An overview of stability measures derived from MAR(1) estimates

Craig Jackson

September 28, 2016

1 Basic MAR(1) Estimates

Here we make the following conventions:

1. p is the number of species in the ecosystem we are studying
2. q is the number of tanks or replicates of the ecosystem
3. r is the number of covariates (nutrient, light, etc.)
4. The state vectors $\vec{X}_t$ for the system will generically be written as row vectors. That is:
$$\vec{X}_t = (x_{t,1}, \ldots, x_{t,p})$$
where $x_{t,j}$ is the log of the abundance (biomass) of species j at time t. This is consistent with how our datasets are structured. We do the same for covariates and the environmental white noise term: $\vec{U}_t = (u_{t,1}, \ldots, u_{t,r})$ and $\vec{\varepsilon}_t = (\varepsilon_{t,1}, \ldots, \varepsilon_{t,p})$.
5. The total length of our time series is $N + 1$. That is, there are initial values $\vec{X}_0$ and $\vec{U}_0$ at time $t = 0$ and N subsequent measurements.

Later we will consider multiple tanks, but for now we will limit our discussion to a single tank. With this in mind, we structure our population dataset into two $N \times p$ matrices:

$$X = \begin{bmatrix} x_{0,1} & x_{0,2} & \cdots & x_{0,p} \\ x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ \vdots & & & \\ x_{t,1} & x_{t,2} & \cdots & x_{t,p} \\ \vdots & & & \\ x_{N-1,1} & x_{N-1,2} & \cdots & x_{N-1,p} \end{bmatrix} \qquad Y = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & & & \\ x_{t+1,1} & x_{t+1,2} & \cdots & x_{t+1,p} \\ \vdots & & & \\ x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{bmatrix}$$

That is, X contains all population data except the last data point at time $t = N$, while Y contains all population data except the first data point at time $t = 0$. Hence, X and Y are related by

$$Y_t = X_{t+1}$$

Our covariate dataset takes the form of an $N \times r$ matrix:

$$U = \begin{bmatrix} u_{0,1} & u_{0,2} & \cdots & u_{0,r} \\ u_{1,1} & u_{1,2} & \cdots & u_{1,r} \\ \vdots & & & \\ u_{t,1} & u_{t,2} & \cdots & u_{t,r} \\ \vdots & & & \\ u_{N-1,1} & u_{N-1,2} & \cdots & u_{N-1,r} \end{bmatrix}$$

Note that we ignore the values of the covariates at the final data point since they are not relevant to any population data in out data set.

The theoretical framework for the MAR(1) process imagines that there are three matrices which describe the stationary distribution of the system. These are as follows:

1. The community matrix: This is a $p \times p$ matrix

$$B = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,p} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{p,1} & b_{p,2} & \cdots & b_{p,p} \end{bmatrix}$$

   with the usual meaning to its entries. Namely, $b_{i,j}$ relates to the direct effect of species j on the abundance of species i in the next generation. This term is 0 if and only if species j has no direct effect on species i from one generation to the next. However, there may be indirect effects through interactions with other species. A further discussion of this is delayed until later.

2. A $p \times 1$ column vector $\vec{A}$:

$$\vec{A} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

   This vector is related to the mean about which the system varies, though it is not equal to the mean as we will see later.

3. A $p \times r$ matrix C:

$$C = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,r} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,r} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p,1} & c_{p,2} & \cdots & c_{p,r} \end{bmatrix}$$

which determines the total effect of the covariates on the species abundances in the next generation.

With all this in mind, our assumption that the ecosystem behaves according to a MAR(1) process results in the following equation:

$$X_{t+1}^T = \vec{A} + BX_t^T + CU_t^T + \varepsilon_t^T \tag{1}$$

which is equivalent to

$$Y_t^T = \vec{A} + BX_t^T + CU_t^T + \varepsilon_t^T \tag{2}$$

We now make the following definitions:

Definition 1 *Let $\mathbb{1}$ be a row vector of length N with all entries equal to 1. We define an $N \times (1 + r + p)$ matrix Z as a block form matrix $Z = [\mathbb{1}^T, U, X]$. Explicitly:*

$$Z = \left[\begin{array}{c|ccc|ccc} 1 & u_{0,1} & \cdots & u_{0,r} & x_{0,1} & \cdots & x_{0,p} \\ 1 & u_{1,1} & \cdots & u_{1,r} & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & u_{N-1,1} & \cdots & u_{N-1,r} & x_{N-1,1} & \cdots & x_{N-1,p} \end{array}\right]$$

Definition 2 *We define a $p \times (1 + r + p)$ matrix D as a block form matrix $D = [\vec{A}, C, B]$. Explicitly:*

$$D = \left[\begin{array}{c|ccc|ccc} a_1 & c_{1,1} & \cdots & c_{1,r} & b_{1,1} & \cdots & b_{1,p} \\ a_2 & c_{2,1} & \cdots & c_{2,r} & b_{2,1} & \cdots & b_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_p & c_{p,1} & \cdots & c_{p,r} & b_{p,1} & \cdots & b_{p,p} \end{array}\right]$$

With these definitions and ignoring the effect of environmental noise, equation (2) is equivalent to the following matrix equation:

$$Y^T = DZ^T \tag{3}$$

In fact, (2) is just a column-specific version of (3).

Since we view Z as a matrix of known data and D as a collection of all the unknown parameters we are seeking to estimate, we transpose equation (3) to obtain an equivalent equation that is in standard form:

$$ZD^T = Y \tag{4}$$

We can then used standard linear least squares to find the matrix D that does the best job of estimating the terms $\vec{A}$, B, and C that define the stationary distribution:

$$D^T = (Z^T Z)^{-1} Z^T Y$$

which is equivalent to

$$D = Y^T Z (Z^T Z)^{-1}$$

Questions:

Q: Starting from standard linear least squares, explain how the expression $D = Y^T Z(Z^T Z)^{-1}$ gives the best/optimal solution to the equation $ZD^T = Y$.

Q: How would you modify these methods to find the optimal solution to $ZD^T = Y$ when one or more entires in the community matrix is constrained to be 0?

2 Stability Measures

Our principal interest here is to investigate various measures of ecosystem stability and how they relate to each other. In general, we can divide these various measures of stability into two classes:

a) Those derived from the community matrix B, which we will call "theoretical" measures of stability

b) Those derived from the data, which we will call "empirical" measures of stability

This differentiation between "theoretical" and "empirical" is rather arbitrary and ambiguous in a couple of ways. First, the community matrix itself is derived from the data so there is a sense in which it is "empirical." On the other hand, all measures of stability are synthetic and involve some measure of statistical or mathematical compression, hence there is a sense in which every measure of stability is "theoretical." In any case, we will forge on in spite of the limitations of these terms.

It is important to understand that in MAR(1) models the environmental perturbations (covariates and noise term) are the sole source of variability. That is, there are no internal predator-prey-like dynamics (stable limit cycles) and absolutely no internal chaotic behavior of any kind. Without a source of noise all MAR(1) model datasets would converge to the mean in a predictable way (since they are just linear models after all). What is of interest, however,

is the degree to which a given MAR(1) model amplifies the environmental perturbations (the variance of the model variables is always at least as much as the noise term) and this is what theoretical notions of stability are trying to measure.

Theoretical Measures of Stability (see Ives 2003 for 1-3 and Arnoldi 2016 for 4-7):

1) **The dominant eigenvalue of the community matrix**, λ_0. Larger means less stable. In order for the model to be stable we need all eigenvalues of B to be less than 1 in absolute value. The largest of these eigenvalues is the so-called dominant eigenvalue also called the slow eigenvalue or the eigenvalue for the least stable mode. What this means is that λ_0 corresponds to one or more eigenvectors along which the system changes the slowest. That is, if the system is started from any initial condition without noise, then the state will quickly approach the least stable mode (meaning the λ_0-eigenspace) after which it will move to the mean along this least stable mode.

2) **The** pth **root of the determinant of** B, $\sqrt[p]{\det B}$. Larger means less stable. Recall that the determinant of B is the product of all eigenvalues of B, hence this measure takes into account the rate by which the system changes in the direction of all eigenvectors. The pth root is to allow comparison between systems with different numbers of species. There is also an argument that this expression measures the fraction of the total variation of the system that is due to the system itself (i.e., amplification of noise), though I am not convinced this is a very good argument.

3) **Relative variance along an eigenvalue**, $\psi^2/(1-\lambda^2)$ (e.g., the dominant eigenvalue). Larger means less stable. Here ψ is the variation of the noise term along the eigenvalue. This measure depends on knowing the distribution of the environmental noise. It would make sense to me to normalize so that $\psi = 1$.

Below are four more measures of stability described in Arnoldi 2016. They are formulated for continuous systems, however. One goal for this summer is to formulate them in the context of a discrete MAR(1) model.

4) **Asymptotic resilience**, R_∞. Slowest asymptotic rate of return to equilibrium.

5) **Initial resilience**, R_0. Slowest initial (instantaneous) rate of return to equilibrium.

6) **Intrinsic stochastic invariability**, I_S. From Arnoldi, "constructed from the stationary response of ecosystems to stochastic perturbations of zero-mean and persisting through time. A linear system that is perturbed by a white-noise signal eventually exhibits Gaussian fluctuations (Arnold, 1974). The larger the variance of the stationary response,the less stable the system. We use the inverse of this variance to define stochastic invariability."

7) **Intrinsic deterministic invariability**, I_D. Similar to stochastic invariability but "constructed from the stationary response of ecosystems to zero-mean periodic perturbations that persist through time."

Empirical Measures of Stability

1) **Species-specific coefficient of variation** (CV). If you consider the population X_j of species j to be a random variable, then the CV is just the standard deviation of X_j divided by the mean of X_j:

$$CV = \frac{\sigma(X_j)}{\mu(X_j)}$$

Of course, these terms are never known precisely and must be estimated. Given a sample (timeseries data) $X_{0,j}, X_{1,j}, \ldots, X_{N,j}$, the terms in the equation for the CV is interpreted as the ratio of the sample standard deviation

$$\sigma^2 = \frac{1}{N}\sum_{t=0}^{N}(X_{t,j} - \mu)^2$$

to the sample mean

$$\mu = \frac{1}{N+1}\sum_{t=0}^{N} X_{t,j}$$

The CV is a dimensionless measure of variation that is often called the relative standard deviation. Typically it is computed directly from biomass data without log transformation.

2) **Community coefficient of variation** (CCV). Here we take the total biomass $X = \sum_j X_j$ as the random variable and compute its CV.

Other data-derived measures of variation are possible. For example, one can compute the **covariance matrix** for the dataset and look at things like variation along principle components. The main limitation of all data-derived (i.e., "empirical") measures of stability is that they depend as much on the environmental noise driving the system as the internal dynamics of the system itself.

Questions:

Q) How are the CVs for population data and log transformed data related?

Q) Show that individual CVs are invariant under scaling, but CCV need not be.

Q)* How are R_0, R_∞, I_S, and I_D defined in the context of a discrete MAR(1) model?

Q)* Arnoldi shows that $R_0 \leq I_S \leq I_D \leq R_\infty$. Is this still true in the discrete case?

Q)* What are some potentially interesting measures of stability (variability) that can be derived from the data covariance matrix?

Q)* Are there any consistent relations between the measures from Q3 and any theoretical notions of stability?

Q)* How does everything apply to Dr. Downing's tank data and what does it tell us about these ecosystems?

3 Normalized Data

The theoretical mean for the system can be derived from equation (1) by taking the expected value as $t \to \infty$:

$$\overline{X}^T = \vec{A} + B\overline{X}^T + C\overline{U}^T \tag{5}$$

which implies that $\overline{X}^T = (I - B)^{-1}(\vec{A} + \overline{U}^T)$. Notice that the noise term disappears from the equation since we assume it to have mean $\vec{0}$.

A standard thing is to express data in terms of a mean and a deviation from the mean (anomaly). With this in mind we write $X_t = X'_t + \overline{X}$ and $U_t = U'_t + \overline{U}$. In this case equation (5) implies that equation (2) will reduce to

$$(Y'_t)^T = B(X'_t)^T + C(U'_t)^T + \varepsilon_t^T$$

which, after ignoring noise, is equivalent to $Z'(D')^T = Y'$ where Z' and D' are obtained from Z and D by deleting the first column.

What this means is that B and C can be derived just as well from normalized data as from non-normalized data. It is interesting to think about the implications of this fact: namely, there is no theoretical reason to assume there will be any relation between the community matrix B and the community coefficient of variation (CCV). In particular, while the normalization of log transformed data produces data (species specific anomalies) with arithmetic mean zero, this corresponds on the biomass data side to normalization by scaling to achieve a geometric mean of 1 for each species. While scaling obviously does not affect the CV of individual species, it can have a huge affect on the CCV.

Questions:

Q)* What other measures of community-wide variation exist on the biomass side which are invariant under scaling of individual species biomass and what is their relation to the community matrix B?

Q)* Alternatively, what measures of community wide variation exist on the log transformed side that might have some relation to CCV? Our argument here implies that these measures will have to incorporate $\vec{A}$ as well as B.

Q) A related mathematical problem: show that all translations of log transformed data, not just the normalization discussed above, have zero effect on the estimated community matrix. Simulations bear this out, but a rigorous proof would be better.

4 Description of the Multitank Composite Estimate

Suppose now we have datasets from multiple replicates (e.g., tanks) $X^1, \ldots, X^q$ which we denote using superscripts. That is, we have data for q ecosystems which, while each consisting of the same species, are responding to possibly different environmental covariates and different

environmental noise. The goal, as before, is to find an estimate for the matrices $\vec{A}$, B, and C which describe the overall ecosystem under the assumption that it follows a MAR(1) model. One approach would be to compute estimates for each tank individually and then use these individual estimates to get an overall estimate for the ecosystem, say by taking an average.

However, another approach would be to view each data point as a different measurement of the same ecosystem. In this case we need only compute one estimate on the "dataset" obtained by stacking the X, Y, and U matrices, respectively, on top of each other (so-called vectorizing them):

$$X = \text{vec}(X^1, \ldots, X^q) \qquad Y = \text{vec}(Y^1, \ldots, Y^q) \qquad U = \text{vec}(U^1, \ldots, U^q)$$

Then we can use equation (2) on these vectorized matrices as usual to compute a single composite estimate.

Notice that what we are NOT doing is just taking all our original datasets and combining them into one super dataset. Doing this would lead to errors at the transitions between the tanks since it would imply that the initial state of tank $k+1$ depends on the final state of tank k.

The obvious question to ask at this point is: which method is better, averaging estimates across multiple tanks or the "single ecosystem" composite? We have already shown that the second method is far better in almost every way. We did this by simulating data from known community matrices and comparing the estimates from each method to see how well they reproduced the known B matrix (or more precisely, the stability estimates from the known B matrix).

Questions:

Q) Using simulations, show the ways in which the multi-tank composite estimate described above gives a better estimate for the community matrix than averaging individual tank estimates.

Q) What do you get when you DO combine all the original datasets into one super dataset? I.e., what is the effect of the errors described above on the estimated B matrix?

5 Some Discussion on Species with No Direct Interaction

Consider the following example: a p species ecosystem with no covariates and with mean $\overline{X} = \vec{0}$ (which can be imposed via normalization). In this case we have

$$X_{t+1}^T = BX_t^T + \varepsilon_t^T$$

so that we can write

$$X_{t+2}^T = B(BX_t^T + \varepsilon_t^T) + \varepsilon_{t+1}^T = B^2X_t^T + (B\varepsilon_t^T + \varepsilon_{t+1}^T)$$

and in general, for any k, we can write

$$X_{t+k}^T = B^kX_t^T + \delta_{k,t}^T$$

where

$$\delta_{k,t}^T = B^{k-1}\varepsilon_t^T + \cdots + B\varepsilon_{t+k-2}^T + \varepsilon_{t+k-1}^T$$

Hence, the MAR(1) process with community matrix B described above gives, for any k, another MAR(1) process with community matrix B^k (it can be shown that if B is a stable community matrix, then so is B^k and, in addition, the noise term δ_k is normally distributed with mean $\vec{0}$). Timeseries for these new processes are obtained from the original ($k = 1$) data set $(X_0, X_1, X_2, \dots)$ just by throwing away intermediate terms $(X_0, X_k, X_{2k}, \dots)$

What this means is that we may start with a MAR(1) description of a system in which species i and j, say, have no direct interaction (meaning the (i, j)th entry in B is zero), but when we consider the same system on a different time scale (by taking $k > 1$), then the community matrix in the new timescale could certainly have a non-zero entry in the (i, j)th place. For example:

$$B = \begin{bmatrix} 0.15 & 0 & 0.15 \\ -0.17 & 0.09 & 0.04 \\ 0 & 0.1 & 0.8 \end{bmatrix} \qquad B^2 = \begin{bmatrix} 0.0225 & 0.015 & 0.1425 \\ -0.0408 & 0.0121 & 0.0101 \\ -0.017 & 0.089 & 0.644 \end{bmatrix}$$

That is to say, just because two species should have zero direct interaction biologically speaking from one generation to the next, doesn't mean that the corresponding entry in the estimated community matrix must be zero. As we have seen, this value is sensitive to the model timescale and, on the empirical side, even if there is a "correct" timescale, it is unlikely that we would know what it is. Hence, I think we can say that constrained least squares estimates should be used only for reasons of model fit, not for biological reasons having to do with known or supposed species interaction strengths.

Questions:

Q) How does this timescale consideration work in reverse, from coarser to finer? That is, given a community matrix B and a noise term ε producing a timeseries $X_0, X_1, X_2, \dots$, can you describe (for any given k) a MAR(1) process (i.e., a community matrix and noise term) that results in a timeseries $X'_0, X'_1, X'_2, \dots$ where $X_t = X'_{kt}$ for all t?