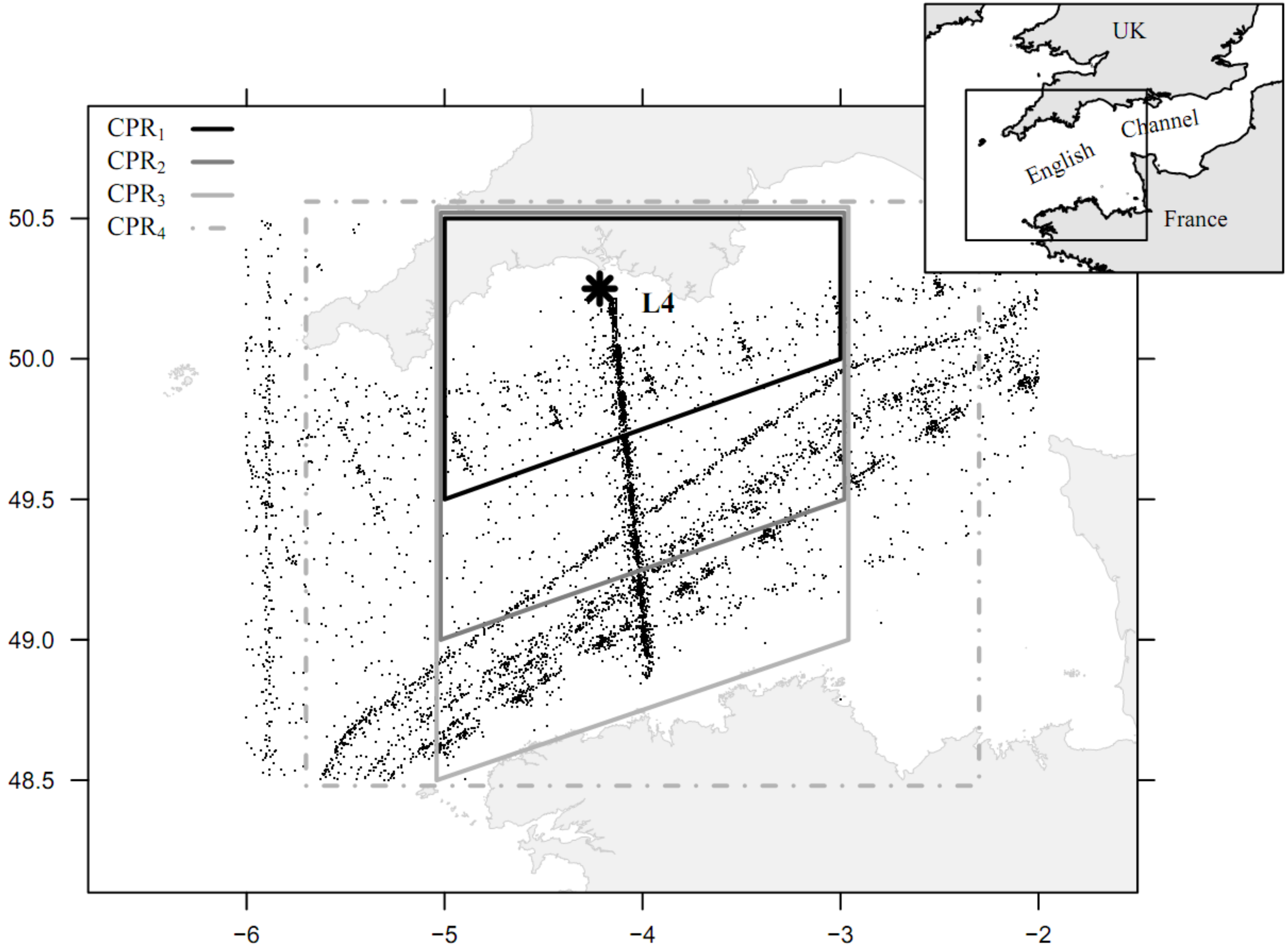
Inferring plankton community dynamics with vector autoregressive state-space models

Northwest Fisheries Science Center
Elizabeth Holmes, Eric Ward, Mark Scheurell, and Dan Pendleton

National Center for Ecological Analysis and Synthesis
Stephanie Hampton and Lindsay Scheef

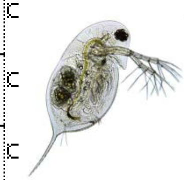
Motivation: understanding plankton dynamics from long-term data sets



14 groups







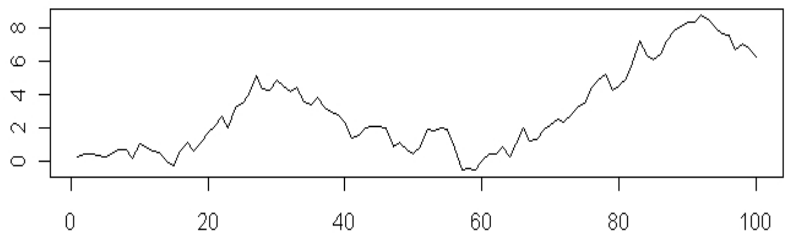

	Group	Proportion of community L4	Proportion of community CPR	Taxa Included	Proportion of group L4 & CPR mean
	Chaetognaths	0.02	0.07	Sagitta spp.	~1.00
	Pteropods	0.01	0.02	Thecosomata	>0.99
	Tunicates	0.03	0.07	Appendicularians Doliolids	0.99 0.01
	Cladocerans	0.05	0.04	Evadne spp. Podon spp.	0.66 0.34
	Amphipods	<0.01	<0.01	Gammarid amphipods Hyperiid amphipods Isopods Mysid shrimp	0.94 0.03 0.02 0.01
	Krill	<0.01	<0.01	Euphausiids	~1.00
Copepods	Large calanoids	0.03	0.08	Calanus spp. Metridia spp. Candacia spp. Eucalanus spp.	0.95 0.03 0.01 0.01
Copepods	Small calanoids	0.38	0.45	Pseudocalanus spp. Acartia spp. Temora spp. Paracalanus spp. Centropages spp. Clausocalanus spp. Ctenocalanus spp.	0.33 0.28 0.15 0.12 0.06 0.02 0.01
Copepods	Cyclopoids	0.12	0.02	Oithona spp.	~1.00
Copepods	Poecilostomatoids	0.19	0.01	Corycaeus spp. Oncaea spp.	0.51 0.49
Copepods	Harpacticoids	0.01	<0.01	Euterpina spp. Clytemnestra spp. Microsetella spp. Alteutha spp.	0.70 0.23 0.05 0.01
Meroplankton	Cirripedia	0.08	0.01	Cirripede larvae	1.00
Meroplankton	Mero. grazers (miscellaneous)	0.06	0.23	Echinoderm larvae Bivalve larvae Cyphonaute larvae Polychaete larvae Gastropod larvae	0.66 0.19 0.05 0.05 0.04
Meroplankton	Decapod larvae	0.01	0.01	Crab & shrimp larvae	1.00

But the talk is about the statistical methods

- What is a multivariate autoregressive state-space model (MARSS or VARSS)?
- A tour of different classes of time series models written as MARSS (more math)
- Estimating parameters using an EM algorithm for MARSS models with linear constraints (more math)
- MARSS R package
- Estimating the species interaction matrix and covariate matrices for PLANKTON (actually more math)
- Stabilitiy metrics (cartoons!)
- Some results from the plankton work

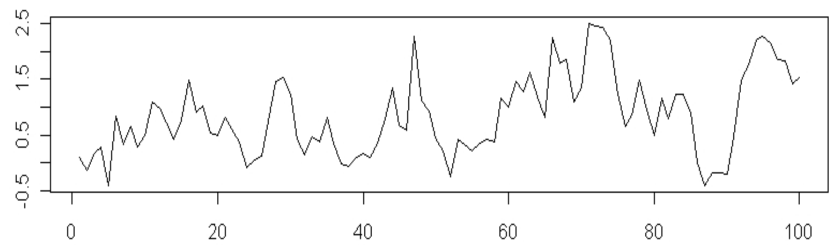
MARSS model

SOME UNDERLYING AUTOREGRESSIVE PROCESS



Random walk

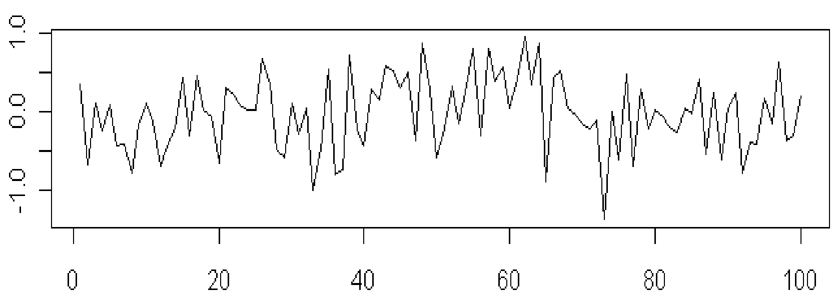
$x(t)=x(t-1)+e(t)$ $+u$



Mean-reverting random walk

$x(t)=bx(t-1)+e(t)$ $+u$

+ OBSERVATION PROCESS



White noise

$x(t)=e(t)$ $+u$

NAMES

Univariate: Autoregressive state-space

Mulitvariate: Vector autoregressive SS

- Multivariate autoregressive SS
- Dynamic linear model
- Structural SS time-series model

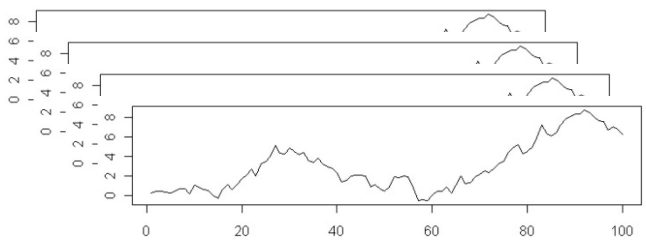
MARSS (or VARSS) model

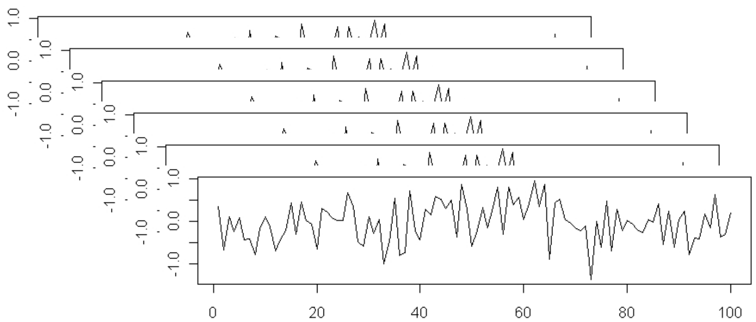
Multivariate autoregressive "random walk"

$$\mathbf{x}_t = \mathbf{B}\mathbf{x}_{t-1} + \mathbf{u} + \mathbf{w}_t, \text{ where } \mathbf{w}_t \sim \mathrm{MVN}(0, \mathbf{Q})$$

$$\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{v}_t, \text{ where } \mathbf{v}_t \sim \mathrm{MVN}(0, \mathbf{R})$$

Multivariate with noise





written out....

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_t = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_{t-1} + \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}_t, \quad \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}_t \sim MVN\left(\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \right)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}_t = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ z_{31} & z_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_t + \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}_t, \quad \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}_t \sim MVN\left(\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}, \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \right)$$

Long history in economics, finance and engineering

Model with lags (lag-p models)

$$\mathbf{x}'_t = \mathbf{B}_1\mathbf{x}'_{t-1} + \mathbf{B}_2\mathbf{x}'_{t-2} + \mathbf{u}' + \mathbf{w}'_t, \text{ where } \mathbf{w}'_t \sim \text{MVN}(0, \mathbf{Q}')$$

x at t-2 affects x at t

In MARSS form, it becomes...

$$\mathbf{x}_t = \mathbf{B}\mathbf{x}_{t-1} + \mathbf{u} + \mathbf{w}_t, \text{ where } \mathbf{w}_t \sim \text{MVN}(0, \mathbf{Q})$$

$$\begin{bmatrix}\mathbf{x}'_t \\ \mathbf{x}'_{t-1}\end{bmatrix} = \begin{bmatrix}\mathbf{B}_1 & \mathbf{B}_2 \\ \mathbf{I}_m & 0\end{bmatrix}\begin{bmatrix}\mathbf{x}'_{t-1} \\ \mathbf{x}'_{t-2}\end{bmatrix}_{t-1} + \begin{bmatrix}\mathbf{u}' \\ 0\end{bmatrix} + \mathbf{w}_t, \ \mathbf{w}_t \sim \text{MVN}\left(0, \begin{bmatrix}\mathbf{Q}' & 0 \\ 0 & 0\end{bmatrix}\right)$$

Multivariate moving average models

$$\mathbf{x}'_t = \mathbf{w}'_t + \mathbf{\Theta}_1 \mathbf{w}'_{t-1} + \mathbf{\Theta}_2 \mathbf{w}'_{t-2}, \text{ where } \mathbf{w}'_t \sim \text{MVN}(0, \mathbf{Q}')$$

In MARSS form, it becomes...

$$\mathbf{x}_t = \mathbf{B}\mathbf{x}_{t-1} + \mathbf{u} + \mathbf{w}_t, \text{ where } \mathbf{w}_t \sim \text{MVN}(0, \mathbf{Q})$$

$$\begin{bmatrix} \mathbf{x}'_{t-2} \\ \mathbf{x}'_{t-1} \\ \mathbf{x}'_t \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{I}_m & 0 \\ 0 & 0 & \mathbf{I}_m \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}'_{t-3} \\ \mathbf{x}'_{t-2} \\ \mathbf{x}'_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \mathbf{w}'_t \end{bmatrix}, \ \mathbf{w}_t \sim \text{MVN}\left(0, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mathbf{Q}' \end{bmatrix}\right)$$

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{\Theta}_2 & \mathbf{\Theta}_1 & 1 \end{bmatrix} \mathbf{x}_t$$

Autoregressive process noise

$$\mathbf{x}'_t = \mathbf{B}\mathbf{x}'_{t-1} + \mathbf{u}' + \boldsymbol{\eta}_t$$

where $\boldsymbol{\eta}_t$ is a AR-1 (or p) process.

We re-write this as a MARSS(1) model by moving the error term into the state process

$$\mathbf{x}_t = \mathbf{B}\mathbf{x}_{t-1} + \mathbf{u} + \mathbf{w}_t, \text{ where } \mathbf{w}_t \sim \text{MVN}(0, \mathbf{Q})$$

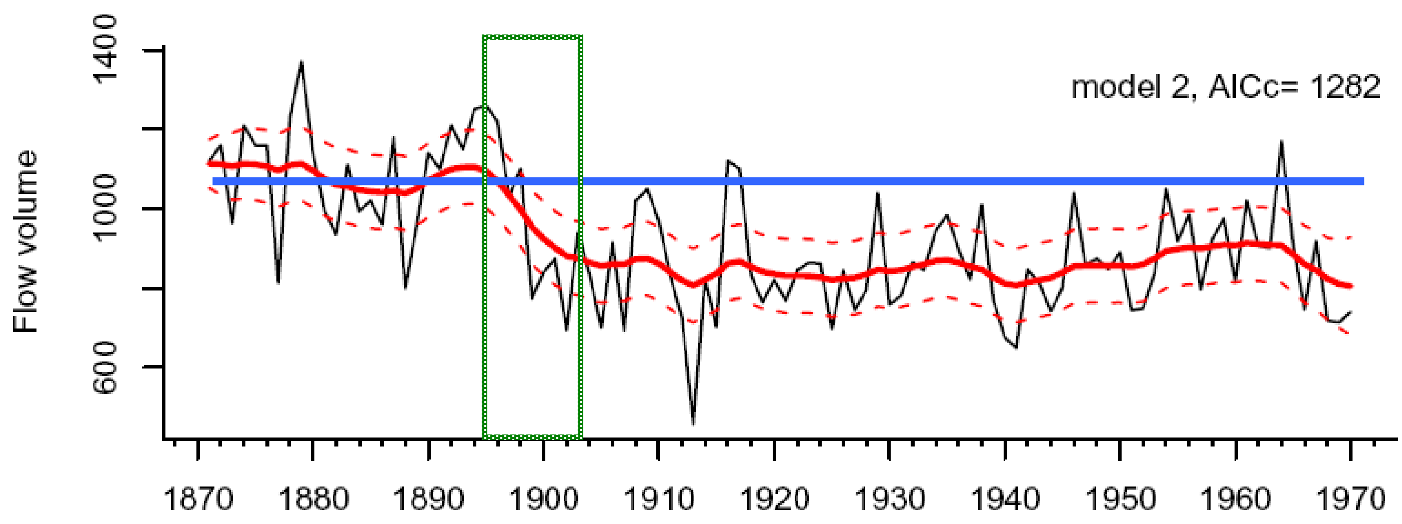
$$\begin{bmatrix} \mathbf{x}' \\ \boldsymbol{\eta} \end{bmatrix}_t = \begin{bmatrix} \mathbf{B}_x & \mathbf{I}_m \\ 0 & \mathbf{B}_\eta \end{bmatrix} \begin{bmatrix} \mathbf{x}' \\ \boldsymbol{\eta} \end{bmatrix}_{t-1} + \begin{bmatrix} \mathbf{u}' \\ 0 \end{bmatrix} + \mathbf{w}_t, \ \mathbf{w}_t \sim \text{MVN}\left(0, \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{Q}_\eta \end{bmatrix}\right)$$

Stochastic level model (used to detect structural breaks)

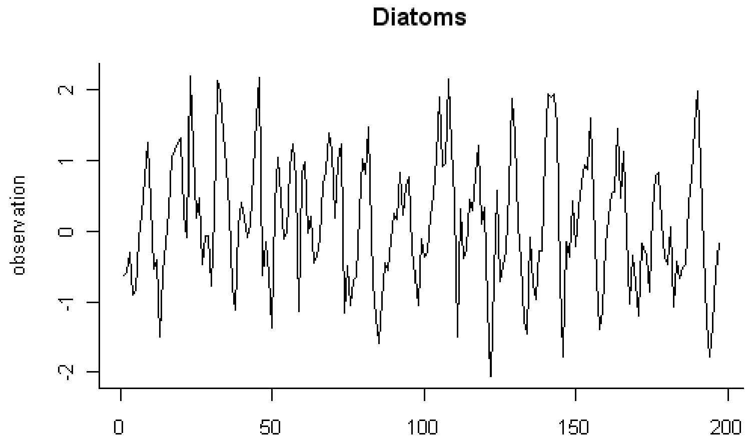
$$x_t = x_{t-1} + w_t$$

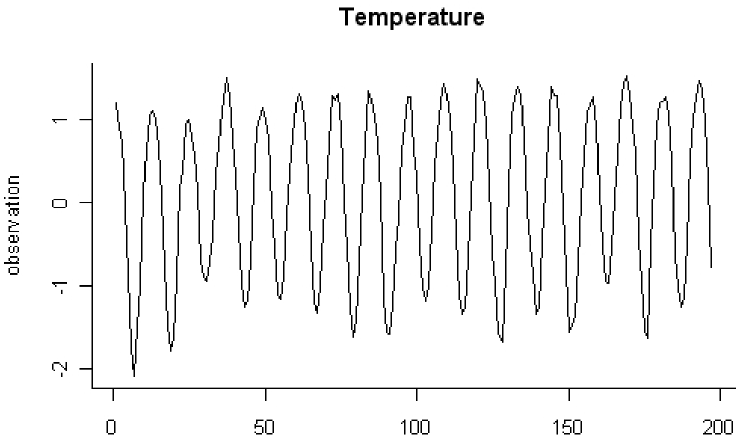
$$y_t = x_t + v_t$$

The mean level is an autoregressive process

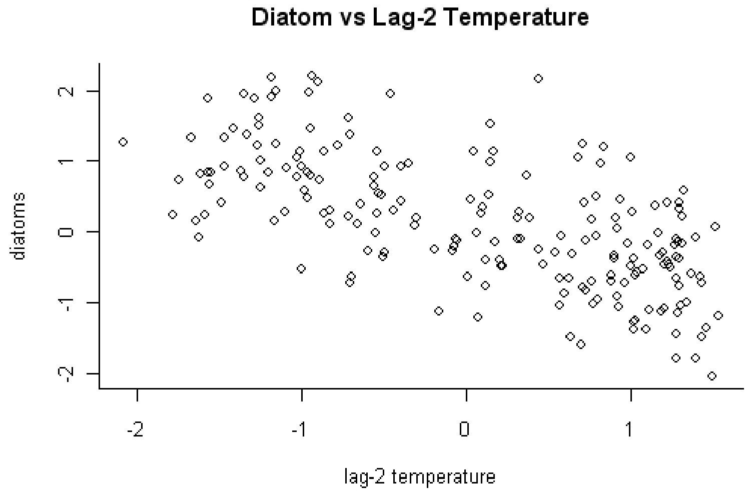


Model with covariates (exogenous variables)





$\mathbf{f}_t$



Effect enters as a process change

$$\mathbf{x}_t = \mathbf{B}\mathbf{x}_{t-1} + \mathbf{u} + \mathbf{C}\mathbf{f}_t + \mathbf{w}_t;$$

$$\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{D}\mathbf{f}_t$$

vs Effect enters as a level change

W. T. Edmondson dataset courtesy of D. Schindler

Model with covariates written as a MARSS by moving covariates into states

$$\mathbf{x}_t = \mathbf{B}\mathbf{x}_{t-1} + \mathbf{u} + \mathbf{C}\mathbf{f}_t + \mathbf{w}_t, \text{ where } \mathbf{w}_t \sim \text{MVN}(0, \mathbf{Q})$$

$$\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{D}\mathbf{f}_t + \mathbf{v}_t, \text{ where } \mathbf{v}_t \sim \text{MVN}(0, \mathbf{R})$$

We can re-write this as a MARSS(1) model

$$\begin{bmatrix}\mathbf{x}^{(v)}\\ \mathbf{x}^{(c)}\end{bmatrix}_t = \begin{bmatrix}\mathbf{B}^{(v)} & \mathbf{C}\\ 0 & \mathbf{B}^{(c)}\end{bmatrix}\begin{bmatrix}\mathbf{x}^{(v)}\\ \mathbf{x}^{(c)}\end{bmatrix}_{t-1} + \begin{bmatrix}\mathbf{u}^{(v)}\\ \mathbf{u}^{(c)}\end{bmatrix} + \mathbf{w}_t, \; \mathbf{w}_t \sim \text{MVN}\left(0, \begin{bmatrix}\mathbf{Q}^{(v)} & 0\\ 0 & \mathbf{Q}^{(c)}\end{bmatrix}\right)$$

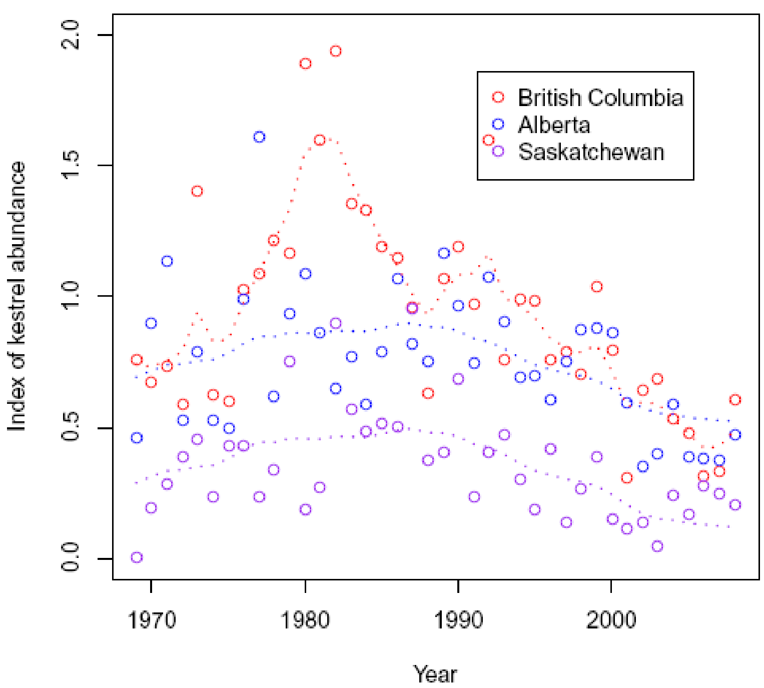
$$\begin{bmatrix}\mathbf{y}^{(v)}\\ \mathbf{y}^{(c)}\end{bmatrix}_t = \begin{bmatrix}\mathbf{Z}^{(v)} & \mathbf{D}\\ 0 & \mathbf{Z}^{(c)}\end{bmatrix}\begin{bmatrix}\mathbf{x}^{(v)}\\ \mathbf{x}^{(c)}\end{bmatrix}_t + \begin{bmatrix}\mathbf{a}^{(v)}\\ \mathbf{a}^{(c)}\end{bmatrix} + \mathbf{v}_t, \; \mathbf{v}_t \sim \text{MVN}\left(0, \begin{bmatrix}\mathbf{R}^{(v)} & 0\\ 0 & \mathbf{R}^{(c)}\end{bmatrix}\right)$$

The covariates can be modeled as a autoregressive process

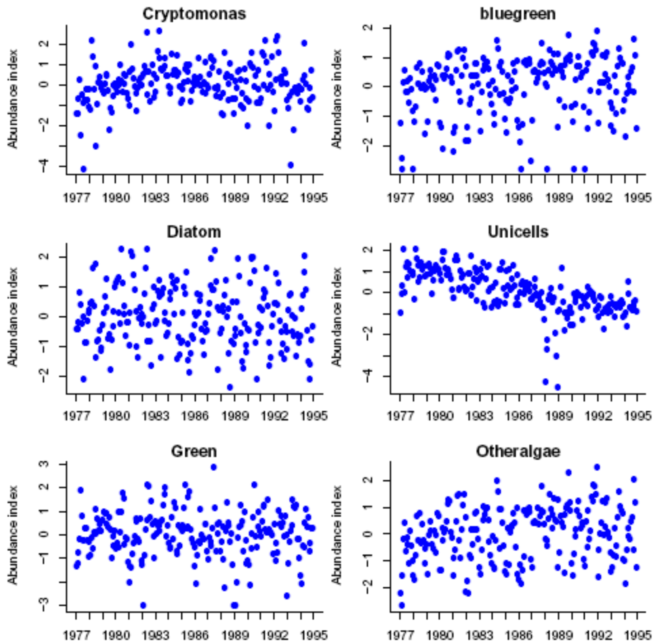
The covariates might have an observation process (to deal with missing values, multiple time series, changing time series)

Other types of models that can be written as MARSS

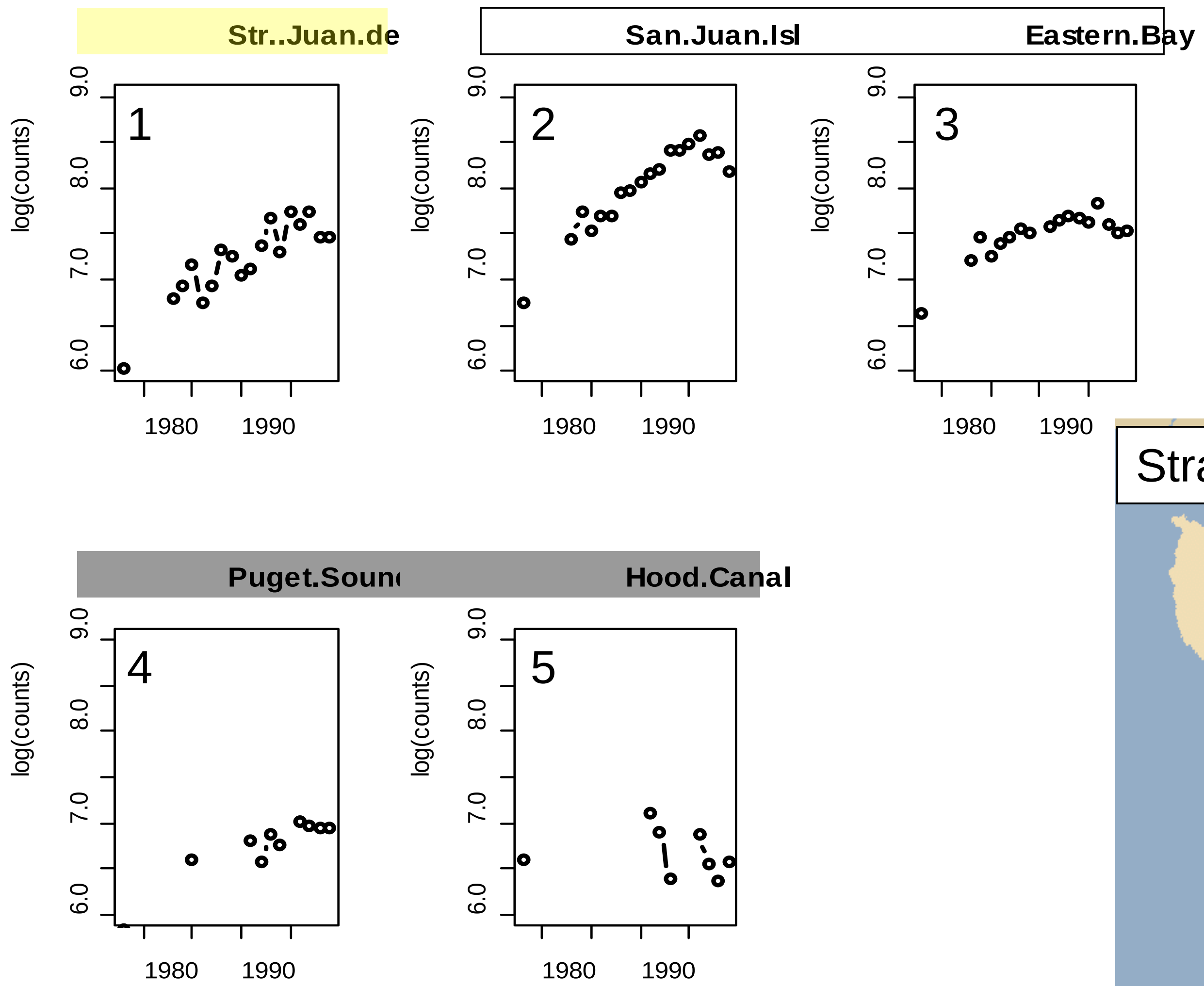
- Link multiple observations of a single process



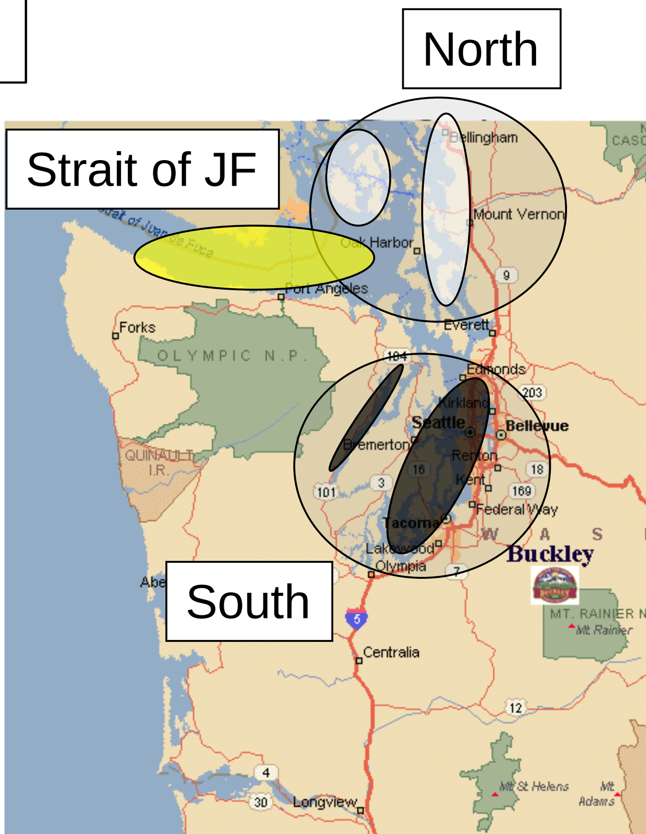
- Try to find a minimum set of stochastic trends to describe a large set of time series



Hierarchical models with shared parameters



5 sampling locations



Jeffries et al. 2003. Trends and status of harbors seals in Washington State: 1978–1999. J. Wildl. Manage. 67(1):208-219

Finding MLE parameters for MARSS models

- What is a multivariate autoregressive state-space model (MARSS or VARSS)?
- A tour of different classes of time series models written as MARSS (more math)
- Estimating parameters using an EM algorithm for MARSS models with linear constraints (more math)
- MARSS R package
- Estimating the species interaction matrix and covariate matrices for PLANKTON (actually more math)
- Stabilitiy metrics (cartoons!)
- Some results from the plankton work

Finding MLE parameters for MARSS models

Joint likelihood of y(data), x(hidden states)

$$\log \mathbf{L}(\boldsymbol{y},\boldsymbol{x};\Theta) = -\sum_{1}^{T}\frac{1}{2}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathrm{a})^\top \mathrm{R}^{-1}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathrm{a}) - \sum_{1}^{T}\frac{1}{2}\log|\mathbf{R}|$$
$$-\sum_{1}^{T}\frac{1}{2}(\boldsymbol{x}_t - \mathrm{B}\boldsymbol{x}_{t-1} - \mathrm{u})^\top \mathrm{Q}^{-1}(\boldsymbol{x}_t - \mathrm{B}\boldsymbol{x}_{t-1} - \mathrm{u}) - \sum_{1}^{T}\frac{1}{2}\log|\mathrm{Q}|$$
$$-\frac{1}{2}(\boldsymbol{x}_0 - \xi)^\top \Lambda^{-1}(\boldsymbol{x}_0 - \xi) - \frac{1}{2}\log|\Lambda| - \frac{n}{2}\log 2\pi$$

- If you can compute the marginal likelihood L(y; Θ), you can maximize that (using some Newton-based method, like BFGS). The Kalman filter will give you the marginal likelihood. Works great for lots of problems. But for many big multivariate problems it doesn't work so great. And it's too easy=boring

- A different approach to finding MLE parameters for problems with hidden states is the Expectation-Maximization (EM) algorithm. And it's elegant and hard=fun

Finding MLE parameters for MARSS models

Joint likelihood of y(data), x(hidden states)

$$\log \mathrm{L}(y, x; \Theta) = -\sum_{1}^{T} \frac{1}{2}(y_t - \mathrm{Z}x_t - \mathrm{a})^\top \mathrm{R}^{-1}(y_t - \mathrm{Z}x_t - \mathrm{a}) - \sum_{1}^{T} \frac{1}{2} \log|\mathrm{R}|$$

$$-\sum_{1}^{T} \frac{1}{2}(x_t - \mathrm{B}x_{t-1} - \mathrm{u})^\top \mathrm{Q}^{-1}(x_t - \mathrm{B}x_{t-1} - \mathrm{u}) - \sum_{1}^{T} \frac{1}{2} \log|\mathrm{Q}|$$

$$-\frac{1}{2}(x_0 - \xi)^\top \Lambda^{-1}(x_0 - \xi) - \frac{1}{2}\log|\Lambda| - \frac{n}{2}\log 2\pi$$

$$\log \mathrm{L}(y,x;\Theta) = f(y,x,\Theta)$$

- If you can compute the marginal likelihood L(y; Θ), you can maximize that (using some Newton-based method, like BFGS). The Kalman filter will give you the marginal likelihood. Works great for lots of problems. But for many big multivariate problems it doesn't work so great.

And when it does work, it's too easy=boring

- A different approach to finding MLE parameters for problems with hidden states is the Expectation-Maximization (EM) algorithm. Very robust.

And it's elegant and hard=fun

EM algorithm

Joint likelihood of y and x is log L(y,x;Θ) = f(y,x,Θ)

The EM algorithm maximizes the expected value of the joint likelihood

$$\mathrm{E}_{\mathbf{XY}}[\log \mathbf{L}(\boldsymbol{Y}, \boldsymbol{X}; \Theta); \boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j]$$

Expected value of the “random variable LL” conditioned on the observed data and a set of parameters

$$\mathrm{E}_{\mathbf{XY}}[\log \mathbf{L}(\boldsymbol{Y}, \boldsymbol{X}; \Theta); \boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j] =$$

g(E(YX), E(XX), E(YY), E(X), E(Y), Θ)

The expectations in this expected joint likelihood can be computed (for MARSS models with the Kalman smoother)

We can maximize g(…., Θ) with respect to Θ to find the Θ that maximizes the expected log likelihood.

EM algorithm for MARSS models

1. Start with Θ_1

2. Compute the expectations involving X and Y conditioned on Θ_1 and the data

3. Put those $\mathrm{E}_{\mathbf{XY}}[\log \mathbf{L}(\boldsymbol{Y},\boldsymbol{X};\Theta);\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j]$ and maximize with respect to Θ to get Θ_2

4. Compute the expectations involving X and Y conditioned on Θ_2 and the data

5. Put those $\mathrm{E}_{\mathbf{XY}}[\log \mathbf{L}(\boldsymbol{Y},\boldsymbol{X};\Theta);\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j]$ and maximize with respect to Θ to get Θ_3

6. Repeat until convergence

EM algorithm for MARSS models

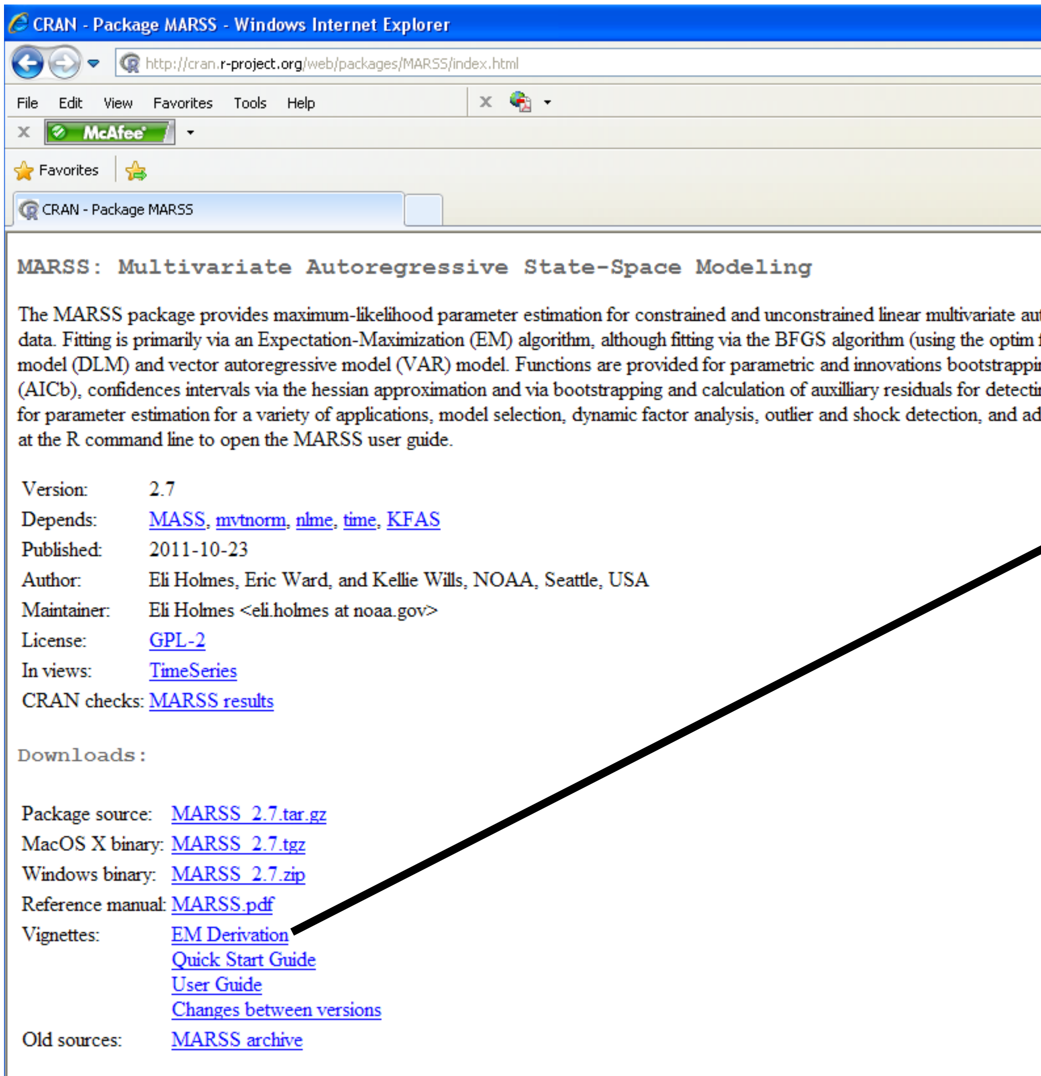
1. Start with Θ_1
2. Compute the expectations involving X and Y conditioned on Θ_1 and the data
3. Put those $\mathrm{E}_{\mathbf{XY}}[\log \mathbf{L}(\boldsymbol{Y},\boldsymbol{X};\Theta);\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j]$ and maximize with respect to Θ to get Θ_2
4. Compute the expectations involving X and Y conditioned on Θ_2 and the data
5. Put those $\mathrm{E}_{\mathbf{XY}}[\log \mathbf{L}(\boldsymbol{Y},\boldsymbol{X};\Theta);\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j]$ and maximize with respect to Θ to get Θ_3
6. Repeat until convergence

What's the point? Seems like a lot of pain!

1) It can make certain types of problems tractable and considerably faster
2) For many of the problems we work on, other approaches grind to a halt
3) The maximization steps (3, 5) and expectation steps (2,4) are analytical

"OMG! EM algorithms sound like fun!"

Google "MARSS cran"



CRAN - Package MARSS - Windows Internet Explorer

http://cran.r-project.org/web/packages/MARSS/index.html

File Edit View Favorites Tools Help

McAfee

Favorites

CRAN - Package MARSS

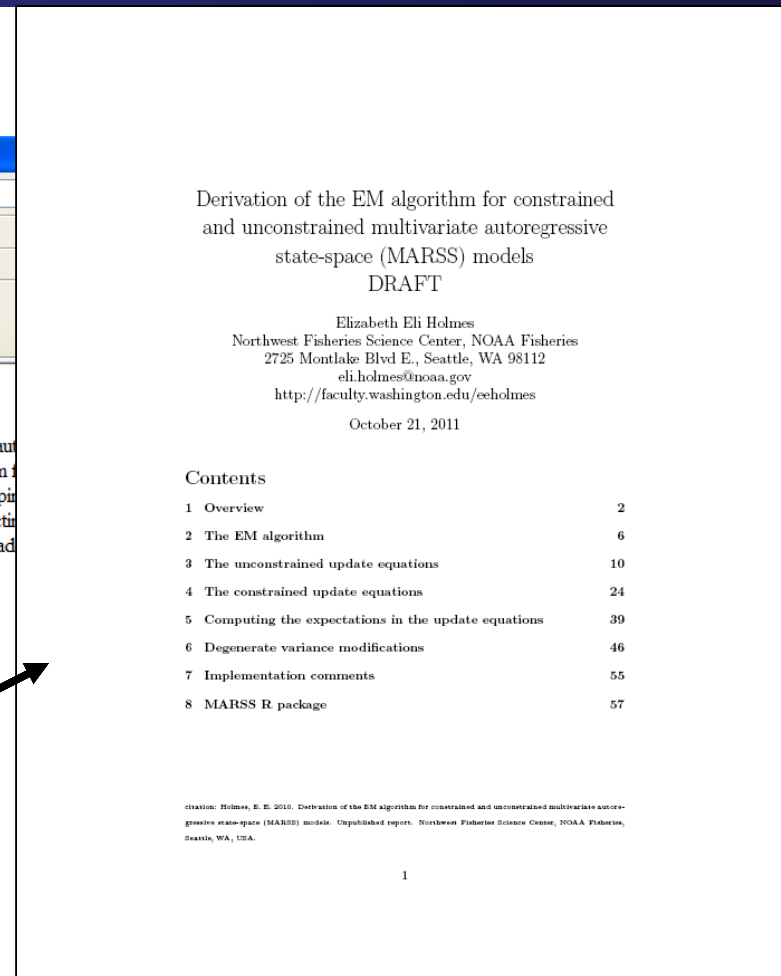
MARSS: Multivariate Autoregressive State-Space Modeling

The MARSS package provides maximum-likelihood parameter estimation for constrained and unconstrained linear multivariate aut
data. Fitting is primarily via an Expectation-Maximization (EM) algorithm, although fitting via the BFGS algorithm (using the optim f
model (DLM) and vector autoregressive model (VAR) model. Functions are provided for parametric and innovations bootstrappi
(AICb), confidences intervals via the hessian approximation and via bootstrapping and calculation of auxilliary residuals for detecti
for parameter estimation for a variety of applications, model selection, dynamic factor analysis, outlier and shock detection, and ad
at the R command line to open the MARSS user guide.

Version:	2.7
Depends:	MASS, mvtnorm, nlme, time, KFAS
Published:	2011-10-23
Author:	Eli Holmes, Eric Ward, and Kellie Wills, NOAA, Seattle, USA
Maintainer:	Eli Holmes <eli.holmes at noaa.gov>
License:	GPL-2
In views:	TimeSeries
CRAN checks:	MARSS results

Downloads:

Package source:	MARSS_2.7.tar.gz
MacOS X binary:	MARSS_2.7.tgz
Windows binary:	MARSS_2.7.zip
Reference manual:	MARSS.pdf
Vignettes:	EM Derivation Quick Start Guide User Guide Changes between versions
Old sources:	MARSS archive



Derivation of the EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models
DRAFT

Elizabeth Eli Holmes
Northwest Fisheries Science Center, NOAA Fisheries
2725 Montlake Blvd E., Seattle, WA 98112
eli.holmes@noaa.gov
http://faculty.washington.edu/eeholmes

October 21, 2011

Contents

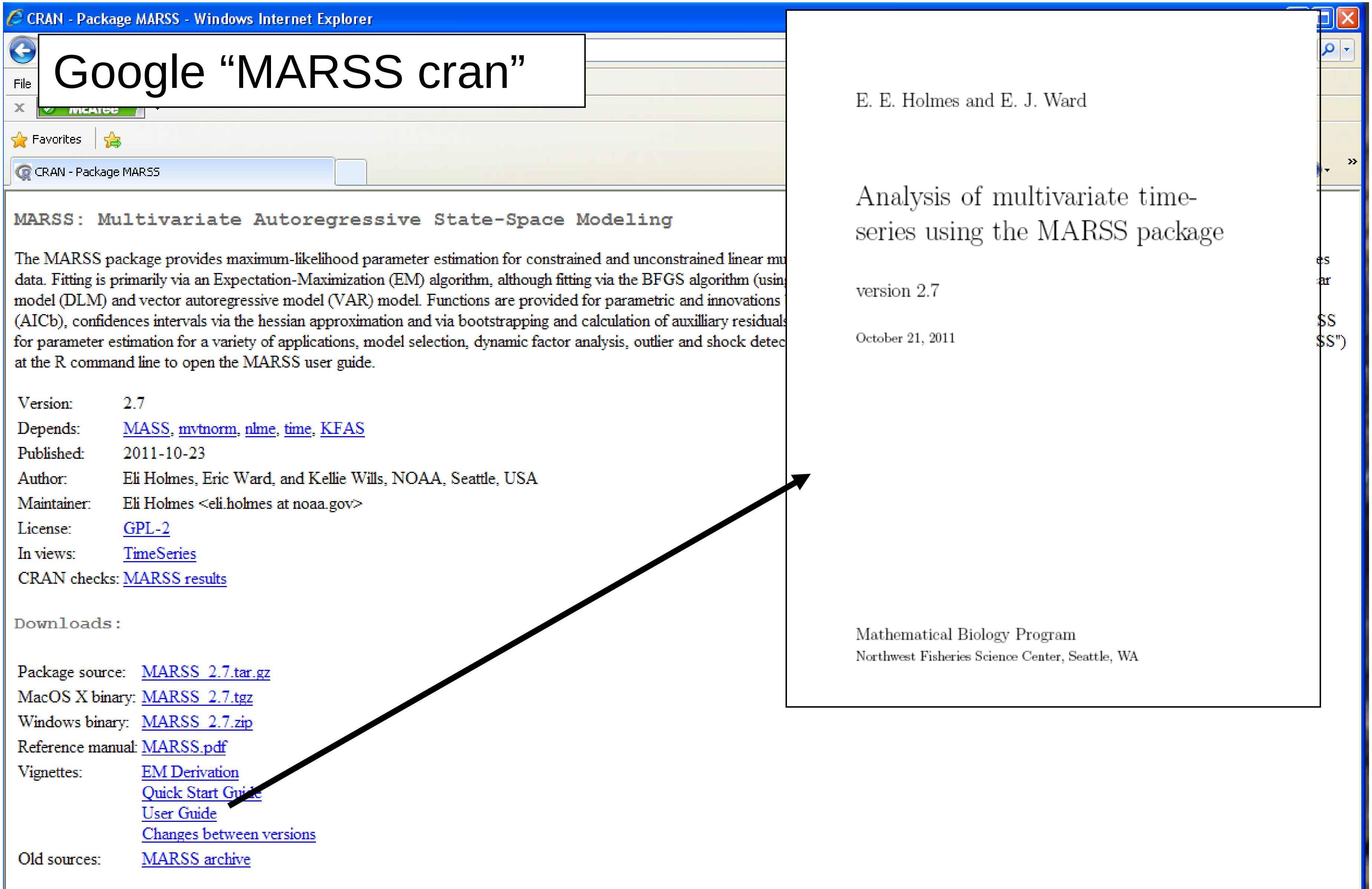
1 Overview 2
2 The EM algorithm 6
3 The unconstrained update equations 10
4 The constrained update equations 24
5 Computing the expectations in the update equations 39
6 Degenerate variance modifications 46
7 Implementation comments 55
8 MARSS R package 57

citation: Holmes, E. E. 2010. Derivation of the EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models. Unpublished report. Northwest Fisheries Science Center, NOAA Fisheries, Seattle, WA, USA.

1

"OMG! Do I really have to do all that math to use MARSS models?

Google "MARSS cran"



CRAN - Package MARSS - Windows Internet Explorer

File

Favorites

CRAN - Package MARSS

MARSS: Multivariate Autoregressive State-Space Modeling

The MARSS package provides maximum-likelihood parameter estimation for constrained and unconstrained linear mu... data. Fitting is primarily via an Expectation-Maximization (EM) algorithm, although fitting via the BFGS algorithm (usin... model (DLM) and vector autoregressive model (VAR) model. Functions are provided for parametric and innovations ... (AICb), confidences intervals via the hessian approximation and via bootstrapping and calculation of auxilliary residuals ... for parameter estimation for a variety of applications, model selection, dynamic factor analysis, outlier and shock detec... at the R command line to open the MARSS user guide.

Version:	2.7
Depends:	MASS, mvtnorm, nlme, time, KFAS
Published:	2011-10-23
Author:	Eli Holmes, Eric Ward, and Kellie Wills, NOAA, Seattle, USA
Maintainer:	Eli Holmes <eli.holmes at noaa.gov>
License:	GPL-2
In views:	TimeSeries
CRAN checks:	MARSS results

Downloads:

Package source:	MARSS_2.7.tar.gz
MacOS X binary:	MARSS_2.7.tgz
Windows binary:	MARSS_2.7.zip
Reference manual:	MARSS.pdf
Vignettes:	EM Derivation Quick Start Guide User Guide Changes between versions
Old sources:	MARSS archive

E. E. Holmes and E. J. Ward

Analysis of multivariate time-series using the MARSS package

version 2.7

October 21, 2011

Mathematical Biology Program

Northwest Fisheries Science Center, Seattle, WA

Lots of case studies and examples from workshops we (Eric Wark, Brice Semmens, Mark Scheuerell, and myself) have taught

12 Case Study 3: Using MARSS models to identify spatial population structure and covariance 103
12.1 The problem 103
12.2 How many distinct subpopulations? 104
12.3 Is Hood Canal separate? 107

13 Case Study 4: Dynamic factor analysis (DFA) using MARSS 111
13.1 Dynamic factor analysis 111
13.2 The data 113
13.3 Setting up the model 114
13.4 Fitting the model. 117
13.5 Using model selection to determine the number of trends 117
13.6 Using a varimax rotation to determine the trend loadings 120

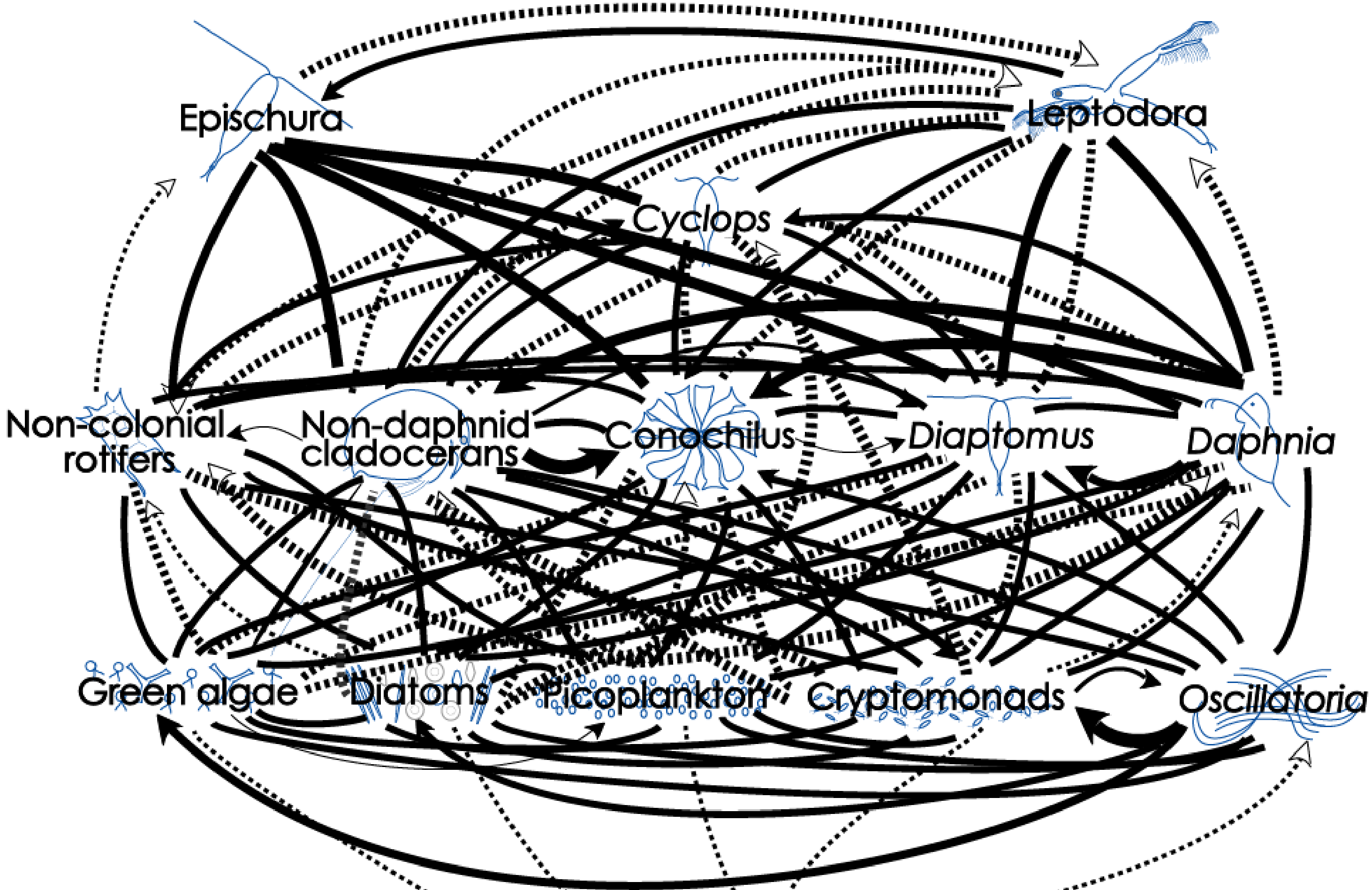
14 Case Study 5: Using state-space models to analyze noisy animal tracking data 123
14.1 A simple random walk model of animal movement 123
14.2 The problem 124
14.3 Estimate locations from bad tag data 124
14.4 Comparing turtle tracks to proposed fishing areas 128
14.5 Using specialized packages to analyze tag data 129

15 Case Study 6: Detection of outliers and structural breaks using MARSS 131
15.1 Detection of outliers and structural breaks 131

But the talk is about the statistical methods

- What is a multivariate autoregressive state-space model (MARSS or VARSS)?
- A tour of different classes of time series models written as MARSS (more math)
- Estimating parameters using an EM algorithm for MARSS models with linear constraints (more math)
- MARSS R package
- Estimating the species interaction matrix and covariate matrices for PLANKTON (actually more math)
- Stabilitiy metrics (cartoons!)
- Some results from the plankton work

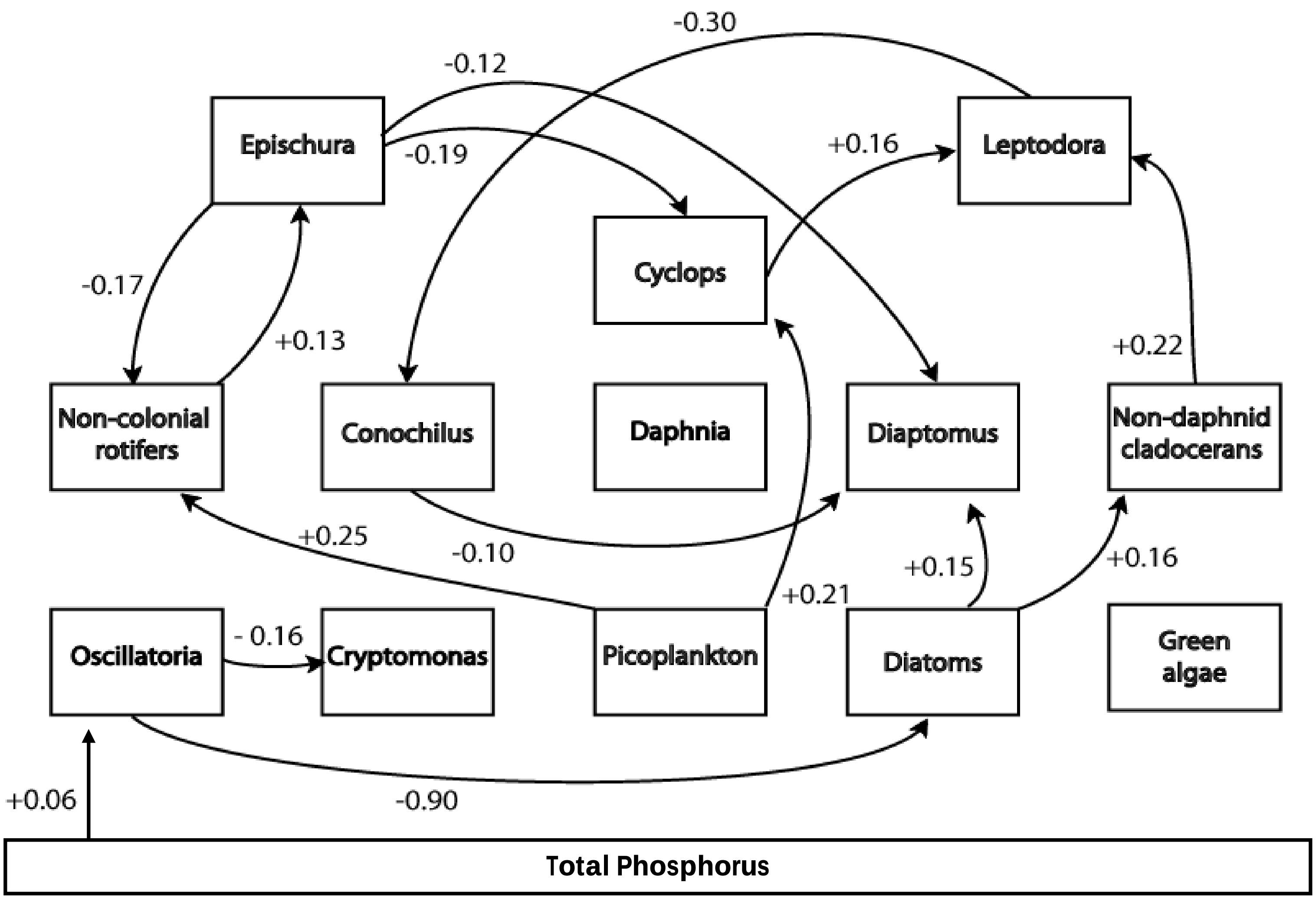
Why use MAR models to study community dynamics?



Temperature ... Nutrients .... Photoperiod ... Storm activity ... Fishing pressure ...

Lake Washington plankton community, S.E. Hampton, NCEAS, UCSB

What are the strong species interactions? How are environmental factors affecting species?



Hampton, Scheuerell, & Schindler 2006

Multispecies Autoregressive Models (MARs) as used in community modeling

Log abundance of species *i* at t+1

Effect of species *j* on species *i*

at t

Effect of environmental variable *k* on species *i*

$$x_i(t+1) = x_i(t) + u_i + \Sigma\, \alpha_{i,j}\, x_j(t) + \Sigma\, c_{i,k}\, \upsilon_k(t)$$

Level of environmental variable *k* at t

Species-specific constant

Log abundance of species *j* at t

Written in matrix form:

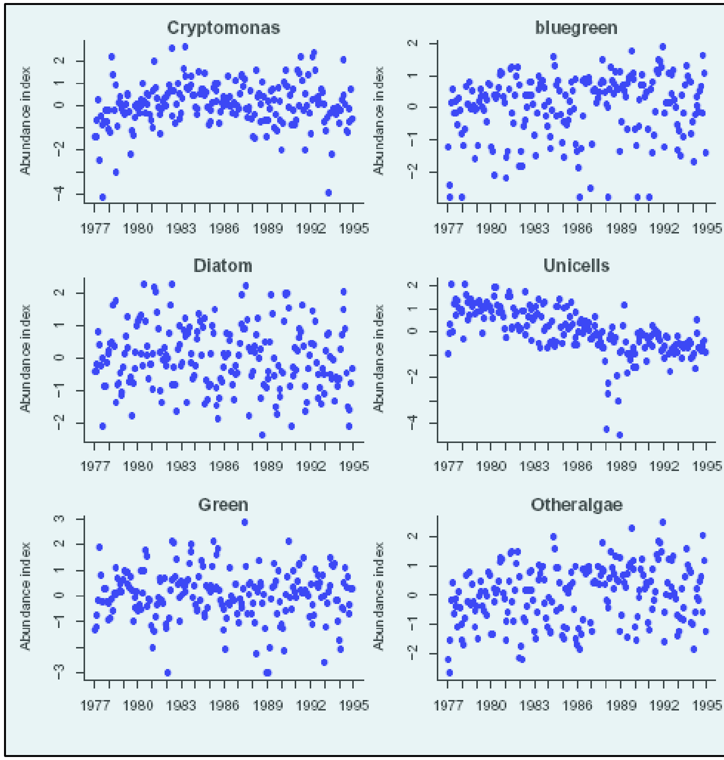
The effect of species 1 on species 2

$$\begin{bmatrix} x_{1,t} \\ x_{2,t} \\ x_{3,t} \end{bmatrix} = \underbrace{\begin{bmatrix} 1+\alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} \\ \alpha_{2,1} & 1+\alpha_{2,2} & \alpha_{2,3} \\ \alpha_{3,1} & \alpha_{3,2} & 1+\alpha_{3,3} \end{bmatrix}}_{\text{B = interaction matrix}} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \\ x_{3,t-1} \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \underbrace{\begin{bmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \\ c_{3,1} & c_{3,2} \end{bmatrix}}_{\text{covariate effect}} \overbrace{\begin{bmatrix} \upsilon_{1,t-1} \\ \upsilon_{2,t-1} \end{bmatrix}}^{\text{covariates}} + \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

The effect of covariate 1 on species 1

Environmental variation (not from covariates)

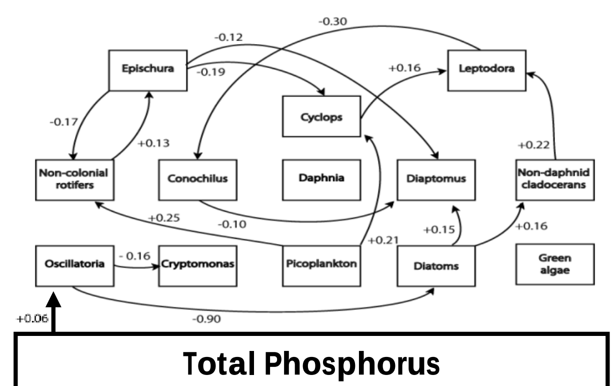
Ives, Dennis, Cottingham, & Carpenter. 2003. Ecol. Monogr. 73(2)



$$\mathbf{x}_t = \mathbf{B}\mathbf{x}_{t-1} + \mathbf{u} + \mathbf{C}\mathbf{f}_t + \mathbf{w}_t;$$

Model selection step to find the "0" interactions

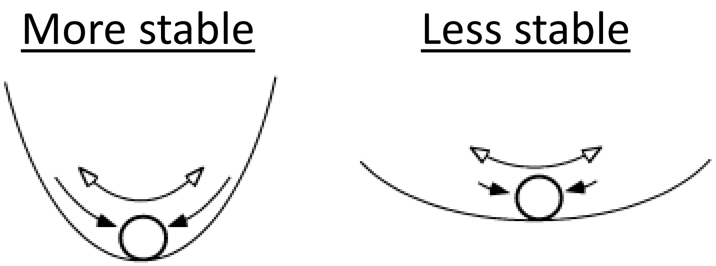
Bootstrap step to find the strong interactions



BUT MAR models
- No observation errors
- Covariates known perfectly
- No (not many) missing values

Stability properties of MAR models

- A variety of different stability metrics can be estimated from the B matrix
- Distinguish underlying 'system' stability from environmentally driven variation



Stability measure	More stable when…
Variance	variance of the stationary distribution is low relative to that for the process error
Return rate	rapid approach to the stationary distribution (i.e. high return rate)
Reactivity	fewer departures from the mean of the stationary distribution (i.e. low reactivity)

Lots of applications to freshwater plankton datasets

Interaction matrix

$$\mathbf{x}_t = \mathbf{B}\mathbf{x}_{t-1} + \mathbf{u} + \mathbf{C}\mathbf{f}_t + \mathbf{w}_t,$$

$$\cancel{\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{D}\mathbf{f}_t + \mathbf{v}_t,}$$

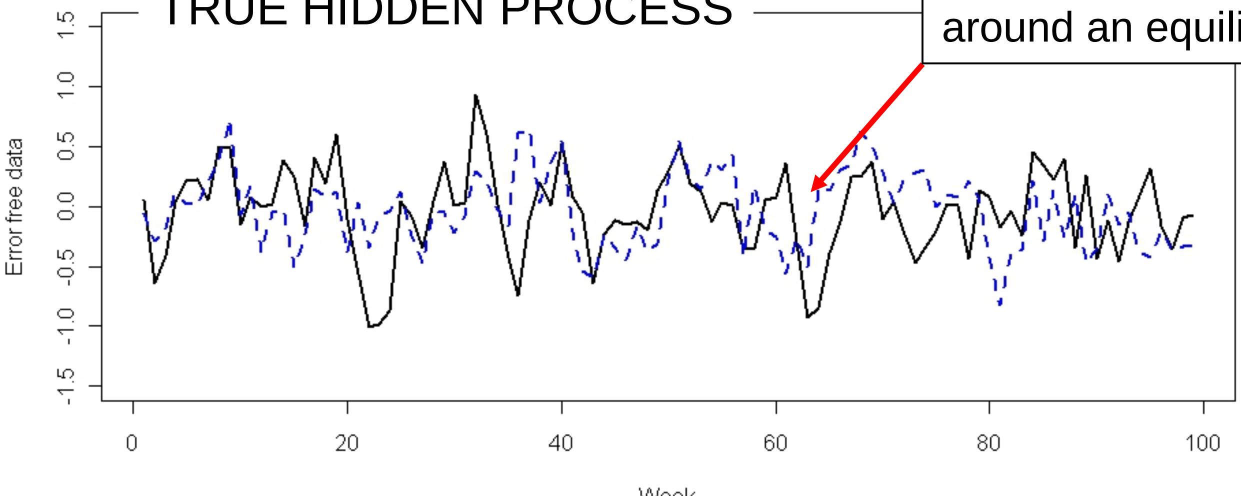
Assume the data have no observation error

Citation	System
Hall et al 2009	Freshwater plankton
Hampton et al 2008	Freshwater plankton
Duffy 2007	Freshwater plankton
Hampton et al 2006	Freshwater plankton
Huber and Gaedke 2006	Freshwater plankton
Hampton and Schindler 2006	Freshwater plankton
Carpenter et al 2005	Freshwater plankton
Beisner et al 2003	Freshwater plankton
Klug et al 2000	Freshwater plankton
Hampton et al 2006	Freshwater plankton
Fischer et al 2001	Freshwater plankton
Ives et al 1999	Freshwater plankton
Ives et al 2003	Freshwater plankton
Klug and Cottingham 2001	Freshwater plankton

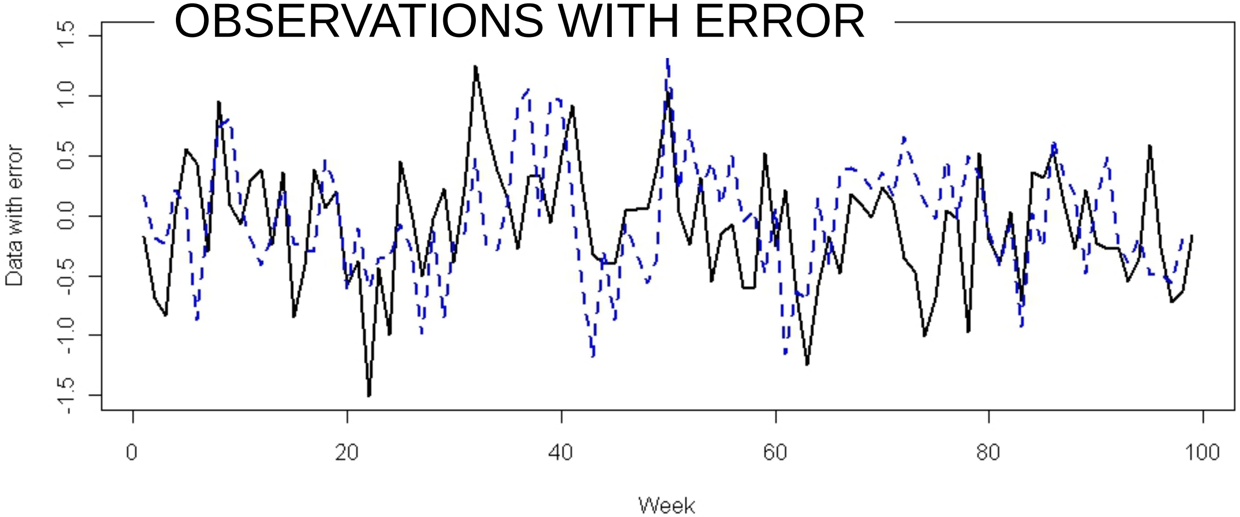
Observation error and spurious density-dependence

Simple predator-prey interaction matrix $\begin{bmatrix} 0.5 & 0.1 \\ -0.1 & 0.5 \end{bmatrix}$

Strong density-dependence leads to tight fluctuations around an equilibrium



No observation error; blue has negative effect on black; black has positive effect on blue

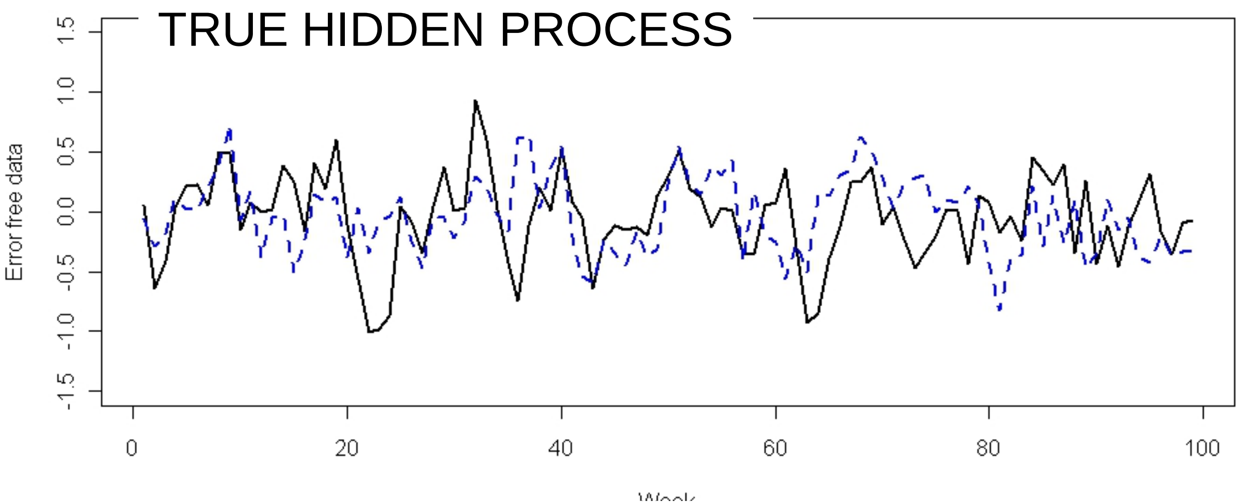


Added (high) observation error; more overall variation; more ‘noisy’

Observation error and spurious density-dependence

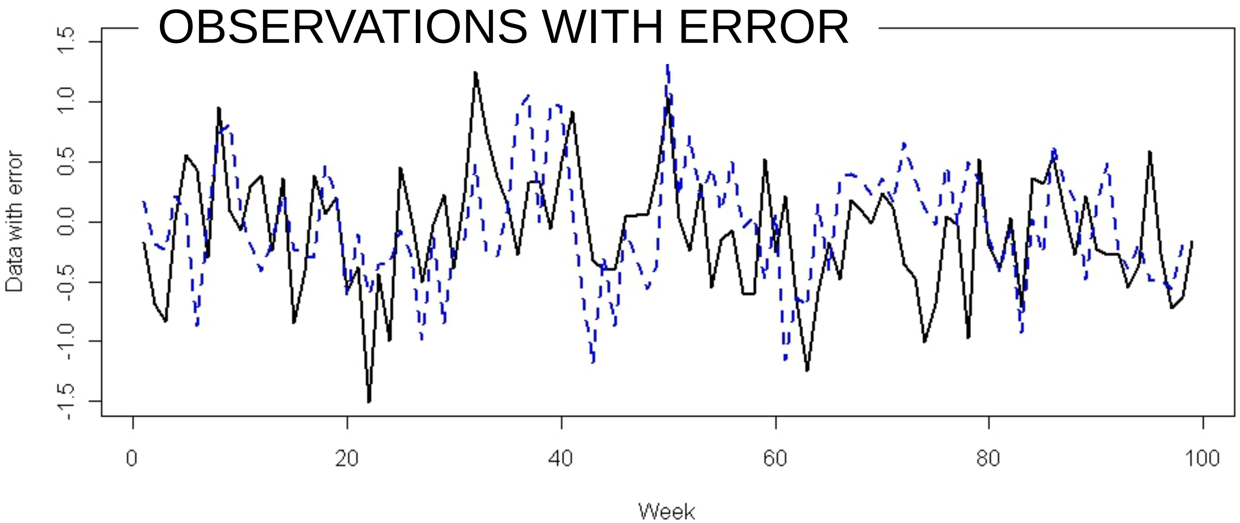
Original interaction matrix

$$\begin{bmatrix} 0.5 & 0.1 \\ -0.1 & 0.5 \end{bmatrix}$$



Ignore observation error and density-dependence looks stronger and interactions weaker

$$\begin{bmatrix} 0.28 & 0.05 \\ -0.05 & 0.28 \end{bmatrix}$$



Include observation error (with MARSS model) and we recover the original matrix…at a cost

$$\begin{bmatrix} 0.5 & 0.1 \\ -0.1 & 0.5 \end{bmatrix}$$

Not a new result but perhaps not widely recognized… unknown obs error = spurious density-dependence

ECOLOGY LETTERS

Ecology Letters, (2011) doi: 10.1111/j.1461-0248.2011.01702.x

LETTER

Are patterns of density dependence in the Global Population Dynamics Database driven by uncertainty about population abundance?

Jonas Knape* and Perry de Valpine
Department of Environmental Science, Policy and Management, 137 Mulford Hall #3114, University of California, Berkeley, Berkeley, CA 94720, USA
**Correspondence: E-mail: jknape@berkeley.edu*

Abstract
Density dependence in population growth rates is of immense importance to ecological theory and application, but is difficult to estimate. The Global Population Dynamics Database (GPDD), one of the largest collections of population time series available, has been extensively used to study cross-taxa patterns in density dependence. A major difficulty with assessing density dependence from time series is that uncertainty in population abundance estimates can cause strong bias in both tests and estimates of strength. We analyse 627 data sets in the GPDD using Gompertz population models and account for uncertainty via the Kalman filter. Results suggest that at least 45% of the time series display density dependence, but that it is weak and difficult to detect for a large fraction. When uncertainty is ignored, magnitude of and evidence for density dependence is strong, illustrating that uncertainty in abundance estimates qualitatively changes conclusions about density dependence drawn from the GPDD.

Keywords
Density dependence, GPDD, observation error, time series.

Ecology Letters (2011)

INTRODUCTION

Density dependence in population growth rates is a fundamental concept for ecological theory as well as for population management. Estimating density dependence in wild populations has, however, proved challenging. Ideally, density dependence in growth rates should be estimated directly from the effects of density acting on the traits contributing to population growth. Given current progress in statistical methods for jointly analysing data on both population size and demographic traits (Besbeas *et al.* 2005), and with long-term population studies involving demographic data becoming increasingly common, this approach holds a bright future. However, the number of such studies is currently limited and they only cover a rather narrow range of taxa. Long-term time series on population abundance are more common and can be used to estimate density dependence in population growth rates. Under this approach, density dependence is defined as a general tendency of per capita growth rates to decrease when population size is large and increase when it is small, and is identified as a statistical pattern not tied to any specific biological mechanism (Wolda & Dennis 1993).

It was noted early that estimates and tests of density dependence based on regressing log transformed current observed population size, y_t, on previous log transformed observed population size, y_{t-1}, are sensitive to uncertainty in the observations (St-Amant 1970; Kuno 1971; Itô 1972; Slade 1977). Similar concerns were aired about estimates from fisheries models of stock-recruitment data (Ludwig & Walters 1981; Walters & Ludwig 1981). Uncertainty inflates the Type I error rate of tests for density dependence (Shenk *et al.* 1998) and tends to bias estimates towards stronger density dependence if dynamics are under-compensatory and towards weaker density dependence if dynamics are over-compensatory (Benson 1973). Bulmer (1975) devised two tests for density dependence taking the time series nature of the data into account. One of those was designed to be robust against uncertainty about population size and has been shown to perform better than density dependence tests ignoring uncertainty in estimates of population abundance (Shenk *et al.* 1998). Simple procedures to correct for effects of uncertainty such as the SIMEX method have been suggested (Solow 1998; Freckleton *et al.* 2006) but typically require that the variance of the uncertainty about population size is known. A more direct approach to account for uncertainty is provided by state space models, first used for modelling population dynamics in the fisheries literature (e.g. Mendelssohn 1988; Sullivan 1992). State space models in these cases consist of a model of a population dynamical process combined with a model of the uncertainty in the abundance estimates, sometimes termed observation, measurement or sampling error, and may be used to estimate the variance of this uncertainty as well as to filter out its effects (de Valpine & Hastings 2002; Calder *et al.* 2003; Buckland *et al.* 2004; Dennis *et al.* 2006). Estimates derived from state space models tend to have smaller bias than estimates ignoring uncertainty about population abundance, but can also have large variances (Knape 2008), and the statistical properties of even simple state space model estimators are not fully understood (Dennis *et al.* 2006; Lebreton 2009).

The Global Population Dynamics Database (GPDD), containing over 5000 time series on population abundances obtained from various forms of population surveys, has provided an opportunity for ecologists to explore population dynamical patterns over a wide range of taxa (Inchausti & Halley 2001). Analyses using data in the GPDD have focused on, e.g., extinction risks (Fagan *et al.* 2001; Inchausti & Halley 2003; Brook *et al.* 2006), population cycles (Kendall *et al.* 1998; Murdoch *et al.* 2002) and effects of weather (Knape & de Valpine 2011) but, arguably, the studies stirring the most attention as well as debate have addressed population regulation and density dependence. These have explored patterns in the shape of density dependence (Sibly *et al.* 2005; Polansky *et al.* 2009) and in the strength of regulation and density dependence (Brook & Bradshaw 2006; Sibly *et al.* 2007;

2010

Ecology, 89(11), 2008, pp. 2994–3000

Reports

ESTIMABILITY OF DENSITY DEPENDENCE IN MODELS OF TIME SERIES DATA

JONAS KNAPE[1]

Department of Theoretical Ecology, Ecology Building, Lund University, SE-223 62, Lund, Sweden

Abstract. Estimation of density dependence from time series data on population abundance is hampered in the presence of observation or measurement errors. Fitting state–space models has been proposed as a solution that reduces the bias in estimates of density dependence caused by ignoring observation errors. While this is often true, I show that, for specific parameter values, there are identifiability issues in the linear state–space model when the strength of density dependence and the observation and process error variances are all unknown. Using simulation to explore properties of the estimators, I illustrate that, unless assumptions are imposed on the process or observation error variances, the variance of the estimator of density dependence varies critically with the strength of the density dependence. Under compensatory dynamics, the stronger the density dependence the more difficult it is to estimate in the presence of observation errors. The identifiability issues disappear when density dependence is estimated from the state–space model with the observation error variance known to the correct value. Direct estimates of observation variance in abundance censuses could therefore prove helpful in estimating density dependence but care needs to be taken to assess the uncertainty in variance estimates.

Key words: density dependence; state–space models; time series analysis.

INTRODUCTION

Density dependence can be loosely defined as a quantitative influence of population size on some life history or population trait of interest. The concept is of central importance to population ecology since it determines both the limiting and the short time behavior of the dynamics of populations. Empirical estimates of density dependence are therefore important from a scientific as well as from a management perspective. Assessment of density dependence in the dynamics of natural populations has however proved to be challenging (Dennis et al. 2006).

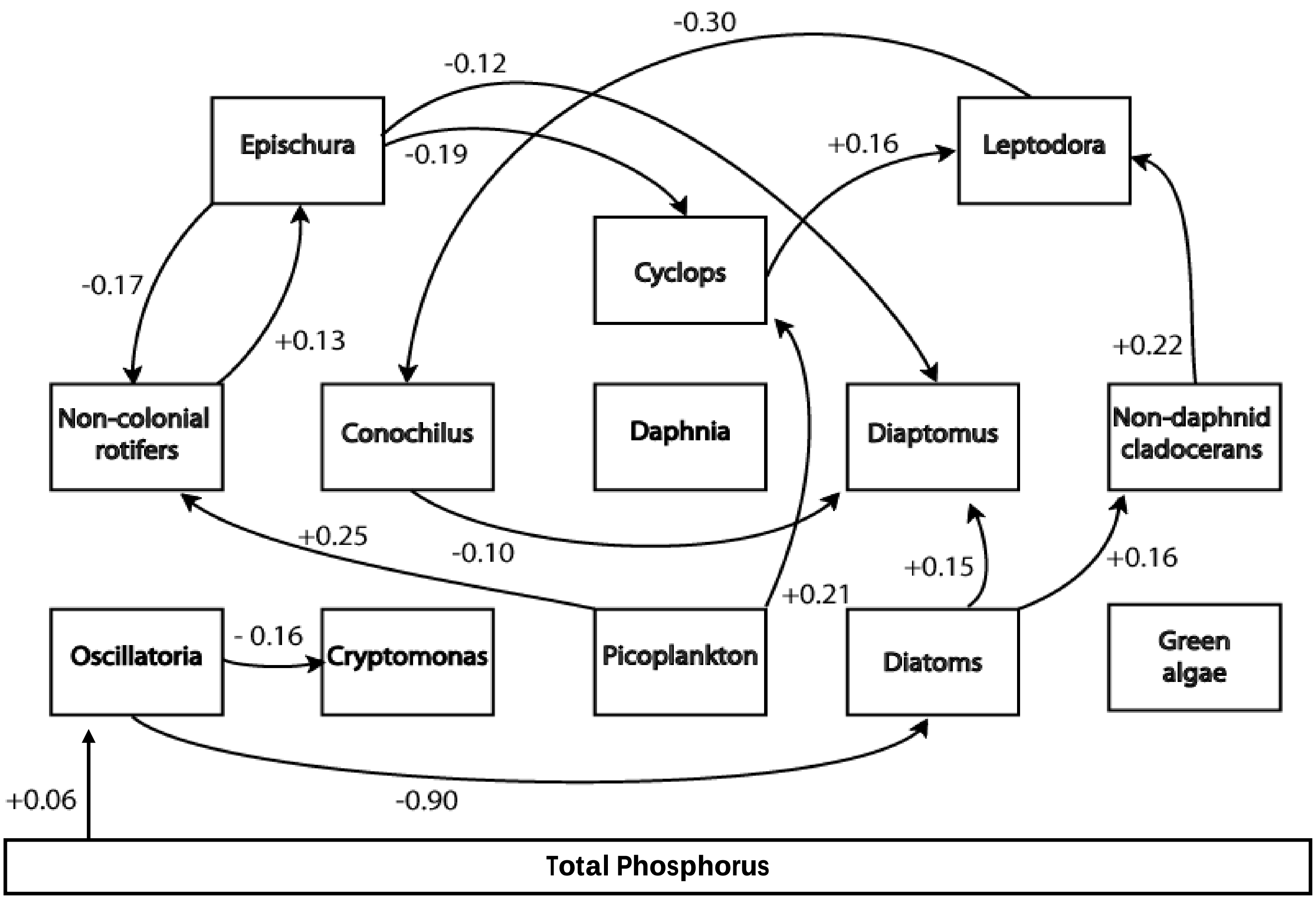
When relevant data are available, effects of density dependence can be directly linked to life history traits. For instance, density dependence in recruitment (e.g., Crespin et al. 2006) and survival (e.g., Festa-Bianchet et al. 2003) have been estimated by mark–recapture analyses and density dependence in fecundity has been inferred from data on reproduction (e.g., Solbreck and Ives 2007). Density dependence in life history traits influences density dependence in population growth rate (Lande et al. 2002). It can be argued that density dependence in population growth is the most important form of density dependence for determining long-term behavior of populations. However, since the link from demographic traits to population change is almost never known with good precision, density dependence in population growth rate is not easily inferred from life history data even if the effects of density dependence on several life history traits are well known. Time series analysis of population abundance data provides an alternative or complementary method that ideally could serve as a more direct way of estimating density dependence in population growth rate.

Estimates of density dependence must rely on measures of population density that are usually difficult to obtain with precision (Freckleton et al. 2006). This problem is particularly relevant to estimates of density dependence in growth rate derived from time series data on population size in that both the dependent and the independent variable are measured with uncertainty. Introducing observation error to dynamical data changes its dynamical structure (Dennis et al. 2006) and estimators relating to the dynamics of the data that do not account for observation errors are therefore often biased. Specifically, tests and estimators of density dependence based on time series data are known to be biased if observation errors are present but ignored for both direct (Kuno 1971, Walters and Ludwig 1981, Shenk et al. 1998, Freckleton et al. 2006) and delayed (Solow 2001) density dependence. An appealing method for overcoming this difficulty is provided by the state–space framework (Harvey 1990), a general term for statistical models of observations of hidden state variables that are dynamically linked through time. For time series data on population abundance, state–space models can be used for explicit modeling of both the observation and the population dynamical processes (Stenseth et al. 2003, Jamieson and Brooks 2004).

Manuscript received 12 January 2008; revised 2 June 2008; accepted 12 June 2008. Corresponding Editor: M. Lavine.
[1] E-mail: jonas.knape@teorekol.lu.se

2008

What are the effects of observation error on estimates of multivariate B matrices?



Hampton, Scheuerell, & Schindler 2006

Lake Washington long-term plankton monitoring

- weekly plankton sampling 1960s to present
- environmental covariate data
- standardized sampling
- basis for lots of MAR-based research into plankton community dynamics

Seattle

Lake Washington

Comparison of the B matrix estimates analysis of Lake WA data mid-1970s on

MARSS

R est=.02 or so

Q est=.6 or so

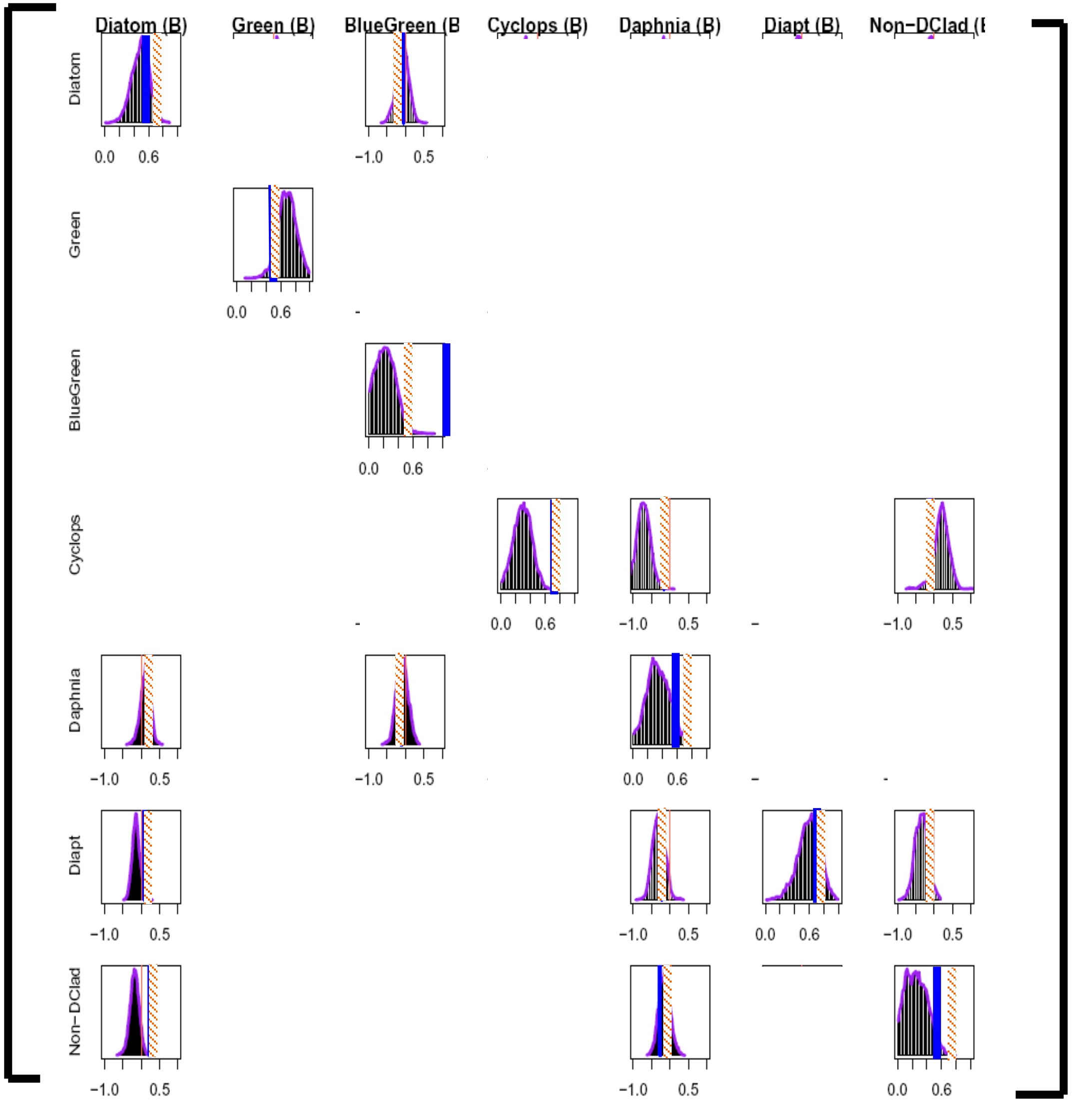
1	0.65		-0.08				
2		0.36					
3			0.34				
4				0.64	-0.17		-0.16
5	0.11		0.01		0.74		
6	0.11				-0.26	0.71	-0.15
7	0.06				-0.13		0.57

ORIGINAL

1	0.52		-0.06					Diatom
2		0.44						Green
3			0.92					Cyano
4				0.66	-0.16		-0.12	Cyclops
5	0.23		-0.11		0.57			Daphnia
6	0.09				-0.22	0.64	-0.14	Diatopmis
7	0.23				-0.27		0.52	Bosmina

Those results assumed we knew where the zeros were. What if we don't know?

This is the same 7x7 interaction matrix. The distributions are posterior distributions. All B elements were estimated but I blocked out the original "zeros"



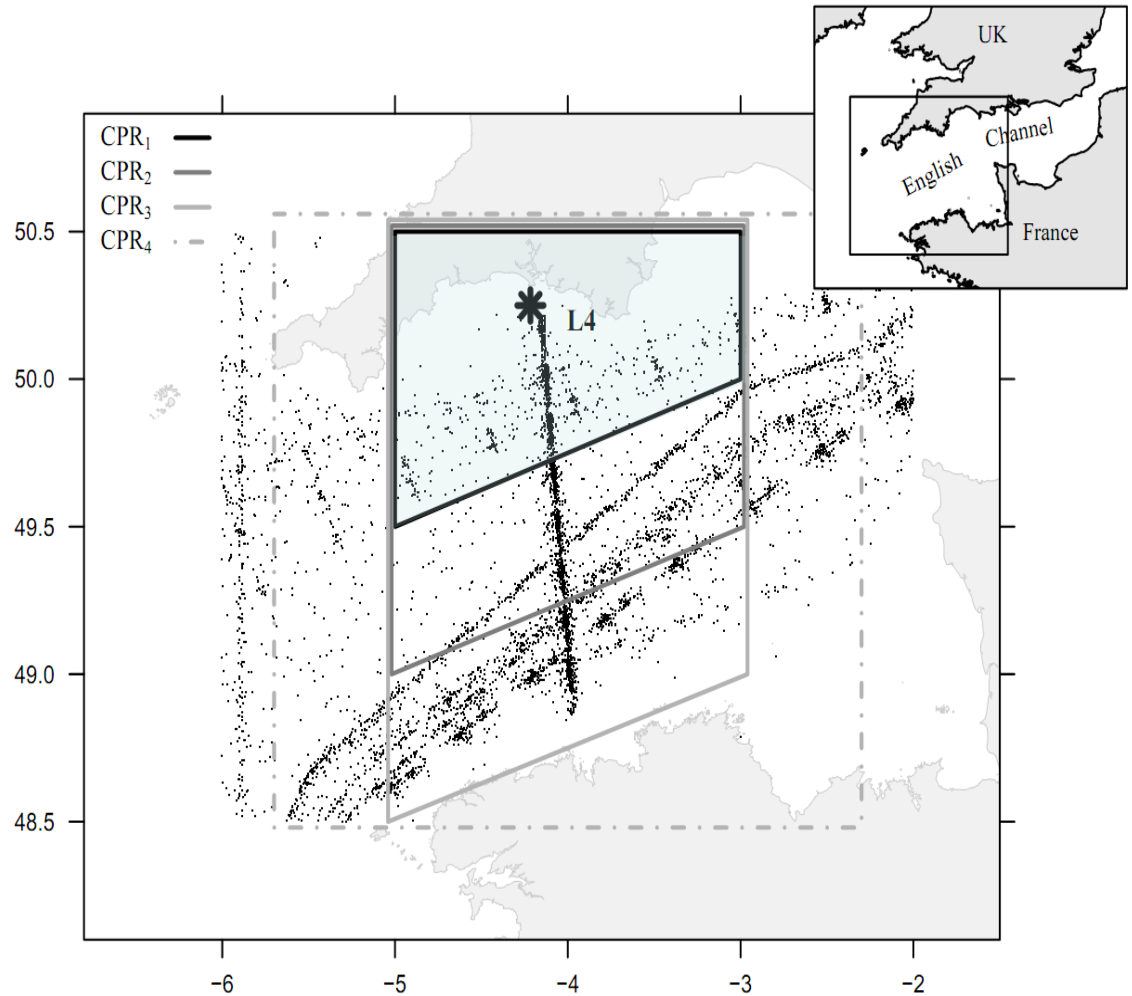
Long-term plankton studies in the English Channel

"L4" data (cleaner)

- **1 location**
- **Weekly samples at standard time of day**
- **Individual counts**
- **Environmental covariate info**
- **Very few missing values**

Continuous Plankton Recorder (CPR) data (noisy)

- **"platforms of opportunity" (ferries, non-research and research ships)= Many locations**
- **Log10 counts**
- **Times of day variable**
- **Lots of missing values**
- **Some spp poorly sampled**



14 groups

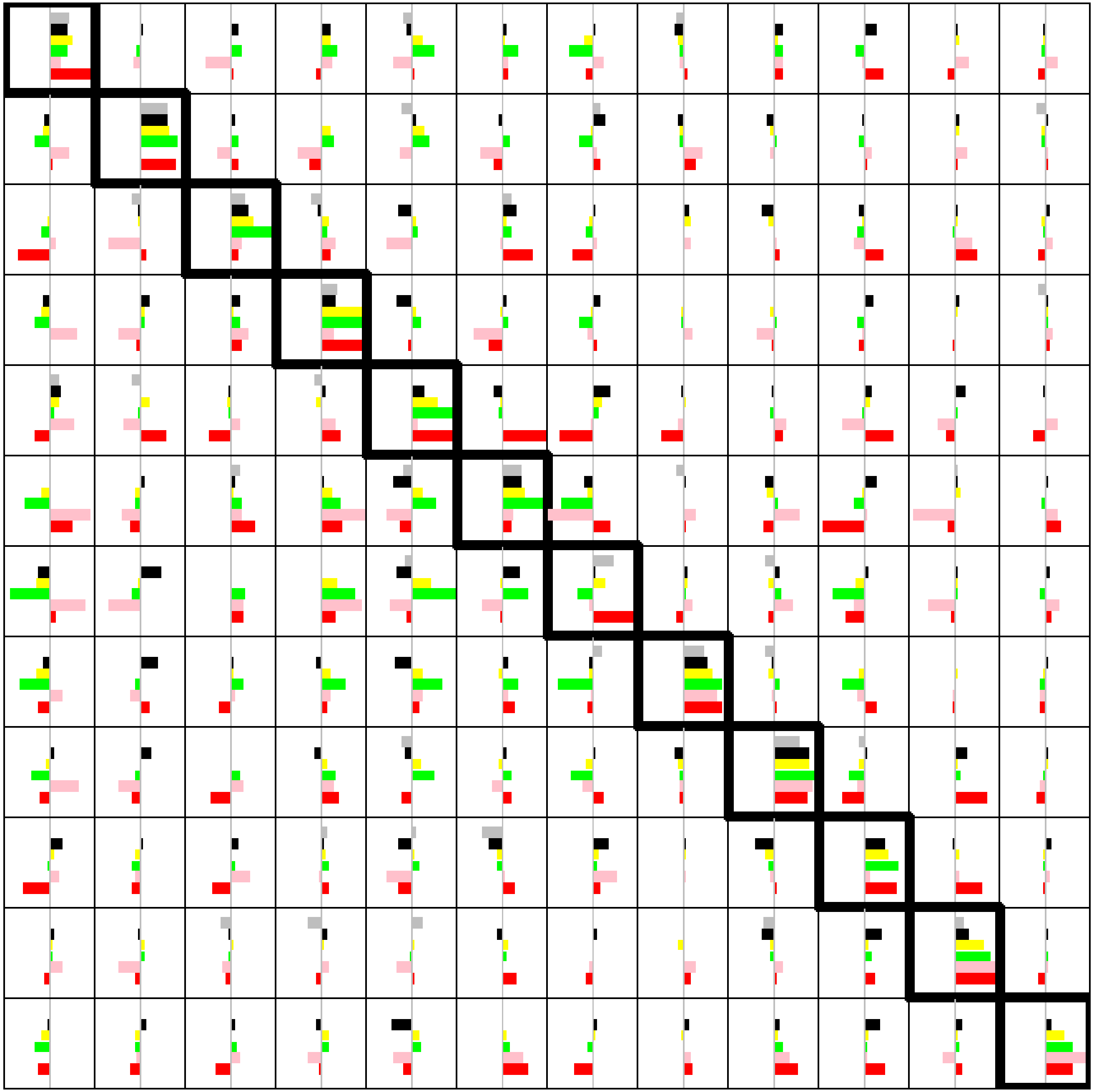
	Group	Proportion of community L4	Proportion of community CPR	Taxa Included	Proportion of group L4 & CPR mean
	Chaetognaths	0.02	0.07	Sagitta spp.	~1.00
	Pteropods	0.01	0.02	Thecosomata	>0.99
	Tunicates	0.03	0.07	Appendicularians Doliolids	0.99 0.01
	Cladocerans	0.05	0.04	Evadne spp. Podon spp.	0.66 0.34
	Amphipods	<0.01	<0.01	Gammarid amphipods Hyperiid amphipods Isopods Mysid shrimp	0.94 0.03 0.02 0.01
	Krill	<0.01	<0.01	Euphausiids	~1.00
Copepods	Large calanoids	0.03	0.08	Calanus spp. Metridia spp. Candacia spp. Eucalanus spp.	0.95 0.03 0.01 0.01
Copepods	Small calanoids	0.38	0.45	Pseudocalanus spp. Acartia spp. Temora spp. Paracalanus spp. Centropages spp. Clausocalanus spp. Ctenocalanus spp.	0.33 0.28 0.15 0.12 0.06 0.02 0.01
Copepods	Cyclopoids	0.12	0.02	Oithona spp.	~1.00
Copepods	Poecilostomatoids	0.19	0.01	Corycaeus spp. Oncaea spp.	0.51 0.49
Copepods	Harpacticoids	0.01	<0.01	Euterpina spp. Clytemnestra spp. Microsetella spp. Alteutha spp.	0.70 0.23 0.05 0.01
Meroplankton	Cirripedia	0.08	0.01	Cirripede larvae	1.00
Meroplankton	Mero. grazers (miscellaneous)	0.06	0.23	Echinoderm larvae Bivalve larvae Cyphonaute larvae Polychaete larvae Gastropod larvae	0.66 0.19 0.05 0.05 0.04
Meroplankton	Decapod larvae	0.01	0.01	Crab & shrimp larvae	1.00

B matrix

- **12 x 12**
- **2 groups removed because they had only 3 levels in the CPR data**

L4 (clean data)
- Green
- Yellow (ignores obs error)

CPR (noisy data)
- Red
- Pink (ignores obs error)



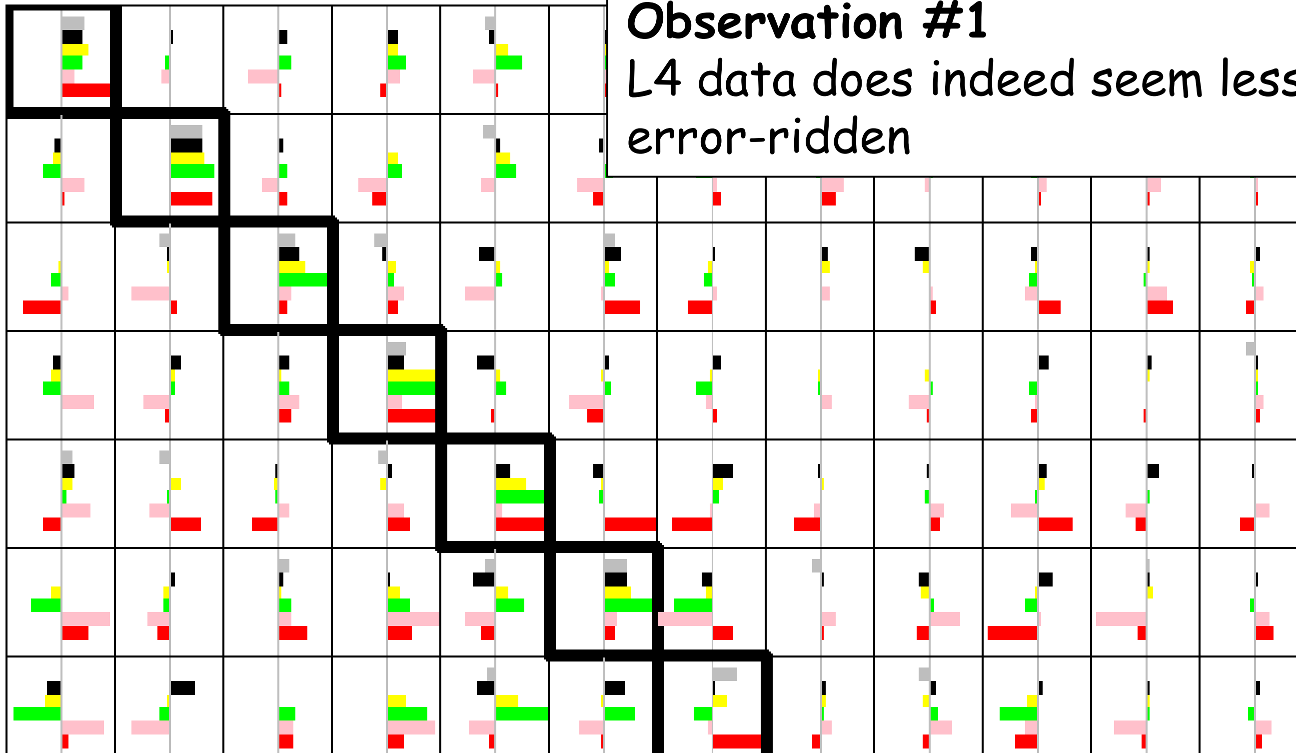
L4 (clean data)

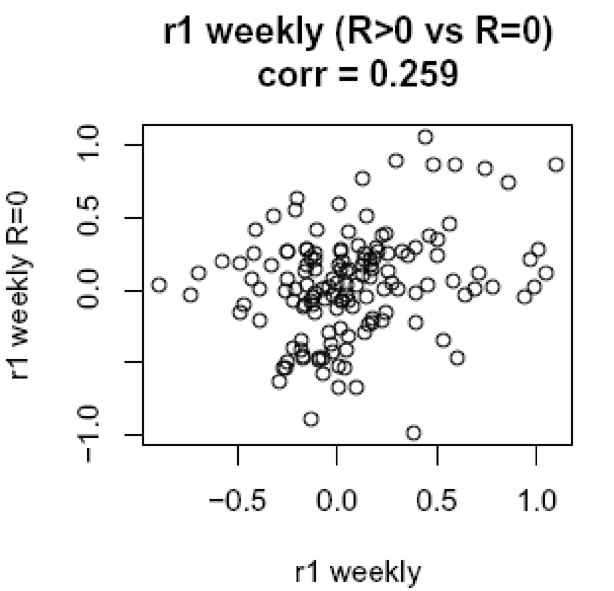
- Green
- Yellow (ignores obs error)

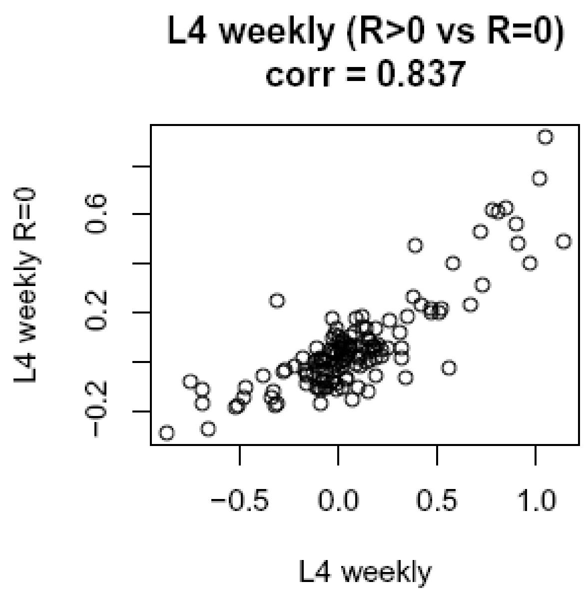
CPR (noisy data)

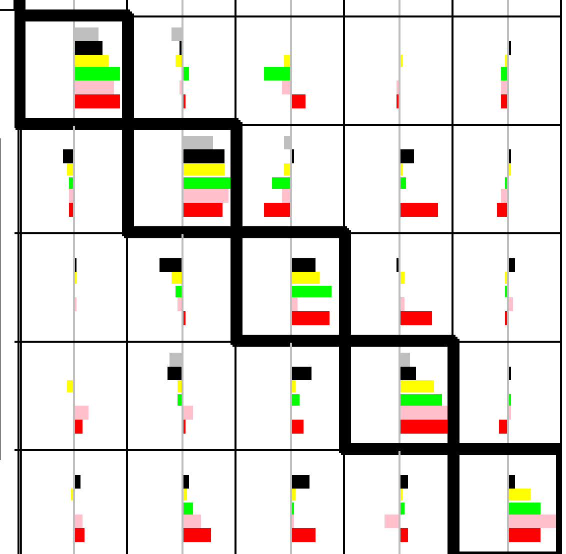
- Red
- Pink (ignores obs error)

Observation #1
L4 data does indeed seem less error-ridden



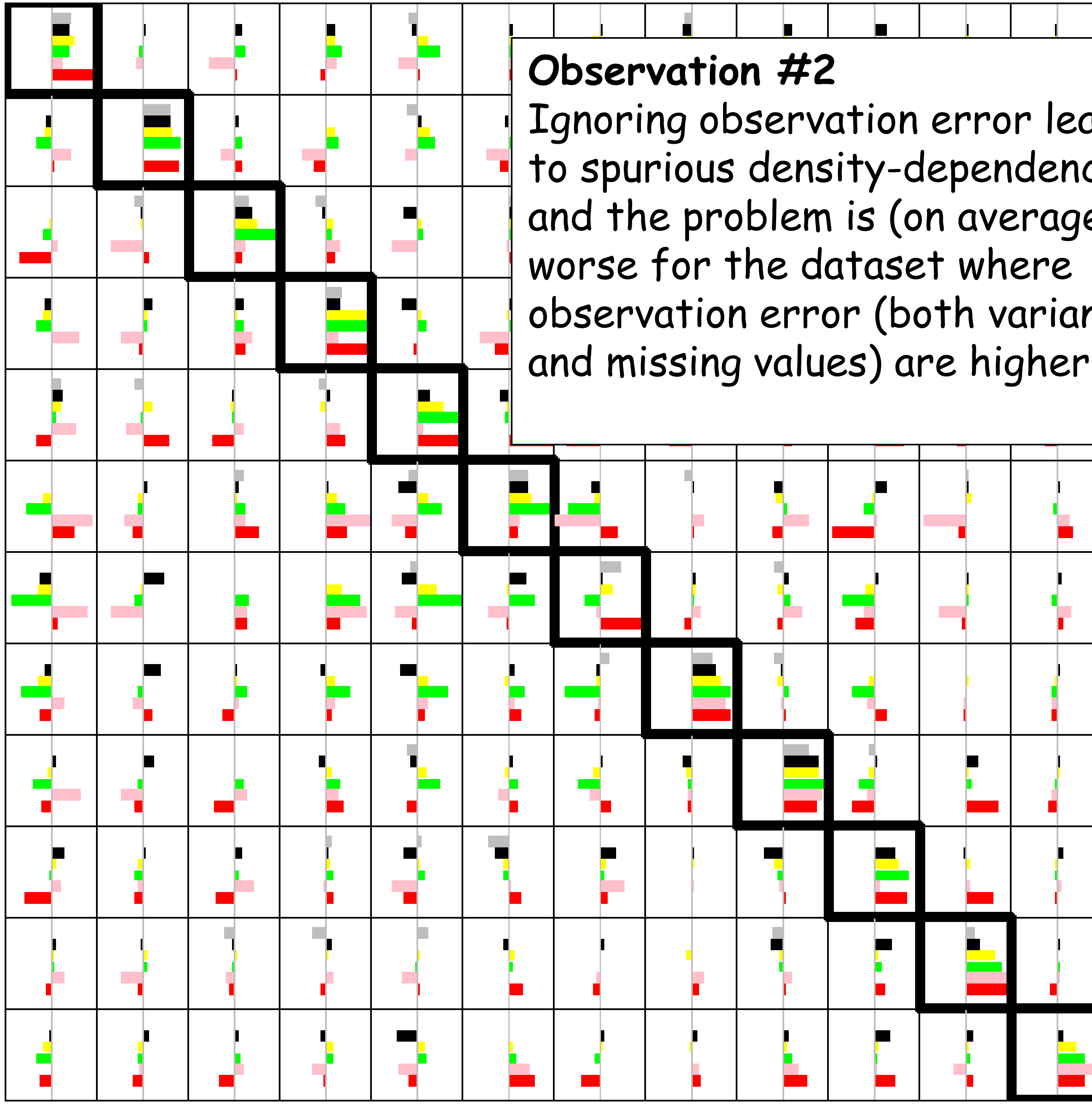






B matrix

- **12 x 12**
- **2 groups removed because they had only 3 levels in the CPR data**



Observation #2
Ignoring observation error leads to spurious density-dependence and the problem is (on average) worse for the dataset where observation error (both variance and missing values) are higher

L4 (clean data)
- Green
- Yellow (ignores obs error)

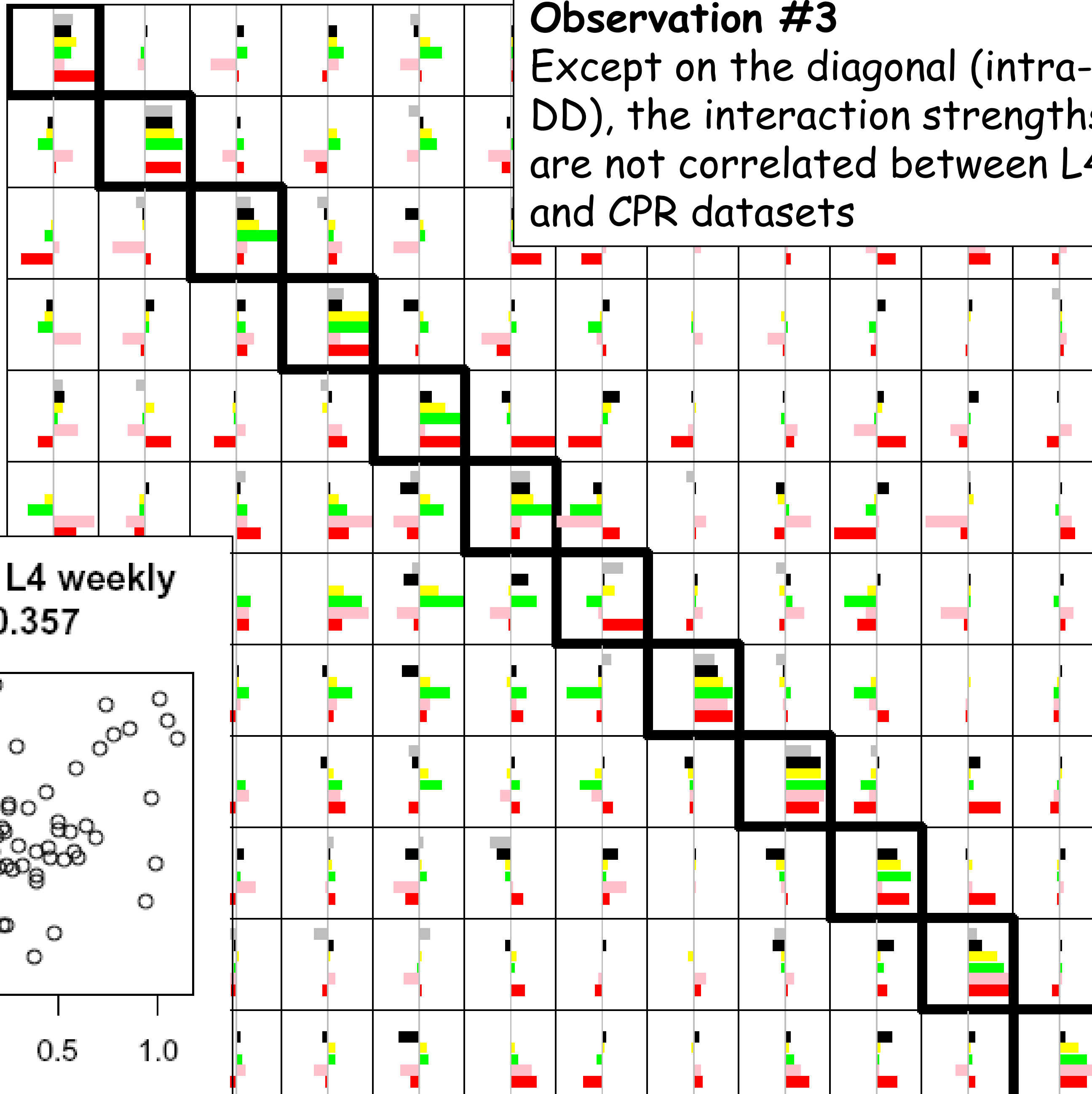
CPR (noisy data)
- Red
- Pink (ignores obs error)

L4 (clean data)

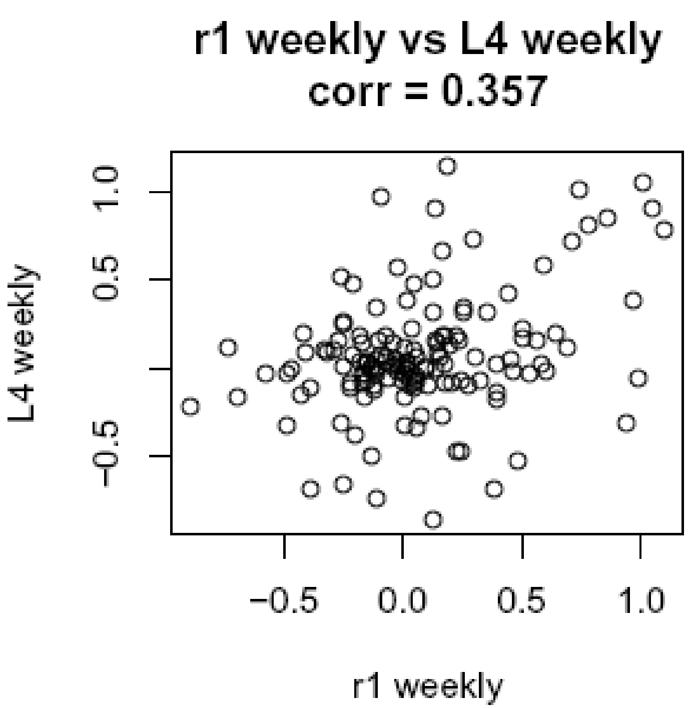
- Green
- Yellow (ignores obs error)

CPR (noisy data)

- Red
- Pink (ignores obs error)



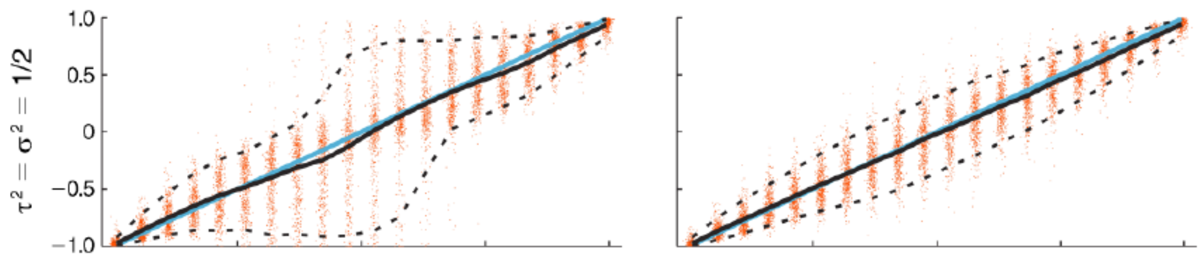
Observation #3
Except on the diagonal (intra-DD), the interaction strengths are not correlated between L4 and CPR datasets



Gets back to the "unknown observation error = poor B estimation" issue

Unknown obs variance | Known obs variance

b est

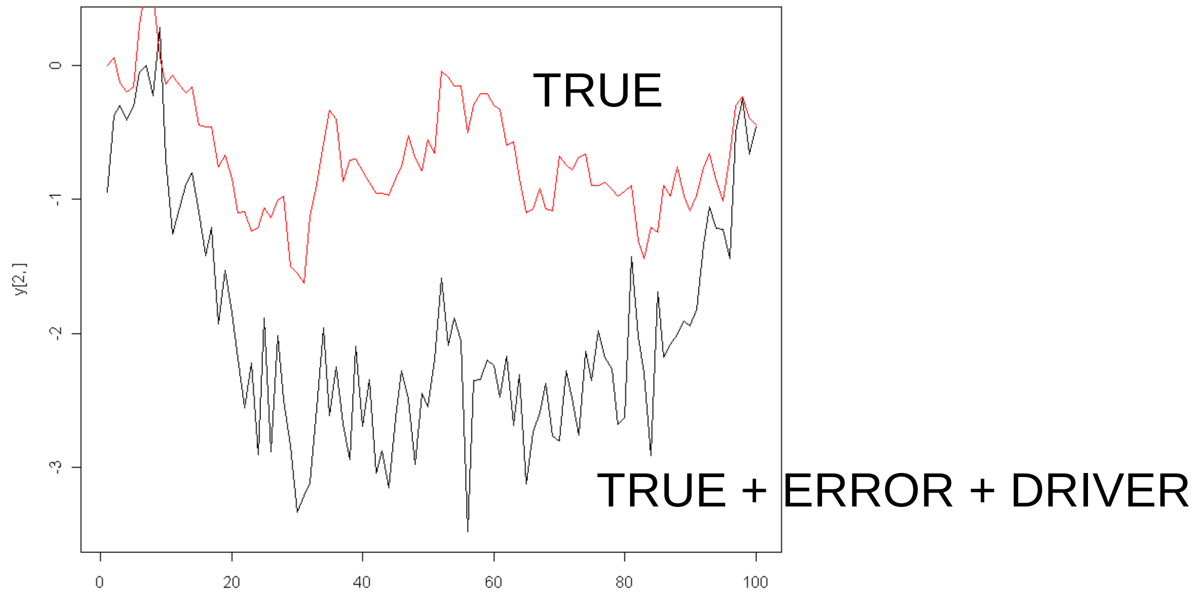


b true | b true

Univariate case (one spp)
- Largely solved by independent samples of same population
- Partially solved with duplicate samples of different populations with same parameters

Multivariate case (community)
- ????
- This is where our current research is focused
- Research depends on simulations (10,000s), so fast algorithms key

Also there is a less recognized issue: the effect of unknown environmental drivers



Univariate case (one spp)

- This is bad unless you can demean your data without removing the true fluctuations.
- If you remove those in your demeaning step, B→0 (spurious DD)
- Datasets much longer than any cycles in the unknown covariate are key.

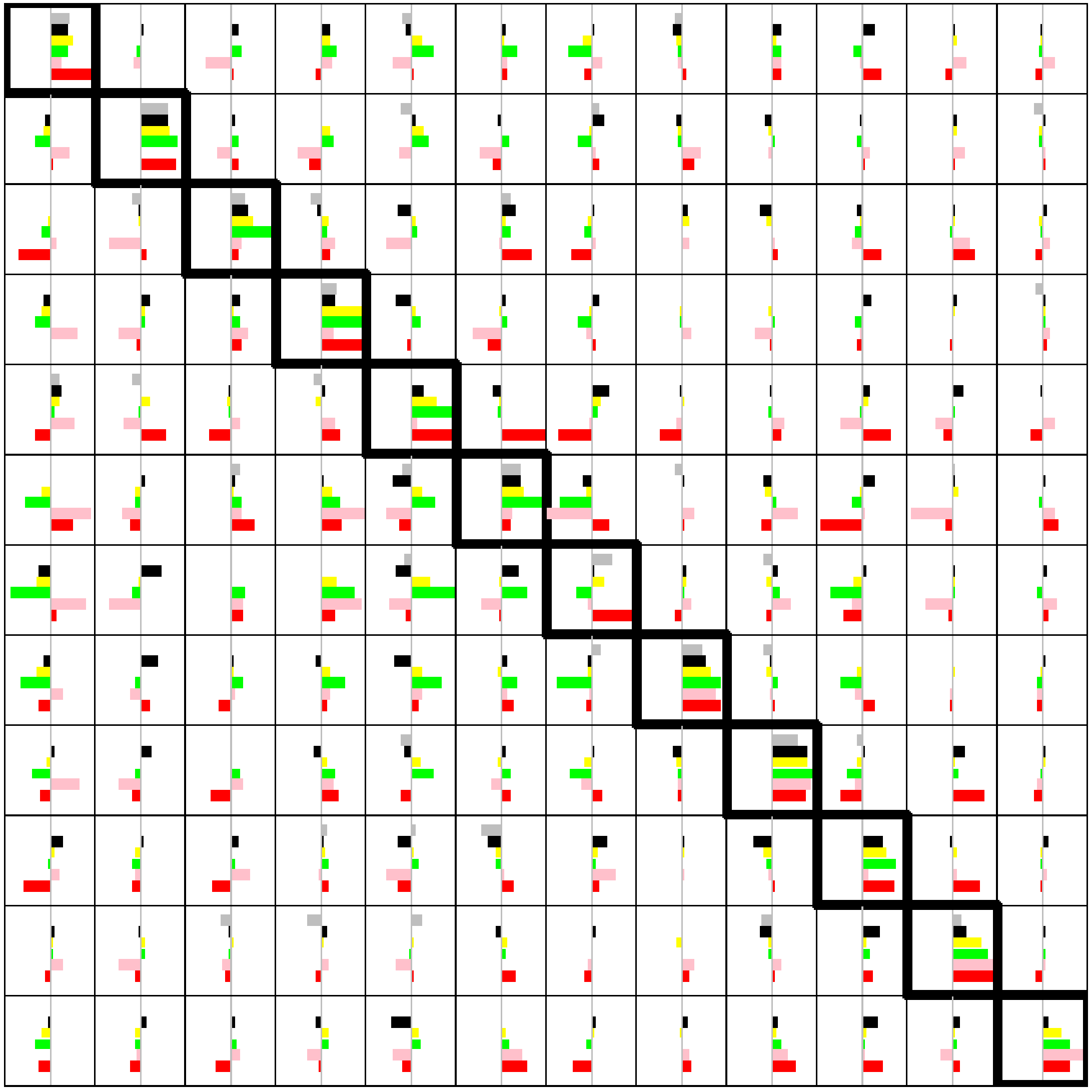
Multivariate case (community)

- It is generally accepted that inclusion of the important environmental drivers is key for good B estimation

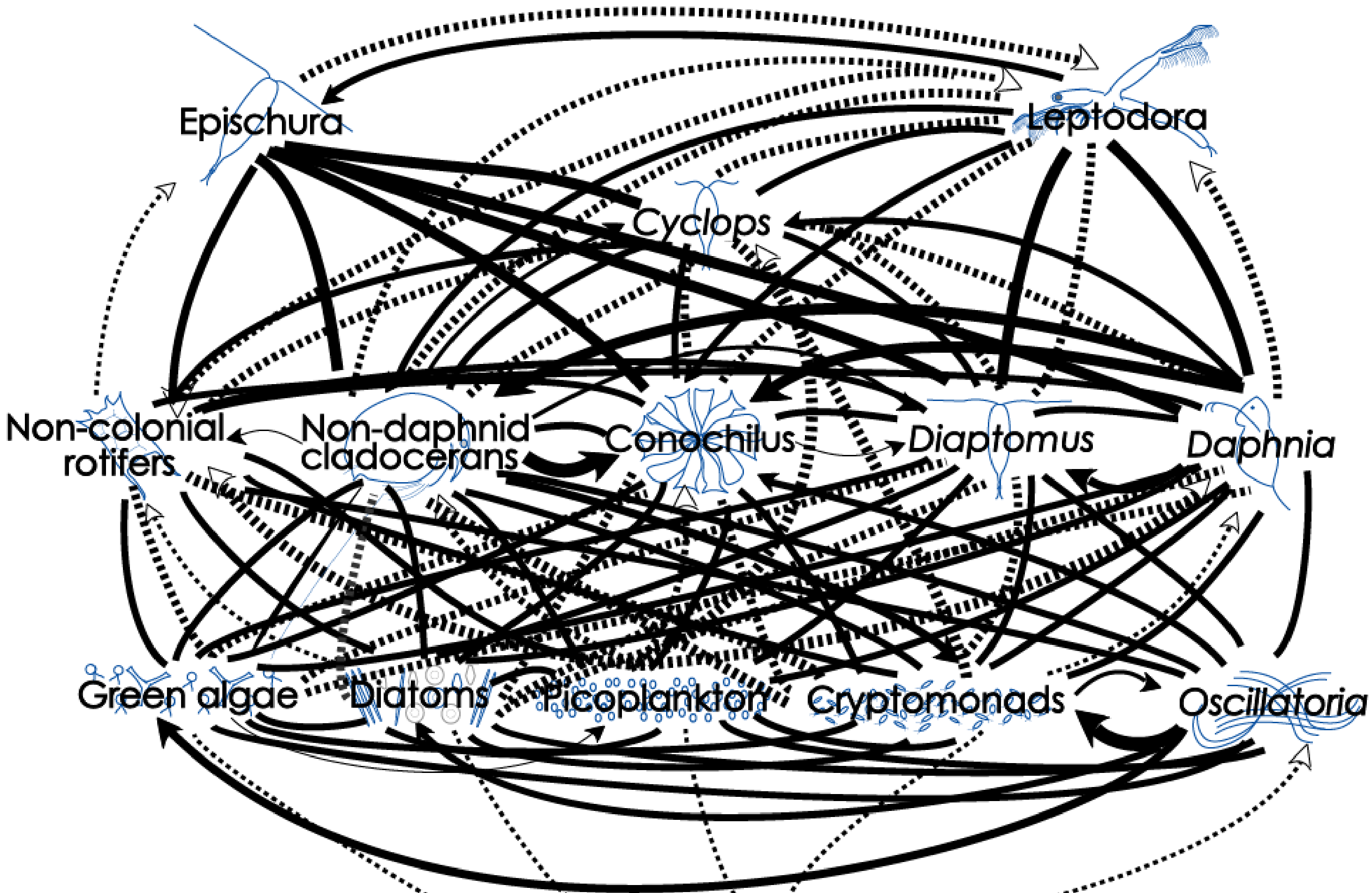
Finally, there are way too many estimated B elements

Constraining B will vastly improve estimation

But current model selection algorithms (for MAR) require searching a huge model space and the fitting step for MARSS is too slow, i.e. model selection steps would = months of computation



Still lots to do.....



Temperature ... Nutrients .... Photoperiod ... Storm activity ... Fishing pressure ...