Choosing the Response Matrix: Generalised Linear Latent Variable Models for Multivariate Ecology and Evolution

Shinichi Nakagawa[1*], Ayumi Mizuno[1], Russell Dinnage[1], Marija Pugar[1], Christine Sosiak[1], Sergio Poo Hernandez[1], Iwo Gross[1], Erick Lundgren[1], Malgorzata Lagisz[1], Eduardo S.A. Santos[1], Santiago Ortega[1]

1 Department of Biological Sciences, University of Alberta, Edmonton, Canada

* email: snakagaw@ualberta.ca

Abstract

Multivariate responses are central to ecology and evolutionary biology, but their covariance is often difficult to model and interpret. Generalised linear latent variable models (GLLVMs) provide a parsimonious way to represent covariance among many responses using a smaller number of latent variables. They are widely used for Site × Species data in joint species distribution modelling and model-based ordination. Here, we argue that their broader value lies in treating the response matrix as an explicit modelling choice. Changing the response matrix changes the biological question: Site × Species, Site × Trait, Individual × Trait, and Species × Trait formulations target different kinds of covariance, even when they use the same basic idea of summarising shared variation with latent variables. We describe the basic structure of GLLVMs, the interpretation of latent variables, loadings, residual covariance or correlation, and communality, which measures how much of a response's variation is shared with other responses. We then use "the fourth-corner problem" to show why the choice of response matrix matters, before developing Unit × Trait as an organising principle for applications in functional biogeography, behavioural syndromes, and phylogenetic trait integration. We conclude that GLLVMs are best viewed not as a method for one data type, but as a general modelling language for multivariate biological covariance, provided that the response matrix, level of inference, and limits of mechanistic interpretation are made explicit.

Introduction

Ecology and evolution are inherently multivariate. We often measure several biological responses on the same observational units: gene expression across tissues, traits across individuals or species, species abundances across sites, or ecosystem functions across communities [1–4]. These measurements are rarely independent. Behaviours may covary because individuals differ consistently in their behavioural tendencies; morphological traits may covary because of shared evolutionary, developmental, or functional constraints; and species may co-occur because they respond to the same environmental gradient. In such cases, covariance is not merely a statistical complication; it is often part of the biological pattern we want to describe and understand.

Modelling this multivariate covariance among measurements creates a practical modelling problem. If we want to model many responses jointly, we need some way to describe how they covary. The most direct approach is to estimate a full covariance matrix among the T response variables (measurements). However, this quickly becomes difficult. A full covariance matrix contains T variances and $T(T-1)/2$ covariances, giving $T(T+1)/2$ variance–covariance parameters in total [3]. Ten responses therefore require 55 variance–covariance parameters; twenty responses require 210 parameters. In many ecological and evolutionary datasets, this is too many parameters to estimate reliably, especially when sample sizes are modest, responses are non-Gaussian, or the model must also account for hierarchical, spatial, temporal, or phylogenetic structure [3, 5, 6]. As a result, researchers often analyse responses one at a time, assume responses are independent, or use dimension-reduction methods (e.g., principal component or factor analysis), which are not directly linked to an observation model for the data [7, 8].

Generalised linear latent variable models (GLLVMs) provide one practical solution. Instead of estimating every pairwise covariance separately, a GLLVM represents shared variation among responses using a smaller number of unobserved dimensions. These dimensions are latent variables: inferred axes of shared variation that are not directly measured, but are estimated as part of the fitted model. In GLLVMs, latent variables describe residual structure among responses after the measured predictors have been included in the model [9–13]. Because GLLVMs are generalised models, they can also be linked to response distributions commonly used in ecology and evolution, including counts, presence–absence data, continuous responses, and ordinal responses. GLLVMs are already familiar in community ecology, especially in joint species distribution models and model-based ordination. In these applications, the response matrix is usually Site $\times$ Species: rows are sites, columns are species, and the latent variables summarise residual structure in species (community) composition [3, 14, 15]. This is a powerful use of GLLVMs, but it is only one possible use. The same logic can be applied whenever we have a multivariate response matrix and want to model shared structure among its columns.

The main point of this article is that the response matrix is itself a modelling choice. Changing the response matrix changes the biological question. Individual $\times$ Trait models ask how traits covary among individuals, and, when repeated observations are available, within individuals. Species $\times$ Trait models ask how traits covary across species. Site $\times$ Trait models ask how trait compositions or functions covary across sites. Site $\times$ Species models ask how species covary across sites. These models may use similar statistical machinery, but they have different units of inference, assumptions, and interpretations. Choosing the response matrix is therefore a biological decision, not only a data-formatting step.

We first describe the fourth-corner problem, which asks how species traits are associated with environmental conditions through patterns of species occurrence [16, 17], to show why response-matrix choice matters. We then explain what a GLLVM is, what

it estimates, and how its main outputs can be interpreted. Finally, we develop Unit × Trait as an organising principle and illustrate it through functional biogeography, behavioural syndromes, and phylogenetic trait integration. Some of these extensions are best viewed as conceptual or emerging applications rather than routine off-the-shelf analyses, because most GLLVM software has been developed primarily for Site × Species data. Our aim is therefore to clarify both what GLLVMs already offer and where further methodological development is needed.

Why the Response Matrix Matters: From the Fourth-Corner Problem to Unit × Trait

The fourth-corner problem provides a useful starting point because it shows that the same biological information can be organised into different matrices, and that the choice of response matrix determines the question being asked. In trait-based community ecology, a common question is whether species with particular traits tend to occur in particular environments. For example, are deep-rooted plant species more common in dry sites, or are large-bodied species more common in productive environments? Answering this question usually involves three observed matrices: a Site × Species matrix describing which species occur at which sites, a Site × Environment matrix describing the conditions at each site, and a Species × Trait matrix describing the traits of each species [16–18].

In the classical fourth-corner setting, the Site × Species matrix is the response. The model asks which species occur where, and whether occurrence or abundance can be explained by environmental variables, species traits, and their interaction [19, 20]. In this formulation, species composition is the outcome, while traits and environments are used to explain that outcome. This is the appropriate formulation when the biological question is about community composition: which species are present, which species are absent, and how species composition changes across environments.

However, the same biological ingredients can support different questions. If the question is not "which species occur where?" but rather "which traits or functions covary across sites?", then the Site × Species matrix is no longer the most direct response. The response may instead be a Site × Trait matrix, where each row is a site and each column is a community-level trait, a trait-state count, a community-weighted mean, or an ecosystem-function measurement such as CO_2 flux. Similarly, if the question is whether behaviours, physiological traits, or morphological traits covary among individuals, the response may be an Individual × Trait matrix. If the question is whether traits covary across species, the response may be a Species × Trait matrix, often with phylogeny incorporated into the covariance structure. Fig. 1 illustrates this broader response-matrix choice.

This distinction matters because the response matrix defines the biological quantity being modelled. A fourth-corner model treats species composition as the multivariate response and asks how traits and environments explain that composition. A Unit × Trait GLLVM instead treats traits, functions, behaviours, or other biological attributes as the multivariate response and asks how these responses covary across a chosen biological or observational unit. The statistical machinery may be similar, but the target of inference is different.

We use *Unit* to mean the biological or observational level at which responses are compared, such as sites, individuals, species, populations, tissues, ecosystems, or communities [1, 21]. We use *Trait* broadly to mean a biological response or attribute, including morphology, behaviour, physiology, gene expression, life-history measures, pathway scores, ecosystem functions, or other biological variables [2, 22–24].

Environmental or contextual variables can enter as predictors, but they are not the multivariate response in a Unit × Trait model.

Under this formulation, rows are units and columns are traits or other biological responses. Each cell must have a clear biological meaning. It may be a direct measurement, a repeated observation, a community-weighted mean, a trait-state count, a pathway score, an ecosystem process rate, or a model-derived estimate [2, 22, 25]. These choices are not interchangeable: they determine the biological quantity being analysed, the appropriate likelihood, and how residual covariance should be interpreted.

The unifying point is not that all Unit × Trait models are identical. They may differ in likelihood, random effects, replication, and spatial, temporal, phylogenetic, or hierarchical structure. Rather, they address a common class of questions: how do multiple biological responses covary across a defined unit of organisation, after accounting for measured predictors? GLLVMs provide a shared latent-variable framework for modelling this covariance once the response matrix and level of inference have been chosen.

What Is a GLLVM?

Once the response matrix has been defined, a GLLVM provides one way to model how its columns covary. The basic idea is simple: rather than estimating a separate covariance for every pair of responses, the model asks whether some of this covariance can be summarised by a smaller number of latent dimensions. These latent dimensions are not directly observed. They are estimated as part of the fitted model and represent shared structure among responses after measured predictors have been included.

A useful way to read a GLLVM is from left to right, as in Fig. 2. We begin with a response matrix, add measured predictors, use these to build a linear predictor, and then use latent variables to summarise any remaining covariance among the responses. Thus, a GLLVM combines two familiar ideas. Like a generalised linear model, it relates responses to measured predictors using an appropriate link function and response distribution. Like factor analysis or ordination, it uses a small number of latent dimensions to describe covariance among many responses [9, 10, 26].

Consider a response matrix $\mathbf{Y}$ with n units and T responses:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1T} \\ y_{21} & y_{22} & \cdots & y_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nT} \end{bmatrix}.$$

Rows are the units of comparison, such as sites, individuals, species, tissues, or ecosystems. Columns are the responses whose covariance we want to model. The entry y_{it} is the observed value of response t for unit i.

A GLLVM does not model $\mathbf{Y}$ directly on the observed scale. Instead, each observation has an expected value, $\mu_{it} = E(y_{it})$, and this expected value is connected to a linear predictor, η_{it}, through a response-specific link function:

$$g_t(\mu_{it}) = \eta_{it}.$$

The link function is simply the translation between the scale of the data and the scale on which the model is additive. For a Gaussian response with an identity link, $\eta_{it} = \mu_{it}$. For a binary response with a logit link, $\eta_{it} = \log\{\mu_{it}/(1-\mu_{it})\}$, where μ_{it} is the probability of occurrence. For a count response with a log link, $\eta_{it} = \log(\mu_{it})$. Thus, η_{it} is not the observed value itself; it is the expected value expressed on the model scale.

For all units and responses together, the linear predictor can be written in matrix form as:

$$\underbrace{\boldsymbol{\eta}}_{n\times T} = \underbrace{\mathbf{X}}_{n\times p}\underbrace{\mathbf{B}}_{p\times T} + \underbrace{\mathbf{Z}}_{n\times d}\underbrace{\boldsymbol{\Lambda}^\top}_{d\times T}.$$

This equation is the algebraic version of the middle part of Fig. 2. The first term, $\mathbf{XB}$, is the measured-predictor part of the model. The matrix $\mathbf{X}$ contains observed predictors, such as environmental variables, treatments, time, body size, or other covariates. The matrix $\mathbf{B}$ contains the response-specific regression coefficients. This part asks how each response changes with measured predictors.

The second term, $\mathbf{Z}\boldsymbol{\Lambda}^\top$, is the latent-variable part of the model. The matrix $\mathbf{Z}$ contains the latent scores for the units. These scores place each unit on a small number of latent dimensions. The matrix $\boldsymbol{\Lambda}$ contains the response loadings. These loadings describe how strongly each response is associated with each latent dimension. The number of latent dimensions is d, and in most applications d is much smaller than the number of responses T; for example, one to three latent dimensions may be used to summarise covariance among many responses [11].

For readers who prefer cell-wise notation, the same model for one unit i and one response t is:

$$\eta_{it} = \mathbf{x}_i^\top \boldsymbol{\beta}_t + \mathbf{z}_i^\top \boldsymbol{\lambda}_t.$$

Here, $\mathbf{x}_i$ is the vector of measured predictors for unit i, and $\boldsymbol{\beta}_t$ is the vector of regression coefficients for response t. Similarly, $\mathbf{z}_i$ is the vector of latent scores for unit i, and $\boldsymbol{\lambda}_t$ is the vector of loadings for response t. Responses with similar loading vectors tend to covary positively; responses with opposite loading vectors tend to covary negatively; and responses with weak or unrelated loadings tend to covary little.

The Gaussian identity-link case gives the clearest connection between latent variables and covariance. In this case, the observed response can be written as:

$$y_{it} = \eta_{it} + \epsilon_{it},$$

where ϵ_{it} is the residual deviation for unit i and response t. This residual is the part of the observation not explained by the measured predictors or by the shared latent variables.

The model-implied residual covariance among responses is:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}.$$

This decomposition is central to the interpretation of a GLLVM. The term $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ is the shared covariance generated by the latent variables. The diagonal matrix $\boldsymbol{\Psi}$ contains response-specific unique variances: variation in each response that is not shared with the other responses.

For example, suppose we have $T = 5$ responses and $d = 2$ latent variables. The loading matrix is then:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \end{bmatrix}.$$

Each row is a response, and each column is a latent variable. The shared covariance between two responses, say response t and response r, is the inner product of their loading vectors:

$$\left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)_{tr} = \boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_r = \sum_{k=1}^{d} \lambda_{tk}\lambda_{rk}.$$

Thus, responses with similar loadings covary positively, responses with opposite loadings covary negatively, and responses with weak or unrelated loadings covary little.

The unique-variance matrix is diagonal:

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & 0 & 0 & 0 & 0 \\ 0 & \psi_2 & 0 & 0 & 0 \\ 0 & 0 & \psi_3 & 0 & 0 \\ 0 & 0 & 0 & \psi_4 & 0 \\ 0 & 0 & 0 & 0 & \psi_5 \end{bmatrix}.$$

Because $\boldsymbol{\Psi}$ is diagonal, it contributes to the variance of each response but not to covariance between different responses. Therefore, $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ creates the shared covariance among responses, whereas $\boldsymbol{\Psi}$ adds response-specific variance to the diagonal of $\boldsymbol{\Sigma}$.

For a single response t, the diagonal entry of the covariance matrix is:

$$\boldsymbol{\Sigma}_{tt} = \left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)_{tt} + \boldsymbol{\Psi}_{tt}.$$

The first term is the part of response t's variance shared with other responses through the latent variables. The second term is the response-specific variance not shared with other responses. This is the basis for communality and uniqueness, introduced below.

For non-Gaussian responses, the same intuition applies, but the covariance is usually interpreted on the link, latent, or distribution-specific scale rather than as ordinary Gaussian residual variation on the original response scale. This distinction is important. A latent variable is not automatically a real environmental gradient, behavioural syndrome, or evolutionary process. It is first a statistical summary of residual covariance among responses after measured predictors have been included.

What Does a Fitted GLLVM Give You?

A fitted GLLVM gives several complementary outputs. Some of these look like familiar regression outputs, whereas others summarise multivariate covariance. These outputs should not be interpreted in isolation, because each describes a different part of the same fitted model.

First, a fitted GLLVM gives regression coefficients for measured predictors. These describe how each response is associated with environmental, temporal, experimental, biological, or other contextual variables. This part of the model is similar to a standard regression, except that multiple responses are analysed jointly and the model also accounts for residual covariance among them.

Second, the model gives the fitted latent-variable structure. This structure has two connected parts: scores for the units and loadings for the responses. Scores place sites, individuals, species, or other units along the fitted latent dimensions. Plotting these scores gives a model-based ordination: units that are close together have similar fitted multivariate profiles. Loadings describe how strongly each response is associated with each latent dimension. In ordination biplots, responses are often shown as arrows. Responses pointing in similar directions tend to covary positively, whereas those pointing in opposite directions tend to covary negatively. Longer arrows indicate stronger associations. These outputs are shown schematically in Fig. 3.

The same fitted latent structure can also be summarised as a model-implied residual covariance or correlation matrix. These matrices are often easier to interpret than the latent axes themselves, because latent axes are not unique: their signs can flip, and rotations can give equivalent covariance structures [26, 27]. In practice, loadings are often rotated before visualisation to obtain a simpler loading pattern, but such rotation does not change the model-implied covariance structure. For this reason, interpretation

should not rely only on naming individual axes; it should also examine the covariance or correlation matrix implied by the model.

A fourth output is a response-level summary called *communality*. Communality answers a simple question: for a given response, how much of its model-implied variance is shared with other responses through the latent structure? This is useful because a correlation matrix tells us about pairwise relationships, but it does not directly tell us whether a response is deeply embedded in the broader multivariate pattern or mostly varies on its own.

For Gaussian models, the communality of response t can be written as:

$$c_t^2 = \frac{\left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)_{tt}}{\left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)_{tt} + \boldsymbol{\Psi}_{tt}}.$$

Here, the subscript tt means the diagonal entry for response t. Thus, $\left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)_{tt}$ is the amount of variance in response t explained by the shared latent structure, and $\boldsymbol{\Psi}_{tt}$ is the unique variance of response t that is not shared with other responses. The denominator is therefore the total model-implied residual variance of response t. Communality is the proportion of this variance explained by the shared latent structure. The complement, $1 - c_t^2$, is the uniqueness proportion [26].

High communality means that a response strongly participates in the shared multivariate structure. Low communality means that much of the response's variation is response-specific, at least relative to the other responses included in the model. The biological interpretation therefore depends on the response matrix. In a Site × Species model, a high-communality species is strongly aligned with shared community structure, such as a latent environmental or compositional gradient. In a Site × Trait model, a high-communality trait or function is strongly embedded in site-level functional covariance. In an Individual × Trait model, high communality indicates that a behaviour, physiological trait, or morphological trait participates strongly in an among-individual syndrome or other integrated (correlated) phenotype. In a Species × Trait model, high communality indicates that a trait is strongly embedded in a cross-species integration axis, possibly after accounting for phylogeny. Conversely, high uniqueness indicates more response-specific variation: a species, trait, behaviour, or function that is less well represented by the shared latent dimensions.

In repeated-measures designs, communality can be calculated for different covariance components. For example, between-individual communality describes how strongly a response participates in stable among-individual integration, whereas within-individual communality describes how strongly a response participates in coordinated change across contexts, time, or state [28, 29]. A behaviour may therefore be highly integrated among individuals but weakly integrated within individuals, or the reverse.

In practice, GLLVM outputs should be reported together: predictor effects, ordination, loadings, covariance or correlation structure, communality, uniqueness, and sensitivity to the number of latent variables. Together, these outputs describe the fitted multivariate pattern. Their biological interpretation, however, depends on the response matrix defined at the start of the analysis, because the same statistical output can have different meanings under Site × Species, Site × Trait, Individual × Trait, and Species × Trait formulations.

Three Applications: Behaviour, Ecology, and Evolution

The same GLLVM logic can be adapted across many areas of biology. Here, we use three examples to show how changing the response matrix changes the biological question: behavioural syndromes, functional biogeography, and phylogenetic trait integration. Table 1 places these examples in a broader sequence of Unit × Trait formulations, from molecular and organismal levels to ecosystems and landscapes.

Table 1. Examples of Unit × Trait GLLVM formulations. The response matrix defines the biological quantity being estimated. The latent structure should be interpreted as a summary of covariance at the chosen level, not as direct evidence of mechanism.

Unit	Possible entries in the response matrix	Main question	Possible interpretation of latent structure
Tissue, cell type, or biological sample	Gene or transcript expression, protein abundance, metabolite abundance, pathway scores, or cell-state markers	Do molecular or functional responses covary across tissues, cell types, or samples?	Regulatory modules, stress-response axes, differentiation, or cell-state dimensions
Individual	Behavioural, physiological, morphological, or performance measures, often repeated across time or context	Do traits covary among individuals, within individuals, or both?	Behavioural syndromes, physiological integration, or coordinated plasticity
Population	Mean traits, plasticity estimates, regulatory traits, or performance measures	Do populations differ in suites of traits or responses?	Local adaptation, plasticity syndromes, or population-level integration
Species	Morphological, life-history, physiological, behavioural, or molecular traits	How do traits covary across species, possibly after accounting for phylogeny?	Phylogenetic trait integration, life-history axes, or residual trait covariance
Host or microhabitat	Microbiome functions, enzyme activities, functional guilds, or abundance-weighted functional summaries	Do microbial functions covary across hosts or environments?	Nutrient-use, host-associated, decomposition, or metabolic axes
Site or plot	Community-weighted means, abundance-weighted totals, trait-state counts, or ecosystem function rates	How do traits or functions covary across environmental gradients?	Functional composition, ecosystem functioning, spatial structure, or unmeasured environmental gradients
Ecosystem	Productivity, decomposition, carbon flux, nutrient cycling, or other process rates	Do ecosystem functions covary across disturbance or environmental gradients?	Carbon-cycling, nutrient-cycling, recovery, or disturbance-response axes
Landscape	Disease-related traits, host traits, vector traits, pathogen measures, or exposure summaries	Do disease-relevant traits covary spatially?	Transmission-risk, susceptibility, resistance, or exposure axes

Behavioural syndromes: Individual × Trait

In behavioural ecology, the unit can be an individual, and the responses can be behavioural, physiological, morphological, or performance traits. An Individual × Trait

GLLVM then asks how these traits covary among individuals. When the focus is animal personality or behavioural syndromes, the main covariance of interest is usually *between-individual covariance*: stable differences among individuals in their average trait values [4, 28, 30, 31]. For example, the model can ask whether individuals that are consistently more aggressive are also consistently more fearful, active, or trainable.

Box 1 illustrates this between-individual version using C-BARQ behavioural scores from dogs. In that example, rows are individual dogs and columns are behavioural categories. The fitted latent variables summarise major axes of stable behavioural integration among dogs, while the model-implied correlation matrix and communality values show which behaviours are most strongly embedded in the shared behavioural structure.

A full multivariate mixed model can estimate a between-individual covariance matrix directly, but the number of parameters increases quickly as the number of traits grows [32]. A reduced-rank Individual × Trait model provides a parsimonious alternative by summarising the main covariance pattern with a small number of latent dimensions. Communality can then be used to ask how strongly each behaviour participates in the shared syndrome: high communality indicates strong integration with the other behaviours in the model, whereas low communality indicates more behaviour-specific variation.

With repeated observations, this framework can be extended to separate between-individual covariance from within-individual covariance, that is, coordinated deviations of an individual from its own expected phenotype across time, context, or state [4, 33]. This extension is useful for studying plasticity or state-dependent change, but it requires enough repeated observations and careful attention to measurement error, observer effects, session effects, batch effects, or assay conditions [28, 29]. Without repeated observations, an Individual × Trait model can describe phenotypic covariance among individuals, but it cannot separate stable individual differences from within-individual plasticity.

Functional biogeography: Site × Trait

In functional biogeography, the unit is often a site, plot, community, or region, and the responses are traits or functions measured or summarised at that level [1, 34–36]. A Site × Trait matrix may contain community-weighted means, abundance-weighted trait totals, trait-state counts or proportions, ecosystem function rates, or other site-level summaries [2, 22, 25]. These choices are not interchangeable: a community-weighted mean targets average functional composition, a trait-state count targets the frequency of trait categories, and an ecosystem process rate targets function rather than organismal phenotype.

Repeated surveys or plot censuses should be represented according to the biological question. If the aim is to compare long-term average functional composition among sites, repeated observations can be summarised to a single Site × Trait matrix. If the aim is to study temporal change, each row can instead represent a site-by-time or plot-by-time observation. A fuller hierarchical GLLVM can also retain repeated observations nested within sites, plots, years, or sampling events. Similarly, when the raw data begin as a Plot × Species matrix, one can either convert each plot or plot-time sample into trait summaries and analyse a Plot × Trait matrix, or retain species composition in a model closer to a fourth-corner formulation. These choices answer different questions.

A Site × Trait GLLVM can model several site-level traits or functions jointly. Environmental variables such as rainfall, temperature, soil nutrients, disturbance, habitat structure, or land-use history can enter as predictors. The latent variables then summarise residual covariance among traits or functions across sites after these predictors have been accounted for, and communality describes how much each trait or

function shares residual variation with the others. This interpretation should remain cautious, because residual covariance may reflect unmeasured environmental gradients, spatial autocorrelation, aggregation choices, detectability, sampling design, or temporal structure [8, 37].

Phylogenetic trait integration: Species × Trait

In macroevolution, the unit is often a species or higher taxonomic unit, and the responses are morphological, physiological, life-history, behavioural, or molecular traits. The question is how traits covary across species, and whether this covariance is structured by shared evolutionary history. Species usually cannot be treated as independent observations, because closely related species often resemble one another due to shared ancestry [38–43].

A Species × Trait GLLVM can include phylogenetic and non-phylogenetic components of covariance. The phylogenetic component describes covariance among traits that follows the expected similarity among species under an assumed phylogenetic covariance structure. The non-phylogenetic component describes remaining species-level covariance not explained by that phylogenetic structure. This residual component is not necessarily "noise": it may reflect shared ecology, convergence, missing predictors, sampling variation, measurement error, or other processes operating at the tips of the phylogeny [43, 44].

Latent variables can provide a reduced-rank summary of high-dimensional covariance within one or both components. For example, a phylogenetic latent variable might summarise a trait-integration axis aligned with shared ancestry, whereas a non-phylogenetic latent variable might summarise coordinated trait deviations among species that are not explained by the tree [44–46]. Interpretation must remain cautious: covariance aligned with phylogeny may reflect shared ancestry, evolutionary constraint, conserved ecology, or other historically structured processes, and the model does not distinguish among these mechanisms on its own. Likewise, non-phylogenetic covariance is best interpreted as residual species-level covariance conditional on the phylogeny and measured predictors, not automatically as adaptation or convergence.

In many comparative datasets, each species contributes a single value per trait. In that setting, phylogenetic and non-phylogenetic covariance can be modelled at the species level, but separating species-level residual covariance from within-species variation, measurement error, or uncertainty in species means requires additional information or explicit modelling assumptions [44]. Replicate measurements within species, uncertainty estimates for species means, or hierarchical models can help separate these sources when such information is available.

Common structure, different biological quantities

These examples use the same latent-variable logic, but they target different biological quantities. In behavioural ecology, the main target may be stable covariance among individuals, with within-individual covariance as an extension when repeated observations are available. In functional biogeography, the target is covariance among traits or functions across sites, plots, or site-time units. In macroevolution, the target is covariance among species traits, often partitioned into phylogenetic and non-phylogenetic components.

The value of GLLVMs is therefore not that one model can be used unchanged everywhere. Rather, the same modelling language can be adapted once the response matrix, unit of inference, observation model, and covariance level have been defined. This is why choosing the response matrix is a biological decision as much as a statistical one.

Implementation, Assumptions, and Scope

Implementation depends on the response matrix, response distribution, and covariance structure. Several `R` packages support GLLVMs or closely related joint-modelling approaches, but they differ in scope. The `gllvm` package is purpose-built for GLLVMs, model-based ordination, and joint species distribution models [11, 13, 47–49]. Broader latent-variable and mixed-model frameworks include `galamm` and `glmmTMB`, which can be useful when smooth terms, mixed responses, hierarchical designs, or reduced-rank random-effect structures are needed [50–52]. Bayesian approaches such as `Hmsc` and `boral` are useful when hierarchical structure, latent factors, or full posterior inference are important [15, 53, 54]. Other packages, including `sjSDM`, `gjam`, and `mvabund`, provide related joint-modelling or multivariate regression approaches [20, 55, 56]. Table 2 gives examples of packages that can be useful for GLLVMs, latent-variable models, or related joint-modelling approaches.

An important practical caveat is that current GLLVM methodology and software have been developed largely for Site × Species matrices in community ecology. Extensions to other response matrices are not always available in a form that matches the biological problem exactly. For example, a functional biogeography model may require spatial covariance at the site level, whereas a Species × Trait model may require phylogenetic covariance to be placed at the correct level of the model. These structures are not always straightforward to combine with latent-variable covariance among traits in existing software. Custom Bayesian modelling in `Stan` offers a promising route, as it allows users to specify the likelihood, latent-variable structure, phylogenetic or spatial covariance, and hierarchical levels directly [57, 58]. However, such models require careful implementation, prior specification, convergence assessment, and posterior predictive checking. Thus, some Unit × Trait applications should currently be treated as promising extensions rather than standard analyses.

A practical workflow is to start with the simplest model that matches the biological question. First, define the response matrix: what are the rows, what are the columns, and what does each entry represent? Second, specify predictors and choose an observation model appropriate for the response scale. Third, fit a small set of models, usually beginning with one or two latent variables. Additional latent variables can be added if they improve fit, reduce residual structure, and remain interpretable. Model comparison should not rely on a single criterion; information criteria, cross-validation, residual diagnostics, ordination plots, and sensitivity to the number of latent variables should be considered together [8, 59].

Model complexity should reflect the study design. Individual × Trait analyses can describe among-individual covariance from one observation per individual, but repeated observations are needed to separate between- and within-individual covariance [28]. Site × Trait analyses require well-defined site-level responses and adequate replication across sites or site-time units. Species × Trait analyses require an explicit treatment of phylogenetic or taxonomic non-independence [38, 42]. Spatial, temporal, phylogenetic, and hierarchical structures should be added only when they are relevant to the question and supported by the data, because each component changes the covariance being estimated.

Missing data require care. Likelihood-based GLLVMs can often accommodate incomplete response matrices without prior imputation, provided that the observation model and missing-data assumptions are appropriate. However, missingness is not automatically benign. If it depends on the unobserved response, sampling process, detectability, or phylogenetic coverage, sensitivity analyses or explicit missing-data models may be needed [44, 60].

GLLVMs are most useful when the goal is to describe multivariate covariance and the number of responses is large enough that an unstructured covariance matrix would

be difficult to estimate. This difficulty can be judged by comparing the number of responses with the number of covariance parameters: an unstructured covariance matrix for T responses requires $T(T+1)/2$ variance–covariance parameters, so 10 responses require 55 parameters and 20 responses require 210. In such cases, a low-dimensional latent-variable structure can provide a more parsimonious summary of residual covariance. GLLVMs are less useful when there are too few units to support even a reduced-rank model, when the response matrix is poorly defined, when residual covariance is weak, or when prediction alone is the main goal and covariance interpretation is secondary. In such cases, simpler alternatives, such as univariate models, standard multivariate regression, classical ordination, phylogenetic comparative methods, or structural equation models, may be more appropriate [7, 8, 27, 61].

Interpretation, Identifiability, and Related Methods

Latent variables are statistical summaries, not mechanisms. They are introduced because they help describe covariance among responses, not because they necessarily represent causal processes. A latent axis may be interpreted as "behavioural syndrome", "functional integration", or "life-history axis" when the loading pattern and study design support that interpretation, but the label is not evidence that the latent axis generated the observed variation. Residual covariance may arise from unmeasured environmental variables, spatial or temporal structure, phylogenetic history, measurement error, aggregation choices, or model misspecification [37, 44, 62].

Latent axes are also not uniquely identified. Their signs can change, and rotations or alternative parameterisations of the loading matrix can give equivalent fitted covariance structures [26, 27]. Interpretation should therefore not rely only on the naming of individual axes. Loadings are useful for describing how responses relate to the fitted latent dimensions, especially when uncertainty can be reported and identification constraints are clear. However, more stable summaries include the model-implied covariance or correlation matrix, communality and uniqueness, and sensitivity to the number of latent variables [48, 59].

GLLVMs are part of a broader family of multivariate methods. Classical ordination methods remain useful for exploration and visualisation, but are not always tied to an explicit model for the observed data [7, 8]. Full multivariate mixed models estimate covariance directly, but can become difficult to fit as the number of responses increases [32]. Joint modelling frameworks such as `Hmsc`, `boral`, `gllvm`, and related packages differ in implementation, estimation, and scope, but share the general aim of modelling multiple responses together, often using latent variables to summarise residual covariance [15, 47, 53, 54]. The value of GLLVMs in the present context is therefore not that they replace these approaches, but that they provide a clear language for linking the chosen response matrix, measured predictors, and shared covariance structure.

Conclusions and Future Perspectives

Our central point is that GLLVMs are not tied to a single data type. Their broader value lies in treating the response matrix as an explicit modelling choice. Individual × Trait, Site × Trait, Site × Species, Species × Trait, and other Unit × Trait formulations ask different biological questions, require different assumptions, and target different aspects of covariance. What they share is the idea that covariance among many responses can often be summarised by a smaller number of latent dimensions plus response-specific variation [3, 9, 13].

Future developments could make Unit × Trait GLLVMs easier to fit with the

covariance structures required by different biological questions. For behavioural ecology, this means repeated-measures models that can separate between- and within-individual covariance while accounting for measurement, observer, or context effects. For functional biogeography, this means models that can combine latent covariance among traits or functions with spatial, environmental, and hierarchical structure at the site or landscape level. For phylogenetic trait integration, this means models that can place phylogenetic covariance at the correct biological level while also estimating shared latent structure among traits. More generally, promising directions include repeated-measures GLLVMs, spatial and phylogenetic latent-variable models, explicit treatment of measurement error, dynamic latent-variable models for time series, and improved diagnostics for choosing the number of latent variables [47, 49]. These developments will be most useful when they remain grounded in the central principle that the response matrix should match the biological question.

In summary, GLLVMs provide a flexible framework for modelling covariance among multiple biological responses when full covariance structures are impractical. Used carefully, they can help ecology and evolutionary biology move from asking whether responses are correlated to asking where, at which biological level, and after accounting for which predictors that covariance arises. The key step is not only choosing a model, but choosing the response matrix that matches the biological question.

Acknowledgments

This work was supported by a Canada Excellence Research Chair (CERC-2022-00074).

Conflict of Interest

Authors declare no conflict of interest.

Table 2. Examples of `R` packages for GLLVMs, latent-variable models, and related joint-modelling approaches. The table is illustrative rather than exhaustive. Not all packages fit GLLVMs in the strict sense; some provide related latent-variable, generalized additive latent, hierarchical community, or joint-modelling frameworks. Package capabilities change over time, so users should consult current documentation when choosing software.

Package	Model class	Main strengths	Key reference
`gllvm`	GLLVM, directly	Purpose-built for generalized linear latent variable models, model-based ordination, joint species distribution models, constrained/concurrent ordination, hierarchical designs, and phylogenetic random effects	[11, 13, 47, 48]
`galamm`	Generalized additive latent and mixed model	Latent-variable and mixed-model framework with smooth terms, crossed random effects, mixed responses, and factor-analytic structures; useful when latent variables, hierarchical structure, and nonlinear covariate effects are all relevant	[50, 51]
`glmmTMB`	GLMM with reduced-rank structures	Flexible mixed-model syntax for many response families and random-effect structures; useful when reduced-rank random effects are embedded in repeated-measures or hierarchical designs	[32, 52]
`Hmsc`	Bayesian hierarchical joint species distribution model	Joint community modelling with environmental covariates, species traits, phylogeny, spatial and temporal structure, and latent factors	[15, 54]
`boral`	Bayesian latent-variable ordination/regression	Bayesian ordination and regression analysis for multivariate abundance or community data, including latent-variable formulations	[53]
`sjSDM`	Scalable joint species distribution model	Fast joint species distribution modelling for large community datasets, especially presence–absence or abundance-type responses	[56]
`gjam`	Generalized joint attribute model	Joint modelling of mixed response types with direct covariance modelling rather than a standard GLLVM loading-matrix formulation	[55]
`mvabund`	Model-based multivariate regression	Multivariate abundance and presence–absence modelling for hypothesis testing, including trait–environment and abundance models; not a latent-variable GLLVM package	[20]

Figures

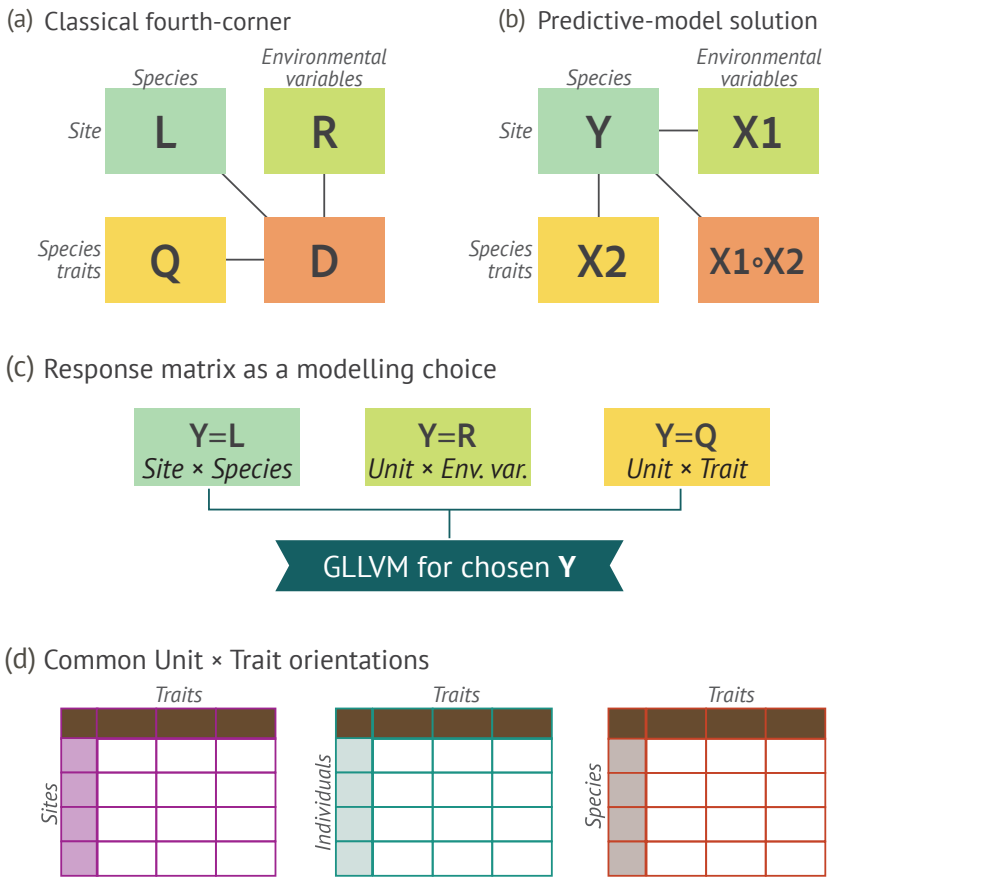

Fig 1. From the fourth-corner problem to Unit × Trait GLLVMs. (a) The classical fourth-corner problem in trait-based community ecology links three observed matrices: the Site × Species composition matrix $\mathbf{L}$, the Site × Environment matrix $\mathbf{R}$, and the Species × Trait matrix $\mathbf{Q}$. The fourth corner, $\mathbf{D}$, represents the inferred association between species traits and environmental variables. (b) In predictive fourth-corner models, species composition is treated as the response matrix ($\mathbf{Y} = \mathbf{L}$). Environmental variables ($\mathbf{X}_1 = \mathbf{R}$), species traits ($\mathbf{X}_2 = \mathbf{Q}$), and their interaction ($\mathbf{X}_{\text{int}}$) are used as predictors to explain which species occur where. (c) This example shows that the response matrix is a modelling choice. Depending on the biological question, $\mathbf{Y}$ may be a Site × Species matrix, a Unit × Environmental-variable matrix, or a Unit × Trait matrix. Changing $\mathbf{Y}$ changes what is being modelled. (d) Unit × Trait formulations keep the same row-by-column logic but change the biological unit. Rows may be sites, individuals, species, populations, ecosystems, tissues, or other units; columns are traits, functions, behaviours, physiological measures, or other biological responses. A GLLVM can then be used to model covariance among the columns of the chosen response matrix. Panels (a) and (b) are redrawn from Brown (2014) [19].

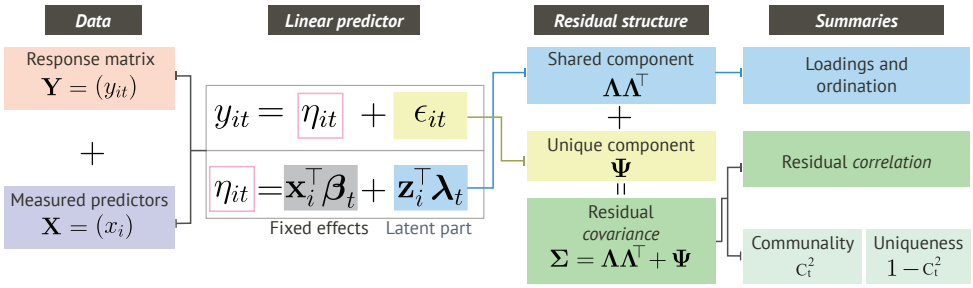

Fig 2. GLLVM architecture: from a chosen response matrix to covariance summaries. A GLLVM starts with a response matrix $\mathbf{Y}$, whose entries y_{it} are observed responses for unit i and response t. The rows define the units of comparison, and the columns define the responses whose covariance is being modelled. Each observation is linked to its expected value through an observation model and link function, $g_t(\mu_{it}) = \eta_{it}$, where $\mu_{it} = E(y_{it})$. Measured predictors $\mathbf{X}$ enter the fixed-effect part of the linear predictor, $\mathbf{x}_i^\top \boldsymbol{\beta}_t$, and estimate response-specific associations with observed covariates. The latent component, $\mathbf{z}_i^\top \boldsymbol{\lambda}_t$, describes shared residual structure among responses through latent scores $\mathbf{z}_i$ and response loadings $\boldsymbol{\lambda}_t$. In the Gaussian identity-link case, the observed response can be written as $y_{it} = \eta_{it} + \epsilon_{it}$, where ϵ_{it} is response-specific residual deviation. This gives the covariance decomposition $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$. The matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ represents shared covariance among responses induced by the latent variables, while the diagonal matrix $\boldsymbol{\Psi}$ collects the variances of the residuals ϵ_{it} and represents unique variation that is not shared with other responses. Together, these components define the model-implied covariance structure, which can be summarised using loadings, ordination, residual covariance or correlation, communality, and uniqueness. Notably, if we have multiple observations per unit, we can construct separate covariance matrices, for example $\boldsymbol{\Sigma}_B$ for between-unit covariance and $\boldsymbol{\Sigma}_W$ for within-unit covariance.

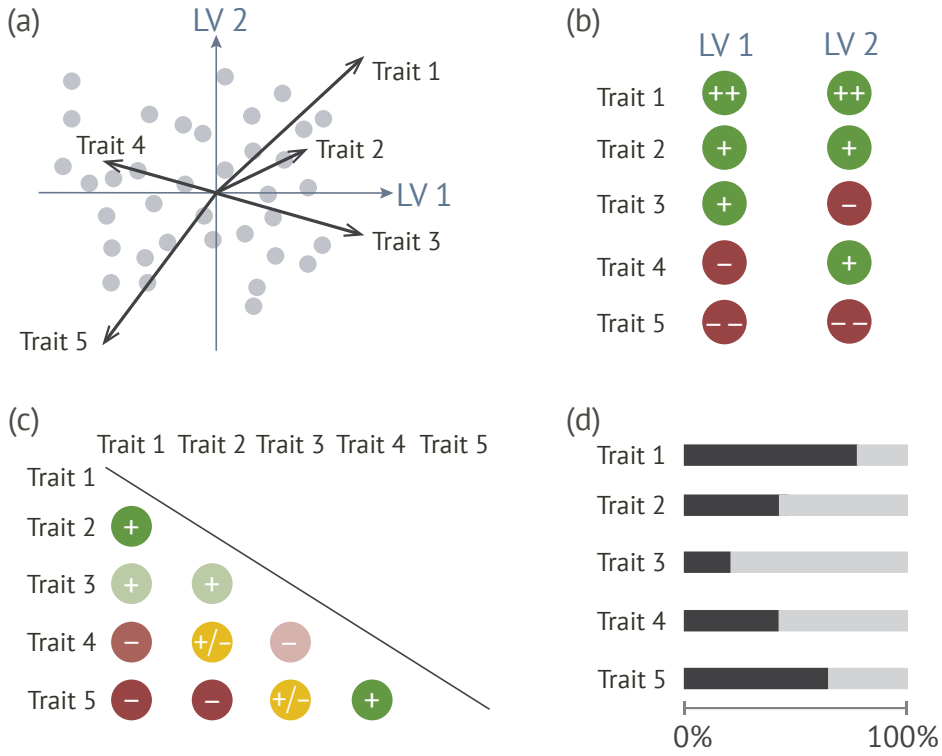

Fig 3. Typical outputs from a fitted GLLVM. (a) The model-based ordination displays units in the latent-variable space. Points represent the fitted positions of sites, individuals, species, tissues, or other observational units. Units that are close together have similar fitted multivariate profiles. Arrows represent trait loadings. A trait pointing strongly along a latent variable is strongly associated with that latent dimension, and traits pointing in similar directions tend to covary positively. (b) The loading matrix gives the same information numerically. Rows correspond to traits and columns correspond to latent variables. Positive and negative signs indicate the direction of association between each trait and each latent variable, while repeated signs indicate stronger loadings in this schematic example. (c) The residual correlation matrix summarises the model-implied residual association among traits after measured predictors have been accounted for. Colours indicate the direction of residual correlations, and colour intensity indicates their strength. Darker colours represent stronger relationships. Because correlation matrices are symmetric, only one triangle is shown. The diagonal represents the correlation of each trait with itself. (d) Communality summarises trait-level integration. For each trait, the filled part of the bar represents the proportion of model-implied variance explained by the shared latent structure, whereas the unfilled part represents trait-specific uniqueness. Together, these outputs help interpret the fitted latent structure, but the latent variables should be treated as summaries of covariance rather than direct evidence of mechanism.

Box 1. Worked example: an Individual × Trait GLLVM for behavioural integration

To illustrate the outputs of a fitted GLLVM, we provide an online worked example using C-BARQ behavioural scores from dogs [63]. The full implementation guide, including code and explanation of the main mathematical objects used in Fig. 2, is available at: `https://santiago-0rtega.github.io/GLLVM_overview/`.
The Canine Behavioral Assessment and Research Questionnaire (C-BARQ) provides quantitative scores for fourteen behavioural categories. We use an Individual × Trait response matrix, where rows are individual dogs and columns are behavioural traits. In the notation of the main text, this means that $\mathbf{Y}$ is a dog-by-behaviour matrix:

$$\mathbf{Y} = \begin{bmatrix} \text{dog 1: aggression} & \text{dog 1: fear} & \cdots & \text{dog 1: energy} \\ \text{dog 2: aggression} & \text{dog 2: fear} & \cdots & \text{dog 2: energy} \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix}.$$

The entry y_{it} is therefore the score of dog i for behavioural trait t.
For simplicity, the worked example focuses on stable among-individual behavioural integration. The model is fitted in long format using `glmmTMB`, with a reduced-rank between-individual covariance structure and trait-specific unique variance. In the notation of Fig. 2, the simplified model can be written as:

$$y_{it} = \mu_t + \mathbf{z}_i^\top \boldsymbol{\lambda}_t + \epsilon_{it}.$$

Here, μ_t is the average score for behavioural trait t, $\mathbf{z}_i$ is dog i's position on the latent behavioural dimensions, $\boldsymbol{\lambda}_t$ is the loading vector for behavioural trait t, and ϵ_{it} is the trait-specific residual deviation. In a model with measured predictors, such as age, sex, breed group, or neuter status, the term $\mathbf{x}_i^\top \boldsymbol{\beta}_t$ would be added to this equation.

Fig 4. Worked example: fitted outputs from an Individual × Trait GLLVM. (a) Model-based ordination of individuals and behavioural trait loadings. (b) Loading matrix for the two fitted latent variables. (c) Between-individual correlation matrix; grey cells indicate the diagonal self-correlations. (d) Trait-level communality and uniqueness.

The four panels summarise complementary parts of the fitted model (Fig. 4). The ordination shows dogs in latent-variable space and behavioural trait loading directions (Fig. 4A). In terms of the model, the points are rows of the latent-score matrix $\mathbf{Z}$, and the arrows are rows of the loading matrix $\mathbf{\Lambda}$. The loading matrix shows the same trait–latent-variable associations numerically (Fig. 4B). For example, a trait with loadings such as

$$\boldsymbol{\lambda}_{\text{Aggression to strangers}} \approx (0.67,\, -0.00)^\top$$

is mainly aligned with latent variable 1, whereas a trait with loadings such as

$$\boldsymbol{\lambda}_{\text{Fear of dogs}} \approx (0.50,\, -0.52)^\top$$

is aligned with both latent variables but in opposite directions.
The between-individual correlation matrix summarises the model-implied behavioural associations after separating shared latent structure from trait-specific uniqueness (Fig. 4C). Mathematically, these associations come from:

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}.$$

The product $\mathbf{\Lambda}\mathbf{\Lambda}^\top$ gives the shared behavioural covariance induced by the latent variables. For two behaviours t and r, their shared covariance is the inner product of their loading vectors:

$$\left(\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)_{tr} = \boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_r.$$

Thus, two behaviours with arrows pointing in similar directions in the ordination tend to have a positive model-implied association; behaviours pointing in opposite directions tend to have a negative association.
The diagonal matrix $\mathbf{\Psi}$ contains trait-specific unique variance: variation in each behaviour that is not explained by the shared latent behavioural dimensions. Communality and uniqueness summarise this decomposition for each trait (Fig. 4D). For behavioural trait t,

$$c_t^2 = \frac{\left(\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)_{tt}}{\left(\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)_{tt} + \mathbf{\Psi}_{tt}}.$$

A high value of c_t^2 means that the behaviour is strongly embedded in the shared among-individual behavioural structure. A low value means that the behaviour is mostly trait-specific in this dataset. This example therefore complements the conceptual overview in Fig. 3 by showing how an Individual × Trait response matrix targets behavioural integration among individuals.

References

[1] McGill BJ, Enquist BJ, Weiher E, Westoby M. Rebuilding community ecology from functional traits. Trends in Ecology & Evolution. 2006;21(4):178–185. doi:10.1016/j.tree.2006.02.002.

[2] Violle C, Navas ML, Vile D, Kazakou E, Fortunel C, Hummel I, et al. Let the concept of trait be functional! Oikos. 2007;116(5):882–892. doi:10.1111/j.0030-1299.2007.15559.x.

[3] Warton DI, Blanchet FG, O'Hara RB, Ovaskainen O, Taskinen S, Walker SC, et al. So many variables: joint modeling in community ecology. Trends in Ecology and Evolution. 2015;30(12):766–779. doi:10.1016/j.tree.2015.09.007.

[4] Dingemanse NJ, Kazem AJN, Reale D, Wright J. Behavioural reaction norms: animal personality meets individual plasticity. Trends in Ecology & Evolution. 2010;25(2):81–89. doi:10.1016/j.tree.2009.07.013.

[5] Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution. 2009;24(3):127–135. doi:10.1016/j.tree.2008.10.008.

[6] Hadfield JD. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of Statistical Software. 2010;33(2):1–22. doi:10.18637/jss.v033.i02.

[7] ter Braak CJF. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology. 1986;67(5):1167–1179. doi:10.2307/1938672.

[8] Roberts DW, Wang Y, Hui FKC. Comparison of distance-based and model-based ordinations. Ecology. 2020;101(10):e03094. doi:10.1002/ecy.2908.

[9] Hui FKC, Taskinen S, Pledger S, Foster SD, Warton DI. Model-based approaches to unconstrained ordination. Methods in Ecology and Evolution. 2015;6(3):399–411. doi:10.1111/2041-210X.12236.

[10] Hui FKC, Warton DI, Ormerod JT, Haapaniemi V, Taskinen S. Variational approximations for generalized linear latent variable models. Journal of Computational and Graphical Statistics. 2016;25(1):105–125. doi:10.1080/10618600.2014.914442.

[11] Niku J, Warton DI, Hui FKC, Taskinen S. Generalized linear latent variable models for multivariate count and biomass data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2017;66(1):142–159. doi:10.1111/rssc.12164.

[12] Warton DI, Hui FKC, Taskinen S. Generalized linear latent variable models for multivariate count and biomass data in ecology. Journal of Agricultural, Biological and Environmental Statistics. 2017;22(4):498–522. doi:10.1007/s13253-017-0304-7.

[13] Niku J, Hui FKC, Taskinen S, Warton DI. gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. Methods in Ecology and Evolution. 2019;10(12):2173–2182. doi:10.1111/2041-210X.13303.

[14] Ovaskainen O, Tikhonov G, Norberg A, Guillaume F, Abrego N, Dunson DB. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. Methods in Ecology and Evolution. 2016;8(5):549–559. doi:10.1111/2041-210X.12502.

[15] Ovaskainen O, Abrego N. Joint Species Distribution Modelling: With Applications in R. 1st ed. Cambridge University Press; 2020.

[16] Dray S, Legendre P. Testing the species traits-environment relationships: the fourth-corner problem revisited. Ecology. 2008;89(12):3400–3412. doi:10.1890/08-0349.1.

[17] Dray S, Choler P, Doledec S, Peres-Neto PR, Thuiller W, Pavoine S, et al. Combining the fourth-corner and the RLQ methods for assessing trait responses to environmental variation. Ecology. 2014;95(7):1752–1760. doi:10.1890/13-0196.1.

[18] Legendre P, Gallagher ED. Ecologically meaningful transformations for ordination of species data. Oecologia. 1997;110(2):277–280. doi:10.1007/s004420050159.

[19] Brown AM, Warton DI, Andrew NR, Binns M, Cassis G, Gibb H. The fourth-corner solution: using species traits to better understand how species traits interact with their environment. Methods in Ecology and Evolution. 2014;5(4):344–352. doi:10.1111/2041-210X.12163.

[20] Wang Y, Naumann U, Wright ST, Warton DI. mvabund: an R package for model-based analysis of multivariate abundance data. Methods in Ecology and Evolution. 2012;3(3):471–474. doi:10.1111/j.2041-210X.2012.00190.x.

[21] Levin SA. The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. Ecology. 1992;73(6):1943–1967. doi:10.2307/1941447.

[22] Cornelissen JHC, Lavorel S, Garnier E, Diaz S, Buchmann N, Gurvich DE, et al. A handbook of protocols for standardised and easy measurement of plant functional traits worldwide. Australian Journal of Botany. 2003;51(4):335–380. doi:10.1071/BT02124.

[23] Funk JL, Larson JE, Ames GM, Feng X, Garofalo-Arevalo S, Geange S, et al. Revisiting the holy grail: using plant functional traits to understand ecological processes. Biological Reviews. 2017;92(2):1156–1173. doi:10.1111/brv.12275.

[24] Schrama M, Bardgett RD, de Bello F, Tsiartsoni M, Wardle DA, Schlaeppi K. Animal functional traits: towards a trait-based ecology for whole ecosystems. Functional Ecology. 2023;37(6):1395–1411. doi:10.1111/1365-2435.14246.

[25] Leps J, Smilauer P, Hadincóvá V. Differences in trait-environment relationships in young and old plant stands. Journal of Ecology. 2023;111(3):550–561. doi:10.1111/1365-2745.14067.

[26] Bartholomew DJ, Knott M, Moustaki I. Latent variable models and factor analysis. 3rd ed. Chichester: John Wiley & Sons; 2011.

[27] Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2009;.

[28] Dingemanse NJ, Dochtermann NA. Quantifying individual differences in behaviour: the evolution and ecology of animal personality. Journal of Animal Ecology. 2013;82(6):1185–1196. doi:10.1111/1365-2656.12103.

[29] Stoffel MA, Nakagawa S, Schielzeth H. rptR: repeatability estimation and variance decomposition by generalized linear mixed-effects models. Methods in Ecology and Evolution. 2017;8(11):1639–1644. doi:10.1111/2041-210X.12797.

[30] Sih A, Bell AM, Johnson JC, Ziemba RE. Behavioral syndromes: an integrative overview. Quarterly Review of Biology. 2004;79(3):241–277. doi:10.1086/422893.

[31] Reale D, Reader SM, Sol D, McDougall PT, Dingemanse NJ. Integrating animal temperament within ecology and evolution. Biological Reviews. 2007;82(2):291–318. doi:10.1111/j.1469-185X.2007.00010.x.

[32] McGillycuddy M, Popovic G, Bolker BM, Warton DI. Parsimoniously fitting large multivariate random effects in glmmTMB. Journal of Statistical Software. 2025;112:1–19.

[33] Bradshaw AD. Evolutionary significance of phenotypic plasticity in plants. Advances in Genetics. 1965;13:115–155. doi:10.1016/S0065-2660(08)60048-6.

[34] Westoby M. A leaf-height-seed (LHS) plant ecology strategy scheme. Plant and Soil. 1998;199(2):213–227. doi:10.1023/A:1004327224729.

[35] Wright IJ, Reich PB, Westoby M, Ackerly DD, Baruch Z, Bongers F, et al. The worldwide leaf economics spectrum. Nature. 2004;428(6985):821–827. doi:10.1038/nature02403.

[36] Kraft NJB, Godoy O, Levine JM. Plant functional traits and the multidimensional nature of species coexistence. Proceedings of the National Academy of Sciences. 2015;112(3):797–802. doi:10.1073/pnas.1413650112.

[37] Dormann CF, McPherson JM, Araujo MB, Bivand R, Bolliger J, Carl G, et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography. 2007;30(5):609–628. doi:10.1111/j.2007.0906-7590.05171.x.

[38] Felsenstein J. Phylogenies and the comparative method. The American Naturalist. 1985;125(1):1–15. doi:10.1086/284325.

[39] Garland T, Harvey PH, Ives AR. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Systematic Biology. 1992;41(1):18–32. doi:10.1093/sysbio/41.1.18.

[40] Pagel MD. Inferring the historical patterns of biological evolution. Nature. 1999;401(6756):877–884. doi:10.1038/44766.

[41] Blomberg SP, Garland T, Ives AR. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution. 2003;57(4):717–745. doi:10.1111/j.0014-3820.2003.tb00285.x.

[42] Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20(2):289–290. doi:10.1093/bioinformatics/btg412.

[43] Garamszegi LZ, editor. Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice. Berlin: Springer-Verlag; 2014.

[44] Ives AR, Midford PE, Garland T. Within-species variation and measurement error in phylogenetic comparative methods. Systematic Biology. 2007;56(2):252–270. doi:10.1080/10635150701313830.

[45] Hardy OJ, Senterre B. Characterizing the phylogenetic structure of communities by an additive partitioning of phylogenetic diversity. Journal of Ecology. 2007;95(3):493–506. doi:10.1111/j.1365-2745.2007.01222.x.

[46] Pearse WD, Kembel SW, Helmus MR. How to define, use, and interpret Pagel's lambda in ecology and evolution. Global Ecology and Biogeography. 2023;32:e14067. doi:10.1111/geb.14067.

[47] Korhonen L, Hui FKC, Niku J, Taskinen S. A review of generalized linear latent variable models and related computational approaches. WIREs Computational Statistics. 2024;16(3):e1700. doi:10.1002/wics.1700.

[48] van der Veen B, Hui FKC, Hovstad KA, O'Hara RB. Concurrent ordination: simultaneous unconstrained and constrained latent variable modelling. Methods in Ecology and Evolution. 2023;14(3):683–704. doi:10.1111/2041-210X.14035.

[49] Taskinen S, Hui FKC, Niku J. Fast and universal estimation of latent variable models using extended variational approximations. Statistics and Computing. 2023;32(4):72. doi:10.1007/s11222-022-10189-w.

[50] Sørensen Ø, Fjell AM, Walhovd KB. Longitudinal modeling of age-dependent latent traits with generalized additive latent and mixed models. Psychometrika. 2023;88(2):456–486. doi:10.1007/s11336-023-09910-z.

[51] Sørensen Ø. Multilevel semiparametric latent variable modeling in R with galamm. Multivariate Behavioral Research. 2024;doi:10.1080/00273171.2024.2385336.

[52] Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R Journal. 2017;9(2):378–400.

[53] Hui FKC. boral–Bayesian ordination and regression analysis of multivariate abundance data in R. Methods in Ecology and Evolution. 2016;7(7):744–750. doi:10.1111/2041-210x.12514.

[54] Tikhonov G, Abrego N, Dunson DB, Ovaskainen O. Joint species distribution modelling with the R-package Hmsc. Methods in Ecology and Evolution. 2020;11(5):652–663. doi:10.1111/2041-210X.13345.

[55] Clark JS, Nemergut DR, Seyednasrollah B, Turner PF, Zhang S. Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. Ecological Monographs. 2017;87(1):34–56. doi:10.1002/ecm.1241.

[56] Pichler M, Hartig F. A new joint species distribution model for faster and more accurate inference of species associations from big community data. Methods in Ecology and Evolution. 2021;12(11):2159–2173. doi:10.1111/2041-210X.13687.

[57] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. Journal of Statistical Software. 2017;76(1):1–32. doi:10.18637/jss.v076.i01.

[58] Stan Development Team. Stan Modeling Language Users Guide and Reference Manual; 2024.

[59] Popovic GC, Warton DI, Hui FKC. Fast model-based ordination with copulas. Methods in Ecology and Evolution. 2022;13(4):989–1003. doi:10.1111/2041-210X.13733.

[60] Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–592. doi:10.1093/biomet/63.3.581.

[61] Lefcheck JS. piecewiseSEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. Methods in Ecology and Evolution. 2016;7(5):573–579. doi:10.1111/2041-210X.12512.

[62] Dormann CF, Schymanski SJ, Cabral J, Chuine I, Graham CH, Hartig F, et al. Correlation and process in species distribution models: bridging a dichotomy. Journal of Biogeography. 2012;39(12):2119–2131. doi:10.1111/j.1365-2699.2011.02659.x.

[63] Swafford B, Labadie J, Knaddison G, Chang J, McVeigh R, Kacenjar L. Behavior Data Commons; 2023. Available from: `https://doi.org/10.62291/UBOQ4994`.