widehat{\mathrm{A}}\right)^{-1} \widehat{\mathrm{D}} \left(\mathrm{I} - \widehat{\mathrm{A}}\right)^{-T}$ with sparse matrix $\widehat{\mathrm{A}}$; hence the precision matrix $\widehat{\mathrm{K}}^{-1} = \left(\mathrm{I} - \widehat{\mathrm{A}}\right)^T \widehat{\mathrm{D}}^{-1} \left(\mathrm{I} - \widehat{\mathrm{A}}\right)$ is also sparse with $O\left(n_y m^2\right)$ non-zero entries. The enhanced computational efficiency of this method is achieved due to the decreased cost of sparse matrix operations compared to their dense counterparts.

Bayesian model fitting algorithms

The devised HMSC model could be fitted with various generic software for Bayesian model fitting. However, due to very high number of parameters in the model when fitting to data on many species sampled in many sites, conventional tools like JAGS are practically inapplicable due to extremely slow convergence to the posterior. Hence, I present another MCMC scheme that can be characterized as full-conditional Gibbs block sampler. This algorithm efficiently utilizes the potential of conjugate priors, conditional independences and is specified by a set of full-conditional updaters. In this subsection I first introduce the algorithm for the baseline HMSC model, following model's matrix notation presented in Figure 2, and then briefly describe how it is modified in Chapters III and IV.

- Latent variable L is sampled elementwise using appropriate data augmentation techniques for the distributions D_j associated with the columns of abundance matrix Y. The implemented data augmentation techniques cover probit data augmentation for presence-absence data (Albert and Chib 1993) and lognormal Poisson augmentation for counts (Zhou et al. 2012), with non-overdispersed Poisson augmentation considered as a limiting case.
- Linear regression coefficients B are sampled species-by-species $\mathrm{B}._j$ following Bayesian scheme for univariate linear regression when no phylogeny matrix C is included to the model. If phylogeny is included, all coefficients are sampled jointly following the formulas of conditional multivariate Gaussian distribution.
- Trait impacts on regression coefficients Γ are sampled following the Bayesian scheme for univariate linear regression.
- The strength of phylogenic signal ρ is sampled from its discrete prior locations, proportional to the prior weights multiplied with the full-conditional likelihood of linear regression coefficients.
- Unstructured random variation scales σ are sampled one-by-one from their Gamma full-conditional posterior distributions.

Parameters associated with different sampling design levels are sampled consecutively, conditional on the values of all parameters of other levels. Therefore, I drop the sampling design level index r from the following formulas for simplicity.

- Latent loadings Λ are sampled species-by-species $\Lambda_{\cdot j}$ following Bayesian scheme for univariate linear regression.
- The priors for latent loadings ϕ_{jh} and δ_h are sampled according to the algorithm proposed in Bhattacharya and Dunson (2011).
- If latent factors H are not assumed to be spatially/temporally structured, latent factors $H_{q\cdot}$ are sampled independently following Bayesian scheme for univariate linear regression. If the spatial/temporal structure is included, all latent factors H are sampled jointly following the formulas of conditional multivariate Gaussian distribution.
- Spatial range parameters α_h are sampled one-by-one from their discrete prior locations, proportional to the prior weights multiplied with the full-conditional likelihood of corresponding latent factor values $H_{\cdot h}$.

I implement the HMSC sampling algorithm in Matlab scientific programming language. This implementation efficiently utilizes the Matlab efficiency for matrix and tensor operations and partially exploits the multithreading potential for parallelizing the parts that could be run simultaneously. Given the strong tendency of ecological community to use R as primary language for scientific computing, a mirroring R implementation of HMSC framework is developed in collaboration with other researchers.

The transition from constant latent loadings to latent loadings that are dependent on covariates $\lambda^r_{hj}(x^r_{q\cdot}) = \sum_{k=1}^{n^r_c} x^r_{qk}\lambda^r_{hjk}$, introduced in Chapter III, requires changes to conditional updaters of H, Λ, ϕ_{jh} and δ_h. Conceptually, the modified formulas do not differ from corresponding counterparts in the baseline HMSC model, although they get algebraically more complicated. For details see Supplement section of Chapter III.

Modifications of Chapter IV concern only the full-conditional updaters of H and α_h. I propose algorithms that utilize the low-rank property of Gaussian predictive process and precision matrix's sparsity of Nearest Neighbor Gaussian process. For details see Supplement section of Chapter IV.

Empirical datasets

Each chapter of my thesis involves statistical analysis of at least one real dataset. In this subsection I provide a brief description for all these datasets, and the more detailed descriptions are given either in the Chapters, or in case of previously published data, in their original publications as cited below.

Chapter I.

a. Bryophytes dataset, presented in Oldén et al. (2014). The dataset contains abundances of 60 bryophyte species, recorded on aspen trees in 14 retention and 14 conservation sites in Central Finland in 2018. Overall, 204 aspens of various age were surveyed, 102 in retention sites and 102 in conservation sites. The species abundance was calculated as the total area (cm^2) of the trunk, covered by this species. The dataset also contains 4

covariates: diameter of the aspen tree, retention/conservation site type, time since logging and stand age.

b. Butterfly dataset, presented in Ovaskainen et al. (2016c). The dataset covers the presence-absence of 55 butterfly species in the Great Britain. The data is based on the 1995-1999 atlas data, with the whole study region being split to 2609 grid cells of size 10×10 km. Alongside the species data, the dataset contains 4 covariates: the number of growing days above 5°C, percentage of broadleaf woodland cover, percentage of coniferous woodland cover and percentage of calcareous substrates.

c. Waterbirds dataset, presented in Sebastián-González et al. (2010). The dataset consists of time-series on 7 species of waterbirds, observed at 221 irrigation ponds in the Vega Baja Valley, southern Spain between 2002 and 2008. In each year, most of these ponds were visually surveyed during daylight for 3-4 weeks in the breeding season and presence-absences of the focal species were recorded. Additionally, the dataset contains information of the following pond characteristics: pond area, distance to closest wetland, connectivity to other ponds, pond construction design, presence of submerged vegetation, presence of shore vegetation, and presence of reed.

Chapter II. Microbiota dataset. The dataset contains normalized operational taxonomic unit (OTU) quantification of bacterial DNA sequences, contained in larval and plant samples from natural populations of the *M. cinxia* and its host plant *P. lanceolata* in the Åland islands. The field sampling was conducted within three-day period in September 2015, following the general framework of the long-term survey of the *M. cinxia* butterfly that is described in Ojanen et al. (2013). DNA consisting of the V5-V6 region in the *rrs* gene was extracted from midgut for larvae samples and from the center of the leaf for plant samples. Alongside the microbiota data the dataset includes information on larvae sex, status of parasitoid infection by *H. horticola* and plant metabolome profile. Overall, the dataset covered 142 larvae samples and 55 host plant samples. The datasets for larvae and plant microbiota included 562 and 610 OTUs correspondingly, for which a phylogenic correlation matrix was obtained with FastTree method assuming the General Time Reversible (GTR) evolution model.

Chapter III. Arctic plant dataset, originally presented in Mod et al. (2014). The dataset contains the projective cover of 18 vascular plant species measured in 960 1-m^2 cells, arranged in six nearby rectangular plots, located approx. 700 m. a.s.l. on a northern slope of the Saana massif, northwest Finland. The fieldwork was conducted in July 2011. The plots had the shape of 8×20 m, and the maximum distance between plots was 110 m. Two environmental covariates were quantified in all 1-m^2 cells: soil moisture and integrated measure of disturbance, representing the cover of disturbed topsoil in the cell.

Chapter IV. Australian plant dataset. The data originate from the Victorian Biodiversity Atlas (https://www.environment.vic.gov.au/biodiversity/victorian-biodiversity-atlas). The subset of this Atlas used in this study involves the occurrences of 1237 herbaceous species, at 30,955 sampling locations within the State of Victoria, Australia, for which presence-absence were recorded. The data were collected in 1984-2014 years on sampling plots of 3900 m^2. The dataset combines survey data undertaken for a range of purposes the predominant being: 1) ecosystem inventory, circumscription and mapping, 2) characterizing the habitats of species of management interest and 3) documenting and describing land subject to development or

land-use change. The sampling design is known to be biased towards public lands, typically less suitable for agriculture and peri-urban areas. Additional to the species data, I used four environmental covariates that were considered potentially important to vegetation and plant distribution: mean maximum temperature in January, measure of hydrology and landscape position, soil properties, and solar radiation and anisotropic heating. Further, 9 species traits were included as binary indicator variables, describing whether the species 1) is annual or perennial, 2) is pollinated by abiotic or biotic means, 3-4) has propagules that are dispersed by wind, invertebrates, or another agent, 5) forms a seed bank that typically persists for two or more years, and is considered vulnerable to or tolerant of 6) fire, 7) prolonged snow cover, 8) protracted waterlogging, or 9) salinity.

Results and discussion

In this section I outline the main results of the four chapters in this thesis and provide a unifying discussion of those. I first present results related to methodological development, and then results related to the ecological examples.

Methodological advances

Unifying framework

Chapter I illustrates, how the recent advances in communities and JSDMs can be incorporated within a single statistical approach. The key value of the resulted Bayesian model, termed with Hierarchical Model of Species Communities (HMSC, Figure 2), is due to its clear, easily-interpretable, but robust design of how its different components relate to processes of community assembly. The HMSC application to three contrasting case studies demonstrates how the inference on assembly processes can be extracted from real data sets. While all examples are based on published studies, the primary novelty of this Chapter is in illustrating how a wide range of questions and data types can be analyzed with the help of this encompassing statistical framework. The types of communities and research questions of these case examples vary greatly: design of the first study is spatially hierarchical, the second study is spatially explicit, and the third study involves time-series data; the first deals with bryophytes, second – butterflies, and third – birds. Still, once the available data is organized according to the structure, presented in Figure 1, the HMSC could be used in seeking answers for questions actual in modern ecological community studies. A selection of such questions is summarized in Table 1, which also describes how HMSC can be applied to obtain an answer.

The developed Matlab and R packages, accompanied by a detailed user manual provide a practical guidance for ecologists on how to apply the HMSC for analysis of their own data on species communities.

Overall, the benefits of choosing HMSC to analyze species communities are summarized in the following list:

(1) HMSC is a unifying framework which encompasses classic approaches such as single-species distribution models and model-based ordinations as special cases.
(2) HMSC provides simultaneous inferences at the species and community levels.
(3) HMSC offers the general advantages of model-based approaches, such as tools for model validation and prediction.
(4) HMSC overcomes previous problems of modeling communities with sparse data.
(5) HMSC overcomes the long-standing challenge in species distribution modeling of how to account for species interactions in explaining and predicting species occurrences.
(6) HMSC allows one to partition observed variation in species occurrences into components related to environmental variation measured vs. random processes at different study design levels – both at the species and community scales.
(7) HMSC tackles the fourth corner problem (the influence of species traits to their occurrences, see Dray and Legendre (2008)) in a way that accounts for the phylogenetic signal in the data.

(8) HMSC can be applied to many kinds of study designs (including hierarchical, temporal or spatial) and many types of data (such as presence–absence, counts and continuous measurements).
(9) HMSC can generate predictions at the species, community or trait levels, while propagating uncertainty in estimated parameter values to the level of the prediction.

However, the core framework presented in Chapter I should be considered just as a starting point: while the HMSC presented here already allows one to address any fundamental and applied questions in community ecology (Table 1), it clearly does not answer all of them. As I have based HMSC on hierarchical generalized linear mixed models, adding additional layers is conceptually and technically straightforward. Below, I cover two further developments that build on the core framework, and discuss the key perspectives that I consider especially fruitful in context of modeling ecological communities.

Table 1. A summary of topical questions in community ecology and an outline of how they can be addressed within the HMSC framework. The presented list is not exhaustive, and the line between fundamental and applied questions is somewhat blurred.

	Question	How to address the question statistically?
FQ1	How much variation in species occurrence is due to environmental filtering, biotic interactions, and random processes?	By assessing the explanatory power of models and by variance partitioning among fixed and random effects.
FQ2	How does the importance of environmental filtering, biotic interactions, and random processes vary across spatial and temporal scales?	By variance partitioning between fixed and random effects operating at different scales.
FQ3	How do species' traits influence ecological niches?	By modelling responses to environmental covariates as a function of species' traits.
FQ4	Do phylogenetic relationships correlate with ecological niches, beyond that explained by traits?	By including the phylogenetic correlation matrix (**C**) when modelling the responses of species to environmental covariates.
FQ5	Are there signals of niche conservatism or niche divergence?	By examining whether the phylogenetic correlation matrix (**C**) helps to explain the data.
FQ6	What are the structures of species interaction networks?	By estimating the species-to-species association matrices **Ω.**
FQ7	How does community similarity depend on environmental similarity and/or geographic distance?	By decomposing community similarity into similarity due to responses to environmental covariates and/or spatial covariance.
FQ8	How does community structure change over time due to predictable succession or stochastic ecological drift?	By including time since environmental perturbation as a predictor, or by including temporally varying random effects.
AQ1	Do some species indicate the presence of others?	By testing how much the predictive power of the model increases for a focal species when accounting for the occurrences of other species.
AQ2	How can geographic areas be classified into communities of common profile?	By clustering of predicted communities based on their similarity.
AQ3	Which processes have been central in determining the response of a community to environmental change?	By decomposing the response to environmental change to components related to species niches and random effects.
AQ4	How can species be classified in terms of their response to abiotic environment?	By clustering parameters or predictions measuring the species responses to environmental covariates.

Variable species associations

One of the fundamental drawbacks of the baseline HMSC model from Chapter I is that it assumes the structure of association networks to be constant across all study. However, the type and strength of ecological interactions may be context-dependent (Poisot et al. 2016). For example, the stress-gradient hypothesis (SGH) predicts that positive interactions are accentuated under stressful abiotic environmental conditions (Callaway and Walker 1997). Hence, in Chapter III I address this challenge and introduce an enhancement to HMSC, aiming to enable the context-dependence of the association matrices by modeling the underlying latent variable structure as a linear function of environmental covariates. After implementing the augmented sampling algorithm, I test the performance of the proposed model with two simulation-based studies.

In the primary study, I use simulated data to test whether the developed framework successfully estimates the dependency of species associations on the environmental context. I generated presence-absence data on species occurrence along a single environmental gradient, called altitude for the sake of illustration (x_{i2}), with two kinds of models: in the null model the species associations were constant along the altitudinal gradient, and in the full model the species associations varied along the altitudinal gradient. Thus, in the null model I assumed the linear predictor

$$L_{ij} = \beta_{j1} + x_{i2}\beta_{j2} + \sum_{h=1}^{n_f} \eta_{ih}\lambda_{jh1} \tag{17}$$

whereas in the full model I assumed the linear predictor

$$L_{ij} = \beta_{j1} + x_{i2}\beta_{j2} + \sum_{h=1}^{n_f} \eta_{ih}\left(\lambda_{jh1} + x_{i2}\lambda_{jh2}\right) \tag{18}$$

The observations were obtained as $y_{ij} = 1_{L_{ij}+\epsilon_{ij}>0}$, where $\epsilon_{ij} \sim N(0,1)$. The β parameters were generated such, that species varied in their responses to the altitudinal gradient and most species were rare. Additionally, I imitated that the species associations depend on altitude in a manner that is in line with SGH: the expected proportions of positive and negative associations at low altitude are equal, while at high altitude most of associations are positive.

For each simulated dataset, I fitted the HMSC with structure equal to the full model and evaluated whether the model was able to correctly capture the variation on the species associations along the environmental gradient. To do so, I computed for each species pair the level of statistical evidence $S_{jj'}(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*)$ that their association was more positive at high altitude than at low altitude. For each species pair I classified the inference obtained from the fitted model as 'correct', 'misleading' or 'lack of statistical power' based on the match between the estimate of $S_{jj'}(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*)$ and underlying truth. I compared the performance of this method with SDM-based methods that were used earlier (Mod et al. 2014).

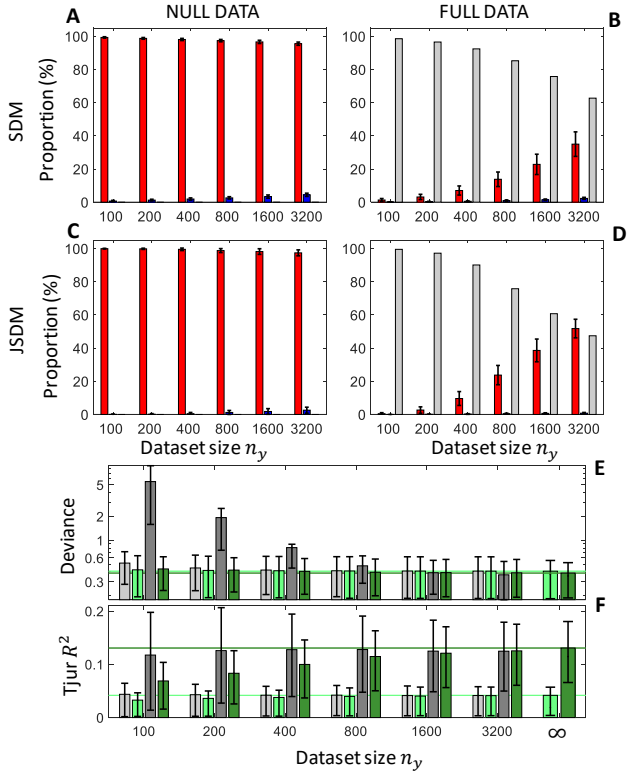


Figure 3. A comparison between single species distribution models (SDMs) and joint species distribution models (JSDMs) in their ability to account for the dependency of species associations on environmental covariates. The heights of the bars in ABCD show the proportions of species pairs (averaged over the replicate data sets; error bars show standard errors) with correct inference (red), misleading inference (blue), and lack of statistical power (grey). Panels E and F compare the predictive powers of the fitted models, measured by deviance (E) and Tjur R2 (F) of the model's predictions for validation data (averaged over the replicate data sets; error bars show 25%, 75% quantiles). The colors correspond to different models: SDM (light grey) and JSDM (light green) that do not account for species associations (unconditional), and SDM (dark grey) and JSDM (dark green) that utilize the information on the 10 most dominant species (conditional). The two rightmost bars and the horizontal lines represent the unconditional and conditional JSDMs with true parameters values that were used in the data simulation.

Next, I examined how much accounting for species associations influenced the predictive powers of the models. I mimicked a situation in which the ecologists would have surveyed the occurrences of all species for the sampling units that make up the training data (used for model fit), but only the few most dominant species for additional sampling units that make up the validation data. The question was on how well different models are able to predict the occurrences of the remaining species in the validation data. I compared the predictive powers of SDMs and HMSC that do or do not account for species associations: predicted the occurrences of the non-surveyed species in the validation data with these four models, and evaluated the predictive power with two measures: Tjur's R^2 and deviance (Tjur 2009).

My results with this simulated data demonstrate that the approach proposed here is capable of finding signals of changing associations in a robust and statistically efficient way. As expected, the ability to classify how species associations depend on environmental conditions increases with the size of the data set. The HMSC approach have more statistical power than the SDM for capturing whether the associations changed with altitude, while the fraction of misleading inferences is in line with expectations based on the threshold criteria used for both approaches. In terms of predictive power, unconditional SDMs and HMSC show almost equal performance. Conditioning the predictions for each focal species on the dominant species considerably improve the models' predictions in terms of the Tjur's R^2 both for SDM and HMSC, with no major differences

between these two approaches, but in terms of the deviance, the HMSC consistently overperform SDM. HMSC perform better than SDMs especially for the sparse data.

As an additional case study, I examine the robustness of the statistical approach by applying it to data simulated by an individual-based model, which data may violate the structural assumptions made by the HMSC. My results indicate that the proposed enhancement is flexible enough to successfully capture how associations vary along the environmental gradient also in this example, which attempts to mimic the nature of real-world assembly processes (see Supplement part of Chapter III).

Determining whether the outcomes of interspecific interactions depend on the environmental conditions is of major interest both for understanding the basic ecology and distribution of species, as well as for predicting changes at macroecological scales due to global change (Tylianakis et al. 2008, Hagen et al. 2012). The HMSC enhancement introduced in this Chapter provides a new tool for ecologists interested in inferring the dependency of interspecific interactions on environmental context from non-manipulative observational community data. Compared to existing methods, the principal advantage of this novel method is that it enables inference about species associations and their changes from sparse data on large communities dominated by rare species.

Numerous spatial data

In contrast to the Chapter III, the Chapter IV introduces no conceptual modifications to HMSC framework but aims to resolve a practical issue of computational challenges that arise in modeling numerous spatial observations. Motivated by increasing availability of high-resolution datasets covering large spatial scales (Graham et al. 2004, Franklin et al. 2017), and the demand for numerically superior approaches, highlighted by previous spatial JSDM-oriented studies (Ovaskainen et al. 2016c), in this Chapter I investigate the advances of spatial statistics that have been shown capable to overcome similar issues for univariate or low-dimensional multivariate modeling, namely Gaussian predictive process and Nearest Neighbor Gaussian process (Banerjee et al. 2008, Finley et al. 2009, Finley et al. 2015, Datta et al. 2016). I devise a way how these methods could be efficiently exploited in the latent factor component of HMSC models, and exemplify my developments with a comprehensive set of analyses to assess their performance and utility.

I use the Australian plant data to (i) test the feasibility to apply the methods presented here to data that are large in terms of both the number of sampling sites and the number of species, and to (ii) examine their performance in comparison to full spatial and non-spatial models, and how the key parameters of the approximate methods (number of knots in GPP and number of neighbors in NNGP) influences such comparisons. I selected four environmental covariates that are essentially uncorrelated and were considered potentially important to vegetation and plant distribution. Further, I included 9 species traits as binary indicator variables, describing the potentially relevant properties of the species. I randomly selected a set of 5000 sites that were used as test data and were not used for model fitting. I restricted the analysis to those $n_s = 623$ species that were observed at least 5 times both in the test part and in the remaining part of the dataset. I randomly selected training datasets of sizes $n_y = 100$, 400, 1600, 6400 and 25955 sites from the remaining locations. To examine how the performance of the

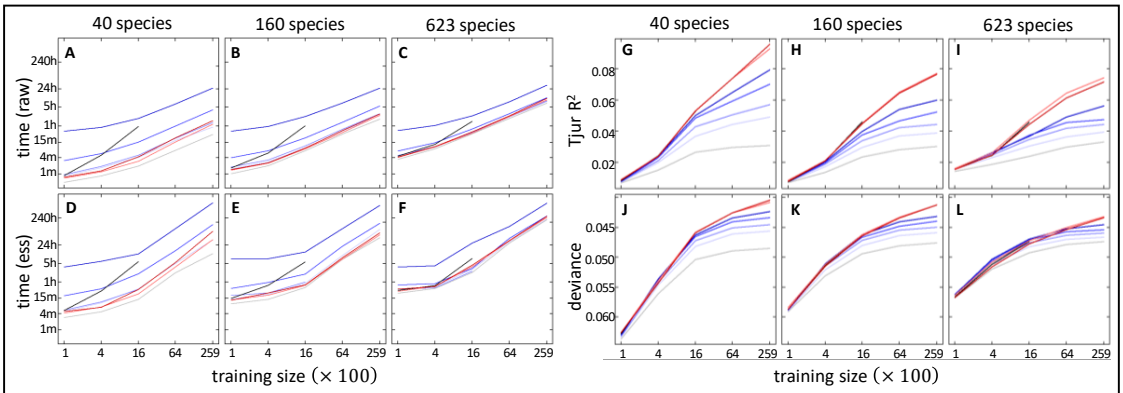


Figure 4. Comparison of non-spatial, GP, GPP and NNGP models. Panels ABC show time elapsed for model fitting to small (n_s=40), medium (n_s=160), and large (n_s=623) species communities using HMSC Gibbs sampler with 10,000 MCMC iterations. Panels DEF depict those results being adjusted for the autocorrelation in the samples and visualize the estimated times to obtain 1000 effectively independent samples from the posterior. Panels GHI show predictive performance measured in terms of Tjur R2 for models fitted and panels JKL in terms of deviance. The colors indicate non-spatial models (grey), GP models (black), GPP models with 16, 64, 265 and 1024 knots (gradation of blue from light to deep), the NNGP models with 10 and 20 neighbors (light and dark red). Note that due to very similar results, the red lines often overlap.

methods depends on the size of the species community, I fitted the model to subsets of $n_s =$ 40, 160 and 623 species.

The combination of five sample sizes and three community sizes yielded 15 datasets, which I used to compare the performance of four kinds of models: non-spatial latent factors (non-spatial), Gaussian process-based latent factors (GP), Gaussian predictive process-based latent factors (GPP), and Nearest Neighbor Gaussian process-based latent factors (NNGP). In the GPP model I repeated all analyses with $m =$ 16, 64, 256 and 1024 knots, and in the NNGP model I tested $m =$ 10 and 20 neighbors. For parity in model comparisons, I fixed the number of latent factors to $n_f = 2$ in each of the models. I fitted all models with equal number of MCMC steps, burn-in and thinning, using the same hardware exclusively to enable fair comparison, and characterized the performances of the models in terms of their computational demand and predictive power on the validation set: Tjur's R^2 and deviance (Tjur 2009).

Predictive performance generally increases with model complexity, so that the non-spatial model performs the worst, and the performance of the predictive process improves with the number of knots. Notably, even very coarse approximation of spatial structure with only 16 knots provides a substantial gain in the predictive performance, as compared to the non-spatial model. Quite strikingly, the performances of the GP and both NNGP models are essentially equal and considerably outperformed the GPP model when the number of knots is lower than number of training points.

The computational times needed for a single Gibbs update step in the models are consistent with the corresponding theoretical expectations: the computational time increases linearly with sample size in non-spatial and GPP models, and the cubic scaling in the full Gaussian model makes it infeasible for applications with large data. My results indicate that the computational burden of NNGP is in line with GPP with 16-64 knots, scaling slightly worse than linear. On the other hand, the effective sampling size substantially decreases with

increased number of training sites, which seriously aggravates the computational load in practical analysis of large datasets. Nevertheless, my results suggest that this undesired behavior is not due to the spatial structure of the models, but known deficiencies of classic probit data augmentation scheme that may be circumvented in following developments (Duan et al. 2017).

Overall, my results consistently fall in line with analogous findings of the original studies that introduced the GPP and NNGP for modeling spatial low-dimensional outcomes (Banerjee et al. 2008, Datta et al. 2016). The results indicate that, among the models compared here, HMSC augmented with the NNGP performs the best in terms of the trade-off between computational time and predictive performance. However, the fact that NNGP outperforms GPP may be partially due to the nature of the data used in this case study: 1) the spatial range of the latent factors is estimated to be rather small and 2) the spatial distribution of sampling sites in this data is spatially clustered. Both these properties seem to be more in favor of NNGP approximation's nature.

The novelty of methods developed in this Chapter is that they overcome the computational limitations for high-dimensional outcomes using big spatial data, such as the context of species-rich ecological communities. This advance facilitates the efficient use of rapidly accumulating high-resolution large-scale ecological datasets towards explaining and predicting how ecological communities are structured and how they respond to ongoing global change. My implementations of GPP- and NNGP-based latent factors to HMSC Matlab package also allows researchers to integrate such analyses with information on species traits and phylogenetic relationships, providing the potential to address a larger number of fundamental and applied questions in community ecology.

Ecological examples

Microbiota of *M. cinxia* and *P. lanceolata*

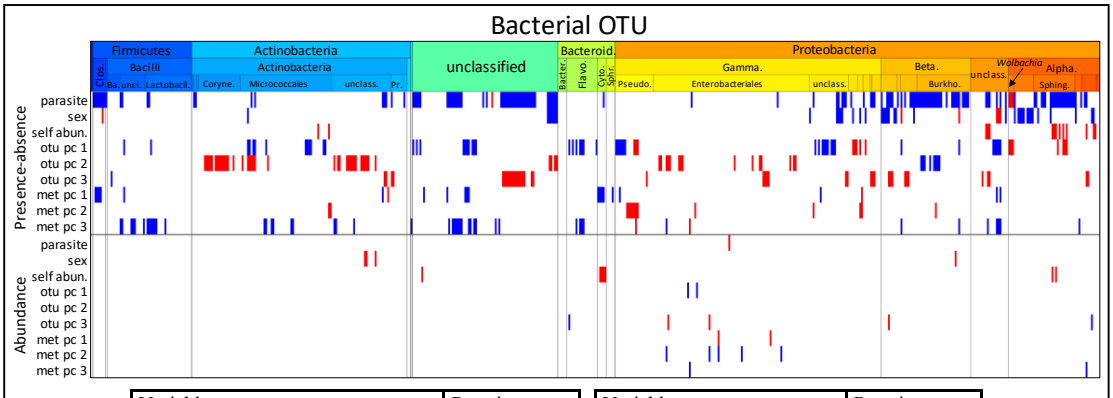
In contrast to other parts of this thesis, Chapter II is organized not around methodological development, but around an applied study of an ecological system. In this study, my colleagues and I pursue to improve the understanding of ecological determinants that influence the associations between insect hosts and their gut symbionts. With *Melitaea cinxia* larvae and their *Plantago lanceolata* host plants sampled across the Åland islands, I aim to assess the structure of midgut microbiota variation across the larvae with respect to available environmental variables and identify the potential drivers of such variation that shape microbiota communities.

I analyzed the data with HMSC model that provides simultaneously species- and community-level inference on how species occurrences and/or abundances relate to environmental covariates, how these relationships are structured with respect to species traits and phylogenetic relationships, and additionally determines those co-occurrence patterns among the species that can't be attributed to responses of the species to the measured covariates. The modeled response variable was the vector of rarified sequence counts of the microbial OTUs. I employed a hurdle approach, in which I first used a probit model for OTU presence-absence, and then a log-normal model for OTU abundances conditional on the presence. In the larval model, my aim was to examine how the OTU composition depended on the properties of the

focal larva, and on the OTU and metabolite compositions of the host plant. I included as fixed effects (1) the sex and (2) the infection status of the individual, (3) the abundance of the focal OTU in the host plant where the individual was residing, (4) the plant OTU community composition, and (5) the plant metabolite composition. I incorporated plant OTU abundance as log-transformed sequence count and described OTU community composition and plant metabolite composition by the first three principal components. To determine to what extent the responses of the species to the explanatory variables show a phylogenetic signal, I included in the analysis a phylogenetic correlation matrix among the OTUs. To examine residual co-occurrence patterns among the OTUs that cannot be attributed to the fixed effects, I further included in the model the level of the larval nest (corresponding to host plant level) as a spatial random effect, and the level of the individual larvae as a non-structured random effect.

I quantified how much of the variation in OTU occurrences can be attributed to the fixed effects and to associations among the OTUs, by evaluating the predictive power of the model in three different ways. All those accounted for the fixed effects but differed on how the random effects were accounted for. Prediction P1 aimed at measuring the predictive power based solely on fixed effects, prediction P2 aimed at measuring the predictive power that can be gained by accounting for species-to-species associations, prediction P3 aimed at measuring the full explanatory power of the model. Thus, the performance of P1 measures the importance of fixed effects, and the difference between P2 and P1 (respectively, between P3 and P1) gives a minimum (respectively, maximum) estimate for the importance of species-to-species associations. I measured predictive powers by Tjur's R^2 (Tjur 2009) for the probit models and standard R^2 for the log-normal models.

Overall, the measured larval microbiota is highly variable both across the larvae individuals and assigned OTUs, with the dominant taxa being Uruburella, Cloacibacterium, Moraxella, Acinetobacter, Dermacoccus, Hymenobacter, Corynebacterium, Paracoccus, Wolbachia, Methylobacterium as well as unclassified Actinobacteria, Enterobacteriaceae and Corynebacteriaceae. The highest prevalence is recorded for an OTU identified as Uruburuella



Variable	Running name
Presence of the *Hyposoter horticola*	parasite
The sex of the larvae	sex
Focal OTU abundance in the plant	self abun.

Variable	Running name
Plant OTU composition	otu pc 1-3
Plant metabolic composition	met pc 1-3

Figure 5. Influence of measured covariates on larval microbiota. Regression coefficients that are estimated to be positive (respectively, negative) with 95% credibility level are shown by red (respectively, blue). The OTUs (columns) are ordered according to their taxonomical classification. The covariates included in the model are listed in the legend alongside with their running names used in axis labelling.

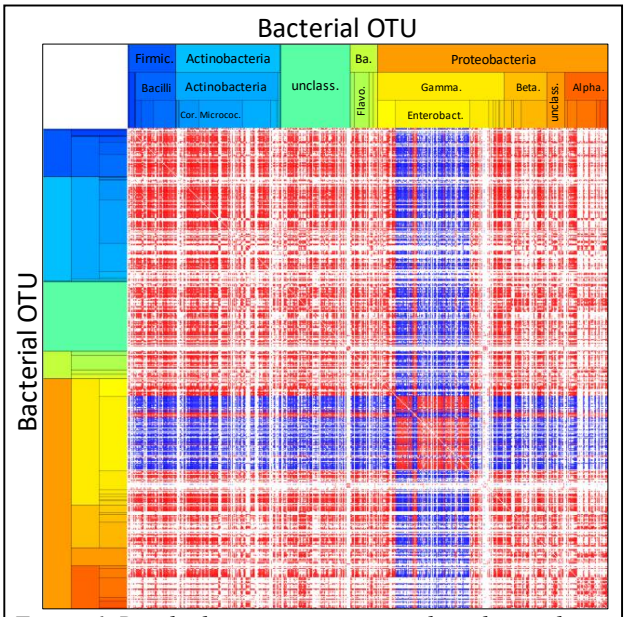


Figure 6. Residual associations among larval microbiota. The figure illustrates the larval-level random effects for the presence-absence part of the larval hurdle model. OTU-pairs for which the residual correlation was estimated to be positive (respectively, negative) with 95% credibility level are shown by red (respectively, blue) color. The ordering of OTUs is identical to that of Figure 5.

which is detected only in 58.8% of the larval samples. Therefore, no core microbiota is evidenced across the larvae.

The presence-absence part of the larval model has only little predictive power through its fixed effects. Accounting for the residual species-to-species associations drastically increases the predictive power, such that it got close to the explanatory power of the fitted HMSC model. This provides evidence that the modeled associations among OTUs represent a true biological signal instead of mere model overfitting. However, the predictive performance is highly variable across the OTUs and remains poor for many species. Contrasting, in the abundance model, both the fixed effects and the species-to-species associations contribute roughly equally. The variance partitioning among the explanatory effects based on the variation of latent variable closely follows the results based on predictive power.

Despite the relatively small contribution of the fixed effects to variance partitioning, substantial proportion of OTUs shows a positive or negative response to many of the fixed effects (Figure 5). The occurrence probabilities of many OTUs decrease with the presence of parasitoid infection and are higher for females than males. Only a minority of the OTUs shows a positive link between the abundance of the focal OTUs in the host plant and the presence of the same OTU in the larvae. In the abundance model a much a smaller proportion of the OTUs shows responses that gains high statistical support (Figure 5). The fitted model estimates a very high phylogenetic signal in how the OTUs respond to the fixed effects for both parts of the model. Hence, closely related OTUs are found to have similar niches and respond similarly to the fixed effects included in the model. The occurrence of the microbial OTUs are phylogenetically structured not only with respect to the measured covariates, but also in their variation that is not attributed to the covariates: the OTUs split into two groups in a markedly pronounced manner (Figure 6). One of these two groups consists, with minor exceptions, of the Enterobacteriaceae family, the other group consists of the remaining OTUs. Given that the majority of explained variation in the presence-absence model is attributed to the random effect at the level of the individual, this pattern is the strongest signal related to OTU occurrences variation in the modeled dataset. In contrast to the strong patterns recorded in the presence-absence model, only few statistically supported associations are found in the abundance model. Concerning the random effect defined at the level of host plants, only few

statistically supported associations are found both for the presence-absence and abundance models.

My findings suggest that the natural midgut microbial community of *M. cinxia* larvae is variable, and that only a small proportion of the variation can be attributed the host plant characteristics that the individual is feeding on. On the other hand, I discovered a strong co-occurrence pattern of OTUs at the level of larval individual that could not be attributed to any of the covariates included to the analyses (Figure 6). These co-occurrence patterns were strongly phylogenetically structured, suggesting two mutually exclusive groups of bacterial communities. One of these groups consisted of mainly *Enterobacteriaceae*, whereas the other group consisted of the remaining taxa. Even though resident microbes may be rare in *Lepidoptera*, the dominance of co-occurring taxa, such as *Enterobacteriaceae* in this study, may be driven by priority effects (dominance of a group of microbes that were the first to colonize the gut), the specific association of bacteria involved in mutualistic interactions, or by a niche overlap among the co-occurring bacteria that grow under similar conditions (Kennedy and Bruns 2005, Sprockett et al. 2018). However, due to the limitation in the available biological material, no absolute quantification (e.g. qPCR) is incorporated in this study. Therefore, the abundances that have been measured are relative and it is not possible to exclude the possibility that the individuals have a uniform microbiota that can be additionally colonized by a very highly-abundant Enterobacteriaceae under certain circumstances.

The microbial variation in neither the presence-absence nor the abundance model could be consistently attributed to the larval or host plant characteristics assessed in a way that would be synchronous across the OTUs. Instead, the effects of the explanatory variables differ greatly among the OTUs, with a high proportion of the variation being explained by the phylogenic relationship between the OTUs (Figure 5). Hence, taxonomically related OTUs responds similarly to the explanatory variables. The occurrence of OTUs belonging to *Rhodobacterales* and *Neisseriales* orders is generally higher in female rather than male larvae. The parasitoid infection is also found to be associated with lower occurrence probability of many taxonomical groups, including *Clostridia*, *Rhizobiales*, *Neisseriales* and *Burkholderiales*. Parasitoid infection may be modifying host's immune/metabolic homeostasis that may further influence the intestinal microbial community. This is possibly because parasitoid infection may modify host's immune or metabolic homeostasis that can further influence the intestinal microbial community (Potter and Woods 2012, Mrinalini et al. 2015). *Wolbachia sp.*, on the contrary, are more likely to occur in parasitized individuals. Previous screening of M. cinxia adults have not found presence of *Wolbachia*, whereas the parasitoid, *H. horticola*, is naturally infected by a *Wolbachia* strain wHho (Duplouy et al. 2015). Thus, the obtained results suggest that Wolbachia could be horizontally transferred by the parasitoid. However, the high mortality of those individuals due to the parasitoid infection might lead to extremely rare infected adults.

Arctic plants

In Chapter III I reanalyze the plant cover data from Mod et al. (2014). In the original study, the authors modeled separately the cover of 17 plant species, including the logarithm of geomorphological disturbance, soil moisture, logarithm of the dominant species' cover (*Empetrum*), and interaction terms as predictors, assuming that soil disturbance and moisture represent stress factors for plant growth.

I modeled all 18 species jointly with a HMSC that included soil moisture, geomorphological disturbance and their interaction as predictors, and allowed the species associations at the sampling unit level to vary with both variables. As the data was collected by hierarchical sampling design, with sampling units located within sites, the model included two levels of random effects to estimate 1) random variation in species occurrences and 2) co-occurrences also at the site level, at which level the associations were assumed to be constant. I predicted the associations $R(\boldsymbol{x}^*)$ at different combinations of disturbance and soil moisture (low and high) and computed the level of statistical support of difference in species associations along the disturbance and soil moisture gradients.

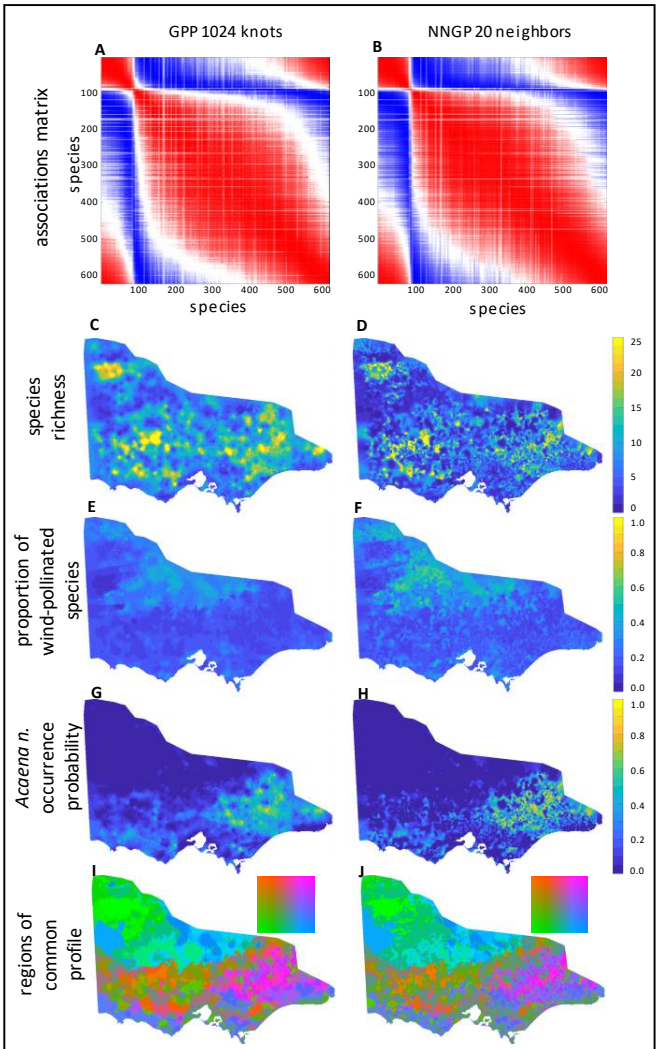


Figure 7. Ecological inference with GPP (left column of panels) and NNGP (right column of panels) models fitted to the maximum training dataset. Panels AB show species association patterns, with red (respectively, blue) depicting species pairs that co-occur more often (respectively, less often) based on the latent factor part of the model. Panels CD visualize predicted spatial distribution of species richness, EF – predicted proportion of wind-pollinated species, GH – predicted occurrence probability of Acaena novae-zelandiae, IJ – predicted regions of common profile, with nodes of 5x5 self-organizing map mapped to YUV color space.

The HMSC-based statistical analysis of the Arctic plant data provides partial support to previous findings of that the plant species associations can be dependent on multiple environmental factors: for many species pairs, the association differed at different combinations of environmental factors, but only for 30% of the species pairs there was high statistical support for a change in the associations along at least one of the environmental gradients. Furthermore, the numbers of species pairs that changes towards more positive and more negative associations are approximately similar for both environmental gradients.

This example demonstrates that the HMSC enhancement proposed in Chapter III can be used to test the SGH using plant community data in an integrated manner. The existing SDM-based method, as in Mod et al. (2014), allows to see only part of the whole picture. However, while SDM method enabled the authors

to relate their results on effect of dominant species to the SGH, considering the species associations among all species within the community, the evidence for SGH based on this data was not clear. I believe that future applications of the method presented in Chapter III, or some related JSDM-based technique, will help to provide more synthetic insights into the SGH in natural communities.

Australian plants

To illustrate the kind of ecological inference that can be derived from the modeling approach presented in Chapter IV, and to highlight the key differences between GPP and NNGP, I use the GPP model with the largest number of knot points ($m = 1024$) and the NNGP model with the largest number of neighbors ($m = 20$) fitted to the entire training partition of Australian plant data ($n_y = 25955$, $n_s = 623$). I visualize the estimates of species associations matrices, predictive distribution maps for individual species, species richness and community weighted mean traits, and divide the study area into regions of common composition profile.

The GPP and NNGP provide essentially identical estimates of species association matrices, revealing numerous positive and negative residual associations (Figure 7). The GPP and NNGP models however differ in their predicted spatial quantities. The NNGP model predicts more fine-scaled patterns and exhibits discontinuities, especially visible in the areas distant from training sites. The GPP model predicts smoother patterns that resembled in some regions the structure of the grid of knots used. Hence, despite its predictive superiority, the NNGP approximation could be criticized by ecologists, since it violates the common understanding that spatial distributions of ecological phenomena are archetypically continuous with potential discontinuities only aligned with distinct habitat edges.

As this example demonstrates, the methods developed in Chapter IV open a great array of possibilities for ecologists working on problems related to fundamental or applied community ecology, conservation biology and macroecology. Most importantly, it is now possible to use spatially extensive data to examine how species occurrences and co-occurrences are associated to environmental variation, how species traits and phylogenies influence such variation, and to generate and validate predictive maps at the levels of species, community composition, and functional traits.

Synthesis, perspectives and conclusions

The ongoing global change, caused by expanding anthropogenic pressure on the environment and biosphere, makes it increasingly important for the humanity to carefully design strategies for sustainable interaction with nature in the upcoming years. To devise robust and successful strategies, we need to acquire a way more comprehensive and integral understanding of the processes and rules that shape the living systems around us. Such understanding at a global scale requires deciphering these underlying processes from the patterns they impose on data that we observe and collect. Currently, due to technical advances of data collection techniques, the amount and quality of scientific data annually obtained keep increasing, and therefore, the availability of analytical methods to exploit the full potential of this data is highly prioritized.

In the last decades, the community ecology has been shifting its aims from the mere description of observed patterns towards a mechanistic perspective, which seeks to understand the processes that shape the observed species communities. These new aims have led to the increased interest in statistical ecology and the development of approaches that jointly model the dynamics and distributions of entire species communities or ecosystems. In particular, joint species distribution models have emerged as efficient tools for modeling data on large numbers of species (Clark et al. 2014, Pollock et al. 2014, Warton et al. 2015a, Ovaskainen et al. 2016a).

This thesis contributes to the ongoing methodological development of analytical tools for the joint species modeling. It combines both practical perspective of an ecologist and methodological/theoretical perspective of a statistician. It starts with synthesis of the recent advances in joint modeling and produce a unifying statistical framework that enables scientists to easily address multiple most common questions in community ecology. Later it provides two important extensions – one more conceptual and on more technical, which considerably expand the potential of the unifying framework. All methodological development is accompanied by examples that use real-world data and demonstrate how these methods can be used in practice.

During my research I've extensively discussed HMSC-related topics both with researchers representing various applied fields and hardcore statisticians (a.k.a. probabilistic machine learning researchers). The following perspectives are based on these discussions and reflect my current vision of the current state of joint species distribution modelling area in general, as well as existing specific challenges for the actual HMSC framework.

The first perspective relates to the need to explicitly account for the observation process when modeling the data, thus allowing one to separate the processes of interest from biases introduced by the observer (Guillera-Arroita et al. 2014, Warton et al. 2016). While currently there are several distributions supported in HMSC framework through known data augmentation schemes, a general approach is still lacking. The four observation functions most desired for ecological research are 1) the ones that would enable to appropriately model presence-only data, 2) multinomial distribution, 3) ordered categorical distribution, and 4) occupancy modeling for repeated samples. Similarly, for the HSMC framework to be robustly applicable for opportunistic spatial ecological data, such as resulted in citizen science projects,

the further development should exploit the advances of statistical literature on modelling non-ignorable sampling designs (Diggle et al. 2010, Gelfand et al. 2012).

The second perspective relates to the need for more versatile treatments of association networks. One of the key strengths of the HMSC approach is that it allows us to estimate species association networks at different spatial or temporal scales, and to utilize the inferred associations in predictions and simulated scenarios. Accompanied by the enhancement of the Chapter III, these associations could additionally vary with respect to environmental factors. However, in all the examples in this Thesis, association matrices were estimated solely from the abundance/occurrence data, even though community ecologists often possess certain a priori knowledge on potential interactions, obtained from direct observations in the field, controlled experiments, or information on how species traits are expected to influence species interactions (Wootton and Emmerson 2005, Schöb et al. 2013). The current HMSC formulation assumes that species traits T and phylogenetic correlations C may directly influence the species responses to the abiotic environment, but ignores the link from these data to prior beliefs on association matrices. Finding sound statistical approaches to incorporate such dependencies in the HMSC framework remains a challenge. A related open question remains on the utilized methodology for variance partitioning – currently HMSC is calculating relative importance of its components based on the latent Gaussian predictor, and therefore it is a subject to the choice of observation model and does not necessarily have an intuitive meaning in the natural scale of observations. To test the sensitivity of such variance partitioning with respect to observation models and link functions, extra numerical experiments are required, which would also preferably be backed up with a qualitative study on how these numerical results correspond to community variation assessment and understanding by ecologists working with various natural systems.

It is also necessary to note that the estimated association networks are conditional on species occurrence – even once two species are estimated to have a strong positive co-occurrence, the association is not truly realized in areas where at least one of the species does not occur at all. This calls for a conceptual revision of how to evaluate species association strengths in a way that would consider variation of species marginal abundances or occurrence probabilities. The shift from the parametric dependence of association matrices on environmental factors as in Chapter III to non-parametric dependence with local shrinkage towards lack of associations (Fox and Dunson 2015) seems to be an attractive option, although could result in excessively cumbersome practicalities.

The third perspective corresponds the ecologists' desire to merge the information on the level of individuals to the framework, as e.g. traits are often measured at the individual level (McGill et al. 2006). A related challenge is to adopt a more micro-evolutionary perspective, e.g. by asking if and how the amount and type of genetic variation influences variation in species occurrence, either among species, or in space or through time.

Forth, for the HMSC to become a robust tool for versatile spatial analysis of species communities, further work is required to expand the available spatial priors for latent factors. While the implemented exponential covariance already provides a substantial improvement compared to the independence assumption, its statistical properties may be undesirable and oversimplistic in many practical applications. First, the anisotropy becomes relevant when the

direction of spatial distances matters, e.g. fish densities can be more correlated with respect to the direction parallel with shore compared to the direction perpendicular to shore (Thorson et al. 2015b). The need for non-stationary covariances arise when the modelled phenomena exhibits various statistical behavior in different parts of the study area/period. As an intuitive example, the species abundance can vary smoothly in flatlands, but very rapidly in mountains. If such variation is not properly captured by fixed effects, the random effects component with stationary assumption for latent factor spatial distribution will fail to properly capture the observed variation.

The fifth and final perspective consists of a broad mixed set of challenges, related to computational and numerical enhancements of the HMSC framework, especially regarding analysis of larger and more comprehensive datasets. The numerical analysis in the Chapter IV indicated that for large datasets HMSC's Gibbs MCMC sampling algorithm does not properly converge within the amount of time that most ecologists consider reasonable to spend on fitting a single statistical model. The exact reason for such misbehavior is not clear yet, although some of the HMSC Gibbs scheme components are known to be potential bottlenecks, e.g. the probit data augmentation scheme of Albert and Chib (1993) has been shown to perform inefficiently for heavily unbalanced presence/absence data (Duan et al. 2017). An appealing solution would be to replace the exact, but slow full-Bayesian MCMC parameter estimation approach with an alternative that would exploit certain approximate solutions in favor of speed and robustness. For example, certain observation models could be efficiently approached with approximate methods (Minka 2001, Rasmussen and Williams 2006, Cunningham et al. 2011). While several JSDMs that exploits these principles has been recently introduced to ecologists (Niku et al. 2017, Thorson 2019), they are based on the maximal likelihood point estimation for the model's fixed effects, which may lead to high sensitivity to model misspecification. However, given the ongoing breakthrough in efficient Bayesian sampling algorithms for seeking solutions to generic probabilistic programming problems with high-dimensional parameter space (Gelman et al. 2013, Hoffman and Gelman 2014, Betancourt et al. 2017), it is worth to investigate whether the combination of approximate methods and proper Bayesian treatment of hyperparameters could be fruitfully exploited in JSDM context.

To conclude, I stress that while the research, presented in this thesis is self-consistent, completed and has already got several applications, is should be seen primarily as a solid starting point for further developments in the field of joint species modeling. Some of these potential developments are related to how more comprehensive ecological questions could be answered with statistical models, while other correspond to the numerical challenges that are posed by new emerging types and amounts of ecological data. I believe that advances and results of my study will enable future research to tackle these challenges and that the JSDM framework will become generally applicable and insightful for a wide array of real-world problems.

Acknowledgements

Finally, I would like to thank those people, who have been around and made a significant impact on me during this road towards the PhD defense:

- First of all, I would like to thank my supervisors – Professor Otso Ovaskainen and Doctor Maria del Mar Delgado. Their supervision, guidance and advice all the way through this four-year journey were crucial for achieving the final goal.
- Next, I express my sincere gratitude to Professor Alan Gelfand for honoring me with accepting the offer to act as an opponent at the defense.
- Pre-examiners Professor Sudipto Banerjee and Doctor James Thorson for their helpful comments and thoughtful feedback on the dissertation. Additional thanks for rigorous adherence to the agreed pre-examination schedule.
- Members of my thesis advisory committee – Professor Anna Kuparinen and Doctor Aleksi Lehikoinen for their professional support and advice.
- LUOVA Doctoral Programme in Wildlife Biology for funding my research, further supporting it with several travel grants and organizing wonderful courses.
- All my co-authors for fruitful collaboration, and particularly Nerea Abrego for great assistance with linking my statistically-oriented research to broad ecological context.
- David Dunson and Leo Duan for hosting my prolonged research visit to Duke University and the United States.
- Jarno Vanhatalo for uncovering the magical world of marginalization and Gaussian processes to me, as well as for other numerous advices and discussions on statistical matters.
- All MRC and REC people for amazing scientific and non-so-scientific environment and especially all my roommates over these years. Further thanks to Viia Forsblom and Bess Hardwick and all other group secretaries for the assistance in handing the practical matters. Huge appreciation to the statistically oriented people – Jukka Siren, Elina Numminen, Panu Somervuo, Malcolm Itter and Torsti Schulz for multiple insightful conversations. And of course, special gratitude to the active HMSC users, namely Anna Norberg, Tad Dallas and Øystein Opedal for keeping poking me in order to make the framework better.
- Turun Metsänkävijät orienteering club, all members of its elite group and especially the elite team manager Juha Suntila for not allowing me to get completely work-crazed during the long dark Finnish evenings.
- Last, but not least, I would like to thank a lot my whole family, which provided invaluable moral support to me during this period. In particular, my mother Elena, father Nikolai and spouse Yulia – I cannot overestimate all your backing and specifically appreciate your immense patience, so that despite of my regular pessimistic grumbling you kept your confidence in me and found the right reasoning to inspire me for pushing harder and aiming higher.

Bibliography

Abrego, N., A. Norberg, and O. Ovaskainen. 2016. Measuring and predicting the influence of traits on the assembly processes of wood-inhabiting fungi. Journal of Ecology **105**:1070-1081.

Agrawal, A. A., D. D. Ackerly, F. Adler, A. E. Arnold, C. Cáceres, D. F. Doak, E. Post, P. J. Hudson, J. Maron, K. A. Mooney, M. Power, D. Schemske, J. Stachowicz, S. Strauss, M. G. Turner, and E. Werner. 2007. Filling key gaps in population and community ecology. Frontiers in Ecology and the Environment **5**:145-152.

Albert, J. H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association **88**:669-679.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang. 2008. Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society. Series B: Statistical Methodology **70**:825-848.

Begon, M., J. L. Harper, and C. R. Townsend. 1996. Ecology: Individuals, populations and communities. Blackwell Science.

Betancourt, M., S. Byrne, S. Livingstone, and M. Girolami. 2017. The geometric foundations of Hamiltonian Monte Carlo. Bernoulli **23**:2257-2298.

Bhattacharya, A., and D. B. Dunson. 2011. Sparse Bayesian infinite factor models. Biometrika **98**:291-306.

Bolker, B., M. Holyoak, V. Křivan, L. Rowe, and O. Schmitz. 2003. Connecting theoretical and empirical studies of trait-mediated interactions. Ecology **84**:1101-1114.

Braak, C. J. F. T., and N. Oct. 1986. Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. Ecology **67**:1167-1179.

Brooker, R. W. 2006. Plant–plant interactions and environmental change. New Phytologist **171**:271-284.

Cadotte, M. W., and C. M. Tucker. 2017. Should environmental filtering be abandoned? Trends Ecol Evol **32**:429-437.

Calabrese, J. M., G. Certain, C. Kraan, and C. F. Dormann. 2014. Stacking species distribution models and adjusting bias by linking them to macroecological models. Global Ecology and Biogeography **23**:99-112.

Callaway, R. M., and L. R. Walker. 1997. Competition and facilitation: a synthetic approach to interactions in plant communities. Ecology **78**:1958-1965.

Chase, J. M., and J. A. Myers. 2011. Disentangling the importance of ecological niches from stochastic processes across scales. Philos Trans R Soc Lond B Biol Sci **366**:2351-2363.

Chen, D., Y. Xue, S. Chen, D. Fink, and C. Gomes. 2016. Deep multi-species embedding. ArXiv e-prints.

Clark, J. S., A. E. Gelfand, C. W. Woodall, and K. Zhu. 2014. More than the sum of the parts: forest climate response from joint species distribution models. Ecological Applications **24**:990-999.

Clark, J. S., D. Nemergut, B. Seyednasrollah, P. J. Turner, and S. Zhang. 2017. Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. Ecological Monographs **87**:34-56.

Clements, F. E. 1936. Nature and structure of the climax. Journal of Ecology **24**:252-284.

Cunningham, J. P., P. Hennig, and S. Lacoste-Julien. 2011. Gaussian probabilities and expectation propagation. ArXiv e-prints.

Datta, A., S. Banerjee, A. O. Finley, and A. E. Gelfand. 2016. On nearest-neighbor Gaussian process models for massive spatial data. Wiley Interdisciplinary Reviews: Computational Statistics **8**:162-171.

Diggle, P. J., R. Menezes, and T. l. Su. 2010. Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics) **59**:191-232.

Dray, S., and P. Legendre. 2008. Testing the species traits-environment relationships: the fourth-corner problem revisited. Ecology **89**:3400-3412.

Duan, L. L., J. E. Johndrow, and D. B. Dunson. 2017. Scaling up data augmentation MCMC via calibration. ArXiv e-prints.

Duplouy, A., C. Couchoux, I. Hanski, and S. van Nouhuys. 2015. Wolbachia infection in a natural parasitoid wasp population. PLOS ONE **10**:1-16.

Elith, J., and J. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics **40**:677-697.

Erland, S., and R. Finlay. 1992. Effects of temperature and incubation time on the ability of three ectomycorrhizal fungi to colonize Pinus sylvestris roots. Mycological Research **96**:270-272.

Finley, A. O., S. Banerjee, and A. E. Gelfand. 2015. spBayes for large univariate and multivariate point-referenced spatio-temporal data models. Journal of Statistical Software **63**:1-28.

Finley, A. O., H. Sang, S. Banerjee, and A. E. Gelfand. 2009. Improving the performance of predictive process modeling for large datasets. Comput Stat Data Anal **53**:2873-2884.

Fox, E. B., and D. B. Dunson. 2015. Bayesian nonparametric covariance regression. Journal of Machine Learning Research **16**:2501-2542.

Franklin, J., J. M. Serra-Diaz, A. D. Syphard, and H. M. Regan. 2017. Big data for forecasting the impacts of global change on plant communities. Global Ecology and Biogeography **26**:6-17.

Garnier, E., M. L. Navas, and K. Grigulis. 2016. Plant functional diversity: organism traits, community structure, and ecosystem properties. Oxford University Press.

Gelfand, A. E., S. K. Sahu, and D. M. Holland. 2012. On the Effect of Preferential Sampling in Spatial Prediction. Environmetrics **23**:565-578.

Gelman, A., H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. Bayesian data analysis. Chapman and Hall/CRC.

Goldberg, D. E., and A. M. Barton. 1992. Patterns and consequences of interspecific competition in natural communities: a review of field experiments with plants. The American Naturalist **139**:771-801.

Golding, N., and B. V. Purse. 2016. Fast and flexible Bayesian species distribution modelling using Gaussian processes. Methods in Ecology and Evolution **7**:598-608.

Gotelli, N. J. 2000. Null model analysis of species co-occurrence patterns. Ecology **81**:2606-2621.

Gotzenberger, L., F. de Bello, K. A. Brathen, J. Davison, A. Dubuis, A. Guisan, J. Leps, R. Lindborg, M. Moora, M. Partel, L. Pellissier, J. Pottier, P. Vittoz, K. Zobel, and M. Zobel. 2012. Ecological assembly rules in plant communities--approaches, patterns and prospects. Biol Rev Camb Philos Soc **87**:111-127.

Graham, C. H., S. Ferrier, F. Huettman, C. Moritz, and A. T. Peterson. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution **19**:497-503.

Gravel, D., C. D. Canham, M. Beaudet, and C. Messier. 2006. Reconciling niche and neutrality: the continuum hypothesis. Ecology Letters **9**:399-409.

Guillera-Arroita, G., J. J. Lahoz-Monfort, D. I. MacKenzie, B. A. Wintle, and M. A. McCarthy. 2014. Ignoring imperfect detection in biological surveys is dangerous: A response to ‘fitting and interpreting occupancy models'. PLOS ONE **9**:1-14.

Guisan, A., and C. Rahbek. 2011. SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. Journal of Biogeography **38**:1433-1444.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters **8**:993-1009.

Hagen, M., W. D. Kissling, C. Rasmussen, M. A. M. De Aguiar, L. E. Brown, D. W. Carstensen, I. Alves-Dos-Santos, Y. L. Dupont, F. K. Edwards, J. Genini, P. R. Guimarães, G. B. Jenkins, P. Jordano, C. N. Kaiser-Bunbury, M. E. Ledger, K. P. Maia, F. M. D. Marquitti, Ó. McLaughlin, L. P. C. Morellato, E. J. O'Gorman, K. Trøjelsgaard, J. M. Tylianakis, M. M. Vidal, G. Woodward, and J. M. Olesen. 2012. Biodiversity, species interactions and ecological networks in a fragmented world. Pages 89-210 *in* U. Jacob and G. Woodward, editors. Advances in Ecological Research. Academic Press.

Harris, D. J. 2015. Generating realistic assemblages with a joint species distribution model. Methods in Ecology and Evolution **6**:465-473.

He, Q., M. D. Bertness, and A. H. Altieri. 2013. Global shifts towards positive species interactions with increasing environmental stress. Ecology Letters **16**:695-706.

Hoff, P. D., and X. Niu. 2012. A covariance regression model. Statistica Sinica **22**:729-753.

Hoffman, M. D., and A. Gelman. 2014. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research **15**:1593-1623.

Holyoak, M., M. A. Leibold, R. D. Holt, and E. S. o. A. Meeting. 2005. Metacommunities: spatial dynamics and ecological communities. University of Chicago Press.

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

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

Ives, A. R., B. Dennis, K. L. Cottingham, and S. R. Carpenter. 2003. Estimating community stability and ecological interactions from time-series data. Ecological Monographs **73**:301-330.

Kennedy, P. G., and T. D. Bruns. 2005. Priority effects determine the outcome of ectomycorrhizal competition between two Rhizopogon species colonizing Pinus muricata seedlings. New Phytol **166**:631-638.

Kraft, N. J. B., P. B. Adler, O. Godoy, E. James, S. Fuller, and J. M. Levine. 2014. Community assembly, coexistence, and the environmental filtering metaphor. Functional Ecology **29**:592-599.

Latimer, A. M., S. Banerjee, H. Sang, Jr., E. S. Mosher, and J. A. Silander, Jr. 2009. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. Ecol Lett **12**:144-154.

Legendre, P., and O. Gauthier. 2014. Statistical methods for temporal and space-time analysis of community composition data. Proc Biol Sci **281**:20132728.

Legendre, P., and L. Legendre. 1998. Numerical Ecology. Elsevier Science Amsterdam, the Netherlands.

Leibold, M. A., and M. A. McPeek. 2006. Coexistence of the niche and neutral perspectives in community ecology. Ecology **87**:1399-1410.

Lewandowski, D., D. Kurowicka, and H. Joe. 2009. Generating random correlation matrices based on vines and extended onion method. Journal of Multivariate Analysis **100**:1989-2001.

Logue, J. B., N. Mouquet, H. Peter, and H. Hillebrand. 2011. Empirical approaches to metacommunities: a review and comparison with theory. Trends Ecol Evol **26**:482-491.

MacDougall, A. S., E. Harvey, J. L. McCune, K. A. Nilsson, J. Bennett, J. Firn, T. Bartley, J. B. Grace, J. Kelly, T. D. Tunney, B. McMeans, S.-I. S. Matsuzaki, T. Kadoya, E. Esch, K. Cazelles, N. Lester, and K. S. McCann. 2018. Context-dependent interactions and the regulation of species richness in freshwater fish. Nature Communications **9**:973.

Maestre, F. T., R. M. Callaway, F. Valladares, and C. J. Lortie. 2009. Refining the stress-gradient hypothesis for competition and facilitation in plant communities. Journal of Ecology **97**:199-205.

Magurran, A. E. 2004. Measuring biological diversity. Wiley.

McGill, B. J., B. J. Enquist, E. Weiher, and M. Westoby. 2006. Rebuilding community ecology from functional traits. Trends in Ecology & Evolution **21**:178-185.

Minka, T. P. 2001. Expectation propagation for approximate Bayesian inference. Pages 362-369 *in* Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc.

Mod, H. K., P. C. Le Roux, and M. Luoto. 2014. Outcomes of biotic interactions are dependent on multiple environmental variables. Journal of Vegetation Science **25**:1024-1032.

Mrinalini, A. L. Siebert, J. Wright, E. Martinson, D. Wheeler, and J. H. Werren. 2015. Parasitoid venom induces metabolic cascades in fly hosts. Metabolomics **11**:350-366.

Niku, J., D. I. Warton, F. K. C. Hui, and S. Taskinen. 2017. Generalized linear latent variable models for multivariate count and biomass data in ecology. Journal of Agricultural, Biological and Environmental Statistics **22**:498-522.

Ojanen, S. P., M. Nieminen, E. Meyke, J. Pöyry, and I. Hanski. 2013. Long-term metapopulation study of the Glanville fritillary butterfly (Melitaea cinxia): survey methods, data management, and long-term population trends. Ecol Evol **3**:3713-3737.

Oldén, A., O. Ovaskainen, J. S. Kotiaho, S. Laaka-Lindberg, and P. Halme. 2014. Bryophyte species richness on retention aspens recovers in time but community structure does not. PLOS ONE **9**:e93786.

Ovaskainen, O., N. Abrego, P. Halme, and D. Dunson. 2016a. Using latent variable models to identify large networks of species-to-species associations at different spatial scales. Methods in Ecology and Evolution **7**:549-555.

Ovaskainen, O., H. J. de Knegt, and M. M. Delgado. 2016b. Quantitative ecology and evolutionary biology: integrating models with data. Oxford University Press.

Ovaskainen, O., J. Hottola, and J. Siitonen. 2010. Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology **91**:2514-2521.

Ovaskainen, O., D. B. Roy, R. Fox, and B. J. Anderson. 2016c. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. Methods in Ecology and Evolution **7**:428-436.

Ovaskainen, O., and J. Soininen. 2011. Making more out of sparse data : hierarchical modeling of species communities. Ecology **92**:289-295.

Ovaskainen, O., G. Tikhonov, D. Dunson, V. Grotan, S. Engen, B. E. Saether, and N. Abrego. 2017. How are species interactions structured in species-rich communities? A new method for analysing time-series data. Proc Biol Sci **284**.

Pickett, S. T. A., and M. J. McDonnell. 1989. Changing perspectives in community dynamics: a theory of successional forces. Trends in Ecology & Evolution **4**:241-245.

Pledger, S., and R. Arnold. 2014. Multivariate methods using mixtures: correspondence analysis, scaling and pattern-detection. Computational Statistics & Data Analysis **71**:241-261.

Poisot, T., A. R. Cirtwill, K. Cazelles, D. Gravel, M.-J. Fortin, and D. B. Stouffer. 2016. The structure of probabilistic networks. Methods in Ecology and Evolution **7**:303-312.

Pollock, L. J., W. K. Morris, and P. A. Vesk. 2012. The role of functional traits in species distributions revealed through a hierarchical model. Ecography **35**:716-725.

Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. Mccarthy. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and Evolution **5**:397-406.

Potter, K. A., and H. A. Woods. 2012. Trichogramma parasitoids alter the metabolic physiology of Manduca eggs. Proceedings of the Royal Society B: Biological Sciences **279**:3572-3576.

Rasmussen, C. E., and C. K. I. Williams. 2006. Gaussian processes for machine learning.:1-248.

Rue, H., A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren. 2017. Bayesian Computing with INLA: A Review. Annual Review of Statistics and Its Application **4**:395-421.

Schöb, C., C. Armas, M. Guler, I. Prieto, and F. I. Pugnaire. 2013. Variability in functional traits mediates plant interactions along stress gradients. Journal of Ecology **101**:753-762.

Sebastián-González, E., J. A. Sánchez-Zapata, F. Botella, and O. Ovaskainen. 2010. Testing the heterospecific attraction hypothesis with time-series data on species co-occurrence. Proceedings of the Royal Society B: Biological Sciences **277**:2983-2990.

Smith, R. L. 1966. Ecology and field biology. New York, USA.

Sprockett, D., T. Fukami, and D. A. Relman. 2018. Role of priority effects in the early-life assembly of the gut microbiota. Nat Rev Gastroenterol Hepatol **15**:197-205.

Stone, L., and A. Roberts. 1990. The checkerboard score and species distributions. Oecologia **85**:74-79.

Thorson, J. T. 2019. Guidance for decisions using the Vector Autoregressive Spatio-Temporal (VAST) package in stock, ecosystem, habitat and climate assessments. Fisheries Research **210**:143-161.

Thorson, J. T., S. B. Munch, and D. P. Swain. 2017. Estimating partial regulation in spatiotemporal models of community dynamics. Ecology **98**:1277-1289.

Thorson, J. T., M. L. Pinsky, and E. J. Ward. 2016. Model-based inference for estimating shifts in species distribution, area occupied and centre of gravity. Methods in Ecology and Evolution **7**:990-1002.

Thorson, J. T., M. D. Scheuerell, A. O. Shelton, K. E. See, H. J. Skaug, K. Kristensen, and D. Warton. 2015a. Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. Methods in Ecology and Evolution **6**:627-637.

Thorson, J. T., A. O. Shelton, E. J. Ward, and H. J. Skaug. 2015b. Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. ICES Journal of Marine Science **72**:1297-1310.

Tjur, T. 2009. Coefficients of determination in logistic regression models—a new proposal: the coefficient of discrimination. The American Statistician **63**:366-372.

Tylianakis, J. M., R. K. Didham, J. Bascompte, and D. A. Wardle. 2008. Global change and species interactions in terrestrial ecosystems. Ecology Letters **11**:1351-1363.

Vanhatalo, J., M. Hartmann, and L. Veneranta. 2018. Joint species distribution modeling with additive multivariate Gaussian process priors and heteregenous data. ArXiv e-prints.

Veech, J. A. 2014. The pairwise approach to analysing species co-occurrence. Journal of Biogeography **41**:1029-1035.

Vellend, M. 2010. Conceptual synthesis in community ecology. Q Rev Biol **85**:183-206.

von Liebig, J. F., and L. P. B. Playfair. 1847. Chemistry in its application to agriculture and physiology. T. B. Peterson.

Warton, D. I., F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. C. Hui. 2015a. So many variables: joint modeling in community ecology. Trends in Ecology and Evolution **30**:766-779.

Warton, D. I., F. G. Blanchet, R. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. C. Hui. 2016. Extending joint models in community ecology: a response to Beissinger et al. Trends in Ecology & Evolution **31**:737-738.

Warton, D. I., S. D. Foster, G. De'ath, J. Stoklosa, and P. K. Dunstan. 2015b. Model-based thinking for community ecology. Plant Ecology **216**:669-682.

Weiher, E., D. Freund, T. Bunton, A. Stefanski, T. Lee, and S. Bentivenga. 2011. Advances, challenges and a developing synthesis of ecological community assembly theory. Philosophical Transactions of the Royal Society B: Biological Sciences **366**:2403-2413.

Williams, R. J., A. Howe, and K. S. Hofmockel. 2014. Demonstrating microbial co-occurrence pattern analyses within and between ecosystems. Front Microbiol **5**:358.

Wisz, M. S., J. Pottier, W. D. Kissling, L. Pellissier, J. Lenoir, C. F. Damgaard, C. F. Dormann, M. C. Forchhammer, J.-A. Grytnes, A. Guisan, R. K. Heikkinen, T. T. Høye, I. Kühn, M. Luoto, L. Maiorano, M.-C. Nilsson, S. Normand, E. Öckinger, N. M. Schmidt, M. Termansen, A. Timmermann, D. A. Wardle, P. Aastrup, and J.-C. Svenning. 2013. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. Biological Reviews **88**:15-30.

Wootton, J. T., and M. Emmerson. 2005. Measurement of interaction strength in nature. Annual Review of Ecology, Evolution, and Systematics **36**:419-444.

Zhou, M., L. Li, D. Dunson, and L. Carin. 2012. Lognormal and gamma mixed negative binomial regression. Proceedings of the International Conference on Machine Learning **2012**:1343-1350.