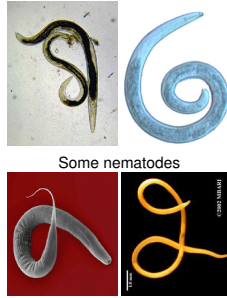


What high dimensional abundance data?

- Counts of the abundances of many taxa, at the same places.
- Commonly collected in ecology: thousands of publications use this type of data!
- Typical study aim – how are multivariate abundances related to environmental factors?
- Typical data properties:
 - Overdispersed count data, many zero counts
 - Many variables, $N < p$
- An example dataset is below.

Nematode abundance at different nutrient levels
Rows: experimental ponds ($N = 12$)
Columns: nematode species ($p = 53$)

control	(1 0 0 1 3 8 0 0 0 0 4 0 0 ... 3)
	(0 1 0 1 8 0 0 0 1 7 3 1 ... 0)
	(0 2 0 0 7 0 0 0 0 1 4 5 0 ... 0)
	(0 1 0 1 6 0 0 1 0 2 3 4 0 ... 0)
low	(0 2 9 0 1 0 1 2 0 1 6 3 0 ... 0)
	(0 0 4 0 2 1 2 0 0 0 1 1 0 0 ... 0)
	(0 1 1 1 1 6 1 2 0 1 5 9 0 ... 0)
	(0 1 0 0 2 6 1 0 0 0 6 9 0 ... 0)
high	(0 1 0 0 4 0 0 0 0 7 0 0 ... 0)
	(0 0 0 1 1 7 1 2 1 0 1 7 0 0 ... 0)
	(0 2 2 4 4 0 0 0 0 4 0 0 ... 0)
	(0 0 0 3 3 2 0 0 0 0 1 0 0 ... 0)



I am interested in developing methods for modelling this type of data. Here I summarise five important new results (labelled a-e) for the analysis of multivariate abundances in ecology:

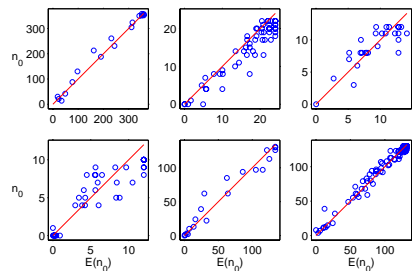
a) Many zeros, but not zero inflated

What types of count distributions tend to best fit this type of data?

- I obtained twenty multivariate abundance datasets from the ecology literature.
- I compared the goodness-of-fit of different parametric models to the count data.
- Distributions considered: Poisson, negative binomial (with $V(\mu) = \phi\mu$ or $V(\mu) = \mu + \psi\mu^2$), zero-inflated Poisson and zero-inflated negative binomial.
- Measured goodness-of-fit using information criteria (AIC, BIC) and graphical diagnostic tools.

model	average AIC
Poisson	298
Negative binomial ($V(\mu) = \mu + \psi\mu^2$)	105
Negative binomial ($V(\mu) = \phi\mu$)	109
Zero-inflated Poisson	221
Zero-inflated negative binomial	121

Observed (n_0) vs predicted ($E(n_0)$) number of zeros for six multivariate abundance datasets

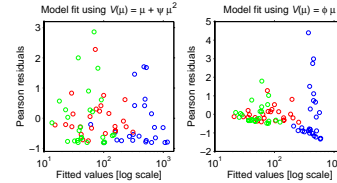


Just because you have lots of zeros doesn't mean your data are zero-inflated.

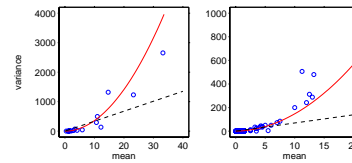
b) $V(\mu) = \phi\mu$ can fit poorly!

- The mean-variance relationship tends to be non-linear (and approximately quadratic).
- The standard "over-dispersed Poisson" or "quasi-Poisson" approach, $V(\mu) = \phi\mu$, tends to over-estimate the variance for rare species and underestimate the variance for abundant species.
- So check your mean-variance relationship!

Residual plots for abundance of three types of benthic macroinvertebrate, Delaware Bay



Mean-variance relationship for abundance of two species:
(left) Zooplankton in experimental ditches, the Netherlands
(right) Terrestrial invertebrates from the Hunter Valley, Australia



c) Try generalised estimating equations

How can we model multivariate overdispersed count data?

- I propose using **generalised estimating equations**.
- GEEs are designed for correlated non-normal data where primary interest is in regression modelling of the mean of the marginal distribution (which is what we are interested in).
- I propose using "negative binomial regression": based on Lawless (1987)

(β) Marginal models, given estimates of $\psi = (\psi_1, \dots, \psi_p)$, are GLM's:

$$\log(\mu_{ij}) = X_i\beta_j \quad V(\mu_{ij}) = \mu_{ij} + \psi_j\mu_{ij}^2$$

(ψ) ψ can be found by moments, given estimates of the μ_{ij} .

Iterate between (β) and (ψ) until convergence, given an initial estimate of β at $\psi = 0$.

This can be generalised to multivariate settings using a "one-step GEE approach":

(β, ψ) Use negative binomial regression to estimate β and ψ by iteration (assuming $\bar{R} = \mathbf{I}$).

(α) Estimate the unstructured correlation matrix $\bar{R}(\alpha)$ by moments, given β and ψ .

Do not iterate between (β, ψ) and (α), because α has too many nuisance parameters.

Acknowledgements

Thanks to Malcolm Hudson (Macquarie University), Wojtek Krzanowski (University of Exeter) and Peter Guttorp (University of Washington), and thanks to their respective institutions for hosting my visits. For funding, thanks to the UNSW Faculty of Science.

References

- Lawless J.F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* **15**, 209-225.
- Warton D.I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics* **16**, 275-289.
- Warton D.I. & Hudson H.M. (2004). A MANOVA statistic is just as powerful distance-based statistics, for multivariate abundances. *Ecology* **85**(3), 858-874.
- Also, I can offer you preprints of d) and e)...

Prepared October 2006.

d) Use regularisation for high dimensional data

How does all this work for high-dimensional data ($N < p$)?

To test hypotheses about (marginal) mean abundance:

- We use a Wald statistic with P -values calculated by **resampling rows** of data.
- This ensures (approximately) exact inference despite possible misspecification of the model.
- But this requires calculating \bar{R}^{-1} , and \bar{R} is singular when $N < p$
→ **regularise \bar{R}**

I propose **ridge regularisation**:

$$\bar{R}_\lambda = \lambda \bar{R} + (1 - \lambda)\mathbf{I} \quad (\propto \bar{R} + \kappa \mathbf{I})$$

This approach has some useful properties:

- It "shrinks" \bar{R} towards \mathbf{I} .
- \bar{R}_λ has the same eigenvectors as \bar{R} , while shrinking the eigenvalues towards 1.
- \bar{R}_λ can be derived as the penalised normal likelihood estimator with penalty term $\kappa \text{tr}(\bar{R}^{-1})$.
- It has good power properties. (Ask David for more details!)

The ridge parameter λ can be estimated by **cross-validation**, using the normal likelihood. Hence

- \bar{R}_λ is consistent for \bar{R} ($\lambda \rightarrow 1$ as $N \rightarrow \infty$).
- Non-singularity is guaranteed: $P(\hat{\lambda} = 1, \text{rank}(\bar{R}) < p) = 0$.
- Simulations suggest that $\lambda \uparrow$ as $N \uparrow$ or $p \downarrow$.

e) Reparameterise to model compositional change

Now consider situations when we are interested in compositional or relative abundance, not absolute abundance.

How can we model compositional abundance?

We can exploit the multiplicative nature of log-linear models to model compositional changes.

Recall that the model for the marginal mean is log-linear:

$$\log(\mu_{ij}) = X_i\beta_j, \quad j \in \{1, 2, \dots, p\}$$

For questions about composition, we are instead interested in

$$\log\left(\frac{\mu_{ij}}{\mu_{ip}}\right) = X_i(\beta_j - \beta_p), \quad j \in \{1, 2, \dots, p-1\}$$

β_j describes effects on absolute abundance.

$\beta_j - \beta_p$ describes effects on compositional abundance.

So we use the same GEE model, but **reparameterise** it as:

$$\log \mu_{ij} = X_i\alpha + X_i\gamma_j, \quad \gamma_p = 0$$

where

α is the "main effect". $\alpha = \beta_p$.

γ_j is the compositional effect of the j th variable, relative to the p th variable. $\gamma_j = \beta_j - \beta_p$.

This allows us to **partition** hypothesis tests about effects on abundance (β_j) into a "main effect" (α) and compositional effects (γ_j), which is great for interpretation!

Further work

I am interested in various other aspects of the analysis of this type of data, such as:

- Can we impose some simpler structure on the correlation matrix?
e.g. sparsity (using covariance selection)?
- What about using a generalised linear mixed modelling approach?
(Problem: How could we model correlation in an adequate yet feasible way?)
- How can we construct biplots to summarise multivariate effects visually?
- Which multiple testing procedures should we apply, and how?