Joint Mean and Correlation Regression Modelling for Multivariate Data

Zhi Yang Tho, Francis Hui and Tao Zou

The Australian National University
Research School of Finance, Actuarial Studies and Statistics

zhiyang.tho@anu.edu.au

Abstract

In the analysis of multivariate or multi-response data, researchers are often not only interested in studying how the mean (say) of each response evolves as a function of covariates, but also and simultaneously how the correlations between responses are related to one or more similarity/distance measures. To address such research questions, we propose a novel joint mean and correlation regression model, which is applicable to a wide variety of correlated discrete and (semi-)continuous responses. Simulations demonstrate the strong finite sample performance of the proposed estimator in terms of point estimation and inference. We apply the proposed joint mean and correlation regression model to a dataset of overdispersed counts of 38 Carabidae ground beetle species sampled throughout Scotland, with results showing in particular that beetle total length and breeding season have statistically important effects in driving the correlations between beetle species.

Motivation

Given the data structure in Figure 1, we would like to simultaneously answer

- What is the relationship between environmental factors and different species?
- Which traits are associated with species correlation?

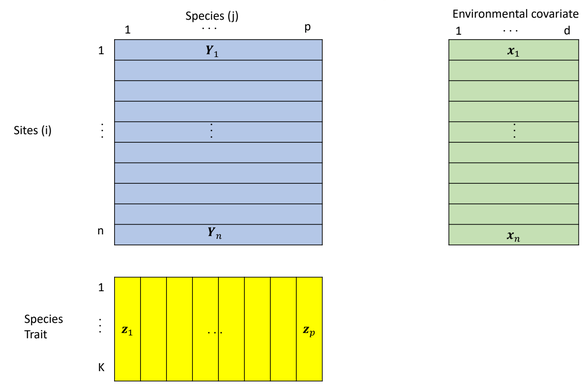

Figure 1: Data structure in community ecology that consists of response matrix, environmental covariate matrix and species trait matrix.

Joint Mean and Correlation Regression Model

Let $\boldsymbol{Y}_i$ be the p-dimensional response (discrete or continuous) vector and $\boldsymbol{x}_i$ be the d-dimensional covariate vector of the i-th cluster and $\mu_{ij}(\boldsymbol{\beta}_j) = \mathrm{E}(Y_{ij})$:

$$g\{\mu_{ij}(\boldsymbol{\beta}_j)\} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j,\ \mathrm{var}(Y_{ij}) = h\{\mu_{ij}(\boldsymbol{\beta}_j); \phi_j\}$$
$$\mathrm{corr}(Y_{ij_1}, Y_{ij_2}) = r_{j_1 j_2}(\boldsymbol{\rho}),$$

where g and h are known link and variance functions. The correlation matrix follows the correlation regression model

$$\boldsymbol{R}(\boldsymbol{\rho}) = \big(r_{j_1 j_2}(\boldsymbol{\rho})\big) = \boldsymbol{I}_p + \sum_{k=1}^{K} \rho_k \boldsymbol{W}_k,$$

where $\boldsymbol{W}_k = (w_{ij}^{(k)})$ are similarity matrices induced from the auxiliary information vectors $\boldsymbol{z}_j$'s, e.g., $w_{j_1 j_2}^{(k)} = \exp(-|z_{j_1 k} - z_{j_2 k}|^2)$. The $\boldsymbol{\rho}$ vector possess a simple and intuitive interpretation as quantifying the impact of the similarities on the correlation between responses. For instance, a higher positive value of ρ_k implies that conditional on other trait values, two species with more similar values in their k-th trait variables are expected to have a stronger positive correlation after accounting for differences in their mean response due to the environmental covariates.

By stacking the response vectors into $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_n^\top)^\top$, we can re-express the model as

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu}(\boldsymbol{\beta}),\ \mathrm{cov}(\boldsymbol{Y}) = \boldsymbol{A}^{1/2}(\boldsymbol{\beta})\{\boldsymbol{I}_n \otimes \boldsymbol{R}(\boldsymbol{\rho})\}\boldsymbol{A}^{1/2}(\boldsymbol{\beta}).$$

where $\boldsymbol{\mu}(\boldsymbol{\beta})$ is the stacked vector of means, $\boldsymbol{A}(\boldsymbol{\beta}) = \mathrm{diag}\{\boldsymbol{A}_1(\boldsymbol{\beta}), \cdots, \boldsymbol{A}_n(\boldsymbol{\beta})\}$ and $\boldsymbol{A}_i(\boldsymbol{\beta}) = \mathrm{diag}\big\{h\{\mu_{i1}(\boldsymbol{\beta}_1); \phi_1\}, \cdots, h\{\mu_{ip}(\boldsymbol{\beta}_p); \phi_p\}\big\}$.

Estimation Approach

In order to obtain a valid correlation matrix estimate $\boldsymbol{R}(\boldsymbol{\rho})$, we consider a slight reparameterisation of the correlation regression model as

$$\boldsymbol{\Sigma}(\boldsymbol{\alpha}) = \alpha_0 \boldsymbol{I}_p + \sum_{k=1}^{K} \alpha_k \boldsymbol{W}_k = \boldsymbol{R}(\boldsymbol{\rho})$$

with $\alpha_0 = 1$ and $\alpha_k = \rho_k$ for $k = 1, \cdots, K$. The resulting covariance equation can then be rewritten as $\mathrm{cov}(\boldsymbol{Y}) = \boldsymbol{A}^{1/2}(\boldsymbol{\beta})\{\boldsymbol{I}_n \otimes \boldsymbol{\Sigma}(\boldsymbol{\alpha})\}\boldsymbol{A}^{1/2}(\boldsymbol{\beta})$, and subsequently the parameters we solve for are now $\boldsymbol{\vartheta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$.

For estimating $\boldsymbol{\beta}$, we solve

$$\boldsymbol{\psi}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{D}^\top(\boldsymbol{\beta})\boldsymbol{A}^{-1/2}(\boldsymbol{\beta})\tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\alpha})\boldsymbol{A}^{-1/2}(\boldsymbol{\beta})\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\} = \boldsymbol{0}_{pd},$$

using Fisher scoring method.

As for estimating $\boldsymbol{\alpha}$, we solve

$$\boldsymbol{\psi}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \Big(\boldsymbol{\epsilon}^\top(\boldsymbol{\beta})\tilde{\boldsymbol{W}}_k\boldsymbol{\epsilon}(\boldsymbol{\beta})\Big) - \big(\mathrm{tr}(\tilde{\boldsymbol{W}}_{k_1}\tilde{\boldsymbol{W}}_{k_2})\big)\boldsymbol{\alpha} = \boldsymbol{0}_{K+1}.$$

using a two-step ADMM approach [6], where

- $\boldsymbol{D}(\boldsymbol{\beta}) = \partial\boldsymbol{\mu}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^\top$
- $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}) = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}(\boldsymbol{\alpha})$
- $\tilde{\boldsymbol{W}}_k = \boldsymbol{I}_n \otimes \boldsymbol{W}_k$
- $\boldsymbol{\epsilon}(\boldsymbol{\beta}) = \big(\boldsymbol{\epsilon}_1^\top(\boldsymbol{\beta}), \cdots, \boldsymbol{\epsilon}_p^\top(\boldsymbol{\beta})\big)^\top$
- $\boldsymbol{\epsilon}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta})\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\}$.

By iteratively solving the two sets of estimating equations, we obtain estimates $\hat{\vartheta} = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\alpha}}^\top)^\top$, which can then be used to obtain the joint estimator of the mean regression coefficients and correlation regression parameters $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\rho}}^\top)^\top$ by setting $\hat{\rho}_k = \hat{\alpha}_k/\hat{\alpha}_0$ for $k = 1, \cdots, K$.

Large Sample Theory

Assume d and K are fixed, $n \to \infty$ and p can either be fixed or grow with increasing n. We introduce a matrix

$$\boldsymbol{\Xi}^{(\mathcal{S})} = \begin{pmatrix} \boldsymbol{E}^{(\mathcal{S})} \otimes \boldsymbol{I}_d & \boldsymbol{0}_{qd\times(K+1)} \\ \boldsymbol{0}_{(K+1)\times pd} & \boldsymbol{I}_{K+1} \end{pmatrix} \triangleq \mathrm{diag}\{\boldsymbol{E}^{(\mathcal{S})} \otimes \boldsymbol{I}_d, \boldsymbol{I}_{K+1}\},$$

which extracts a finite dimensional sub-vector of $\hat{\boldsymbol{\vartheta}}$, e.g., $\boldsymbol{\Xi}^{(\mathcal{S})}\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top, \hat{\boldsymbol{\alpha}}^\top)^\top$ when $S = \{1, 2\}$. Let $\boldsymbol{\vartheta}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ be the true parameter values, then we have the following asymptotic results:

Theorem 1. Under some regularity conditions, we have

$$\boldsymbol{\Omega}^{-1/2}(\boldsymbol{\Xi}^{(\mathcal{S})}) \begin{pmatrix} \sqrt{n}\boldsymbol{I}_{qd} & \boldsymbol{0} \\ \boldsymbol{0} & \sqrt{np}\boldsymbol{I}_{K+1} \end{pmatrix} \boldsymbol{\Xi}^{(\mathcal{S})}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^{(0)}) \xrightarrow{d} N(0, \boldsymbol{I}_{qd+K+1}),$$

as $n \to \infty$, where p is either fixed or tends to infinity but satisfies $p = o(n^{\min\{1/2, \eta/4\}})$, and $\eta > 0$ is related to a moment condition of the responses.

Theorem 2. Under some regularity conditions, we have

$$\overline{\boldsymbol{\Omega}}^{-1/2}(\overline{\boldsymbol{\Xi}}^{(\mathcal{S})}) \begin{pmatrix} \sqrt{n}\boldsymbol{I}_{qd} & \boldsymbol{0} \\ \boldsymbol{0} & \sqrt{np}\boldsymbol{I}_{K} \end{pmatrix} \overline{\boldsymbol{\Xi}}^{(\mathcal{S})}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(0)}) \xrightarrow{d} N(0, \boldsymbol{I}_{qd+K}),$$

as $n \to \infty$, where p is either fixed or tends to infinity but satisfies $p = o(n^{\min\{1/2, \eta/4\}})$, and $\eta > 0$ is related to a moment condition of the responses.

Simulation Studies

- Marginal distributions of response: Bernoulli, Poisson and Gaussian;
- $n \in \{50, 100, 200, 400\}$, $p \in \{10, 25, 50\}$, $d = 4$ and $K = 5$;
- $\beta_{jl}^{(0)} \overset{i.i.d}{\sim} N(0, 0.5)$, $\rho_k^{(0)} \overset{i.i.d}{\sim} U(-0.05, 0.05)$, $\phi_j = 1$;
- $\boldsymbol{x}_i$ and $\boldsymbol{z}_j$ vectors are obtained from standardised environmental covariates and species traits, respectively, in the real data application;
- A total of 1000 replications for each setting.

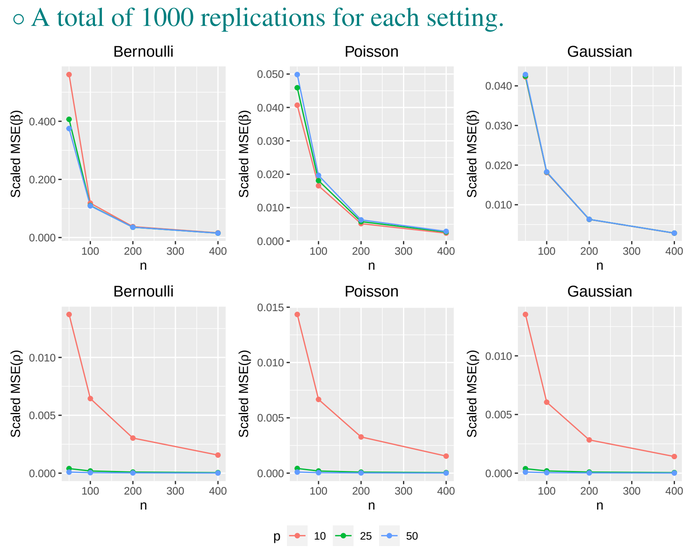

Figure 2: Scaled MSE for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\rho}}$ by marginal distributions of responses.

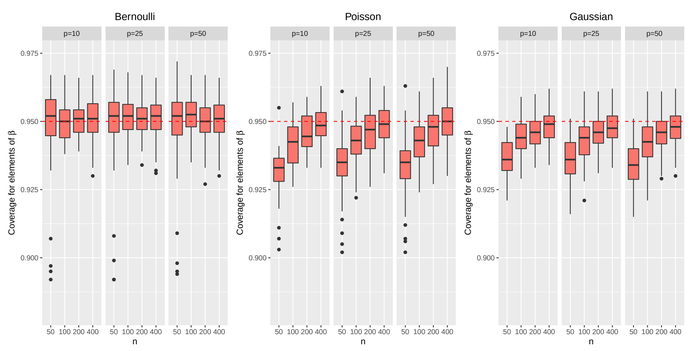

Figure 3: Boxplots summarising coverage of 95% approximate confidence intervals for each of the mean regression coefficients in the simulation studies where the total number of coefficients depends on the dimension of responses p.

Real Data Application

- Data comes from [4], which consists of the counts of $p = 38$ carabid ground beetle species: $\boldsymbol{Y}_i$, for $i = 1, \cdots, n$, on $n = 87$ sites spread across nine main areas in Scotland.
- Study the effects of environmental processes on different ground beetle species, and quantify how different traits influence residual correlation between species after accounting for species-environmental responses.

Estimation of β_j				
Species j	Intercept	Soil pH	Elevation	Land management
A.Mueller	**2.444 (2.015, 2.872)**	**1.382 (0.631, 2.132)**	**-0.730 (-1.396, -0.064)**	-0.022 (-0.725, 0.682)
A.Apricaria	-0.697 (-1.528, 0.135)	-0.151 (-1.186, 0.883)	**-1.649 (-2.904, -0.393)**	**1.615 (0.692, 2.538)**
A.Bifrons	**-1.349 (-2.462, -0.236)**	**1.415 (0.134, 2.696)**	**-2.145 (-3.720, -0.570)**	0.314 (-0.775, 1.404)
A.Communis	**0.604 (0.145, 1.062)**	-0.080 (-0.876, 0.715)	-0.439 (-1.058, 0.181)	**-0.970 (-1.741, -0.198)**
A.Familiaris	0.221 (-0.382, 0.823)	0.454 (-0.593, 1.500)	-0.371 (-1.287, 0.544)	0.474 (-0.509, 1.458)
A.Lunicollis	-0.016 (-0.533, 0.502)	-0.714 (-1.644, 0.217)	**-0.841 (-1.600, -0.082)**	-0.542 (-1.401, 0.317)
A.Plebeja	**3.074 (2.667, 3.480)**	**1.576 (0.860, 2.291)**	**-0.933 (-1.567, -0.298)**	-0.390 (-1.063, 0.283)
A.Dorsalis	0.106 (-0.513, 0.725)	0.305 (-0.544, 1.153)	-0.144 (-1.064, 0.776)	**2.557 (1.720, 3.394)**
B.Aeneum	**2.010 (1.265, 2.756)**	**1.467 (0.262, 2.673)**	**-2.993 (-4.314, -1.672)**	-0.236 (-1.345, 0.873)
B.Guttula	**1.948 (1.560, 2.336)**	**0.921 (0.264, 1.577)**	**-1.024 (-1.672, -0.376)**	0.448 (-0.159, 1.055)

Estimation of ρ_k				
		Traits k		
Total length	Leg colour	Wing development	Overwintering	Breeding season
0.061 (0.023, 0.098)	-0.002 (-0.022, 0.018)	0.021 (-0.001, 0.044)	0.004 (-0.018, 0.026)	**0.036 (0.008, 0.064)**

Table 1: Point estimates and 95% confidence intervals (in parentheses) for mean regression coefficients of the first 10 species, and correlation regression parameters for all five different trait variables.

- Carabid ground beetle species possess differing relationship with the environmental factors e.g., *A.Muelleri* prefers higher levels of soil pH, while *A.Communis* presents no clear evidence of being influenced by this habitat factor.
- The correlation regression parameter estimate for total length presented the strongest magnitude and significance, suggesting that conditional on other trait values, the abundances of carabid species with similar total lengths are more positively associated after accounting for differences in their mean abundance due to environmental processes.

Figure 4: *A.Muelleri* (left) , *A.Communis* (right) ground beetles.

Conclusion

- Different from general fourth corner analyses in community ecology that aim at identifying where traits mediate species responses to the environment, which only deal with the mean model [1, 3], our model links the trait information of different species to the correlation between their abundances.
- The model can be applied to handle a wide variety of discrete and continuous responses, e.g., overdispersed counts, binomial, etc.
- Other than community ecology, the model can potentially be used in quantitative genetics where traits across individuals vary with demographics, and covariance between traits of different people is linked to the associated genotype matrices [5]. Also, it can be applied to sociology where the number of publications for authors vary with journals, while covariation between publications of different authors is related to the co-authorship and citation networks [2].

References

[1] Brown, A. M., Warton, D. I., Andrew, N. R., Binns, M., Cassis, G., and Gibb, H. (2014). The fourth-corner solution – using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution*, 5(4):344–352.

[2] Ji, P. and Jin, J. (2016). Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812.

[3] Pollock, L. J., Morris, W. K., and Vesk, P. A. (2012). The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*, 35(8):716–725.

[4] Ribera, I., Dolédec, S., Downie, I. S., and Foster, G. N. (2001). Effect of land disturbance and stress on species traits of ground beetle assemblages. *Ecology*, 82(4):1112–1129.

[5] Zhou, H., Hu, L., Zhou, J., and Lange, K. (2019). Mm algorithms for variance components models. *Journal of Computational and Graphical Statistics*, 28(2):350–361.

[6] Zou, T., Lan, W., Wang, H., and Tsai, C.-L. (2017). Covariance regression analysis. *Journal of the American Statistical Association*, 112(517):266–281.