Smoothed factor analysis for multivariate time series

Eric J. Ward and …

Conservation Biology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Blvd E, Seattle WA, 98112, USA

## Abstract

## Introduction

Ecological data is characterized by multiple sources of variability, including stochastic natural variation, and errors associated with data collection (observation, sampling, and measurement errors). Disentangling these sources of variability is often challenging, and necessitates the use of complex statistical methods, including state space models. Such approaches have become ubiquitous in ecology, particularly for time series data (Auger-Méthé et al. 2020) – in part because these models allow researchers to make inferences about ecological processes that aren’t directly observable via corrupted observations. Applications of these models include estimating population change over time (Clark and Bjørnstad 2004), understanding movement dynamics (Patterson et al. 2008), and understanding spatiotemporal variation (Anderson and Ward 2019).

Estimating multiple sources of variation in state space models is numerically complex, and can be constrained explicitly or implicitly in ecological models via model assumptions. For example, discrete time state-space models of population trajectories generally assume latent population size can be approximated by an autoregressive process in log-space , where and are normally distributed process deviations representing stochastic variability of the natural system (Dennis et al. 2006). The autoregressive assumption is critical here; without such a constraint, the variance of the stochastic noise is not estimable in the presence of an observation or data model. If model inference is not dependent on parameters of ecological interest (e.g. growth rates, density dependence), there is a wide range of alternative semi-parametric approaches that can be used to model the trajectory of x\_t, including generalized additive models (GAMs, (Wood 2011)) and Gaussian process models (Roberts et al. 2013). Because these models are not autoregressive with a constant time step, the ‘wiggliness’ of the model can be adjusted as part of the model fitting. In addition to their flexibility, these alternative models of may be better suited for situations when data are patchily distributed in time or unequally spaced (making estimation of process and observation errors more difficult).

Challenges posed by univariate time series models also apply to multivariate time series models, though an additional complexity in the multivariate setting is that the number of latent time series may be variable, k=1,…,m, where m is the number of time series observed. At one extreme, k=m, and each time series can be thought of as corresponding to a unique latent state. Motivating questions for these kinds of data include estimating correlated latent processes or trends, or estimating effects of environmental covariates (Hovel et al. 2016). At the other extreme, k=1, where each time series can be thought of representing multiple measurements of the same trajectory of states, or trend, with optional offsets or coefficients included for each time series (offsets allowing for differing detectability). Applications focused on estimating a single trend from multivariate data include the development of ecological indicators. Models with intermediate numbers of latent states 1<k<m require mapping of time series to latent trends. These may be specified a priori (Ward et al. 2010) or estimated within the modeling framework using dimension reduction techniques.

Many statistical approaches have been proposed in recent years for clustering or estimating common signals in multivariate time series (Liao 2005). Examples include clustering based on similarities among time series features (Sardá-Espinosa 2019), identifying common patterns in the frequency domain (Holan and Ravishanker 2018), and clustering based on neural networks (Cherif et al. 2011). Application of these methods to ecological data has been limited, however, in part because many of these approaches identify clusters from raw data and don’t explicitly account for observation error. An alternative approach that has been used in ecology to map collections of multivariate time series to latent states, while accounting for observation error, is dynamic factor analysis (DFA) (Zuur et al. 2003b, 2003a). DFA is an extension of factor analysis for time series data, and estimates a small number of common trends that can describe observed data. Mapping time series to trends is done via an estimated matrix of factor loadings – these allow each time series to be modeled as a mixture of the estimated latent trends, rather than assigning each time series to a single trend.

The objective of this analysis is to introduce a new class of DFA models for multivariate time series. Just as the univariate autoregressive model described above can be approximated with smooth functions, DFA models may be extended to use smooth functions in lieu of autoregressive processes. Recent work has highlighted the application of hierarchical GAMs for multiple data sources (Pedersen et al. 2019). These approaches are flexible and likely to provide similar inference to DFA for single latent trend, however these methods have not been extended to include more than one process. We illustrate two options for modeling smooth functions for latent trends: basis splines (‘b-splines’) and Gaussian process models. Both approaches are compared to conventional autoregressive DFA models for two datasets on marine fishes from the west coast of the USA. All data and code are for replicating our analysis are on Github, and in our existing R package ‘bayesdfa’ (Ward et al. 2019).

## Methods

## Acknowledgments

## Author Contribution Statement

# References

Anderson, S. C., and E. J. Ward. 2019. Black swans in space: Modeling spatiotemporal processes with extremes. Ecology 100:e02403.

Auger-Méthé, M., K. Newman, D. Cole, F. Empacher, R. Gryba, A. A. King, V. Leos-Barajas, J. M. Flemming, A. Nielsen, G. Petris, and L. Thomas. 2020. A guide to state-space modeling of ecological time series. arXiv:2002.02001 [q-bio, stat].

Clark, J. S., and O. N. Bjørnstad. 2004. Population Time Series: Process Variability, Observation Errors, Missing Values, Lags, and Hidden States. Ecology 85:3140–3150.

Dennis, B., J. M. Ponciano, S. R. Lele, M. L. Taper, and D. F. Staples. 2006. Estimating Density Dependence, Process Noise, and Observation Error. Ecological Monographs 76:323–341.

Patterson, T. A., L. Thomas, C. Wilcox, O. Ovaskainen, and J. Matthiopoulos. 2008. State–space models of individual animal movement. Trends in Ecology & Evolution 23:87–94.

Roberts, S., M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain. 2013. Gaussian processes for time-series modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371:20110550.

Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73:3–36.