BIOS 735 Project Proposal

Abby Foes, Zheng Lan, Justin Landis, Yu Liu, Alec Reinhardt

March 2025

1 Introduction

Data Description

We identified a dataset for pasta sales from an Italian grocery store from January 2014 to December 2018. The dataset contains 116 time series with 1798 time points each, including the quantity of sales for specific pasta items and whether or not a promotion was placed on each item. The time series are hierarchical, with 4 brands and up to 45 unique items per brand (see Table 1). Based on some preliminary data analysis, there is evidence of common temporal patterns in items sales within each brand, as well relationships between promotions and sales (see Figure 2).

Brand	n
B1	42
B2	45
B3	21
B4	10

Figure 1: Number of items within each pasta brand

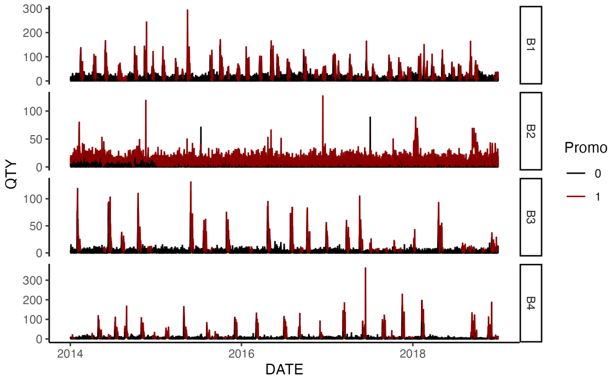

Figure 2: Time series of pasta sales per brand and item. Red lines indicate dates in which that brand's item was promoted

Research Questions

We will consider the following questions:

- Can we improve forecasting accuracy by accounting for the hierarchical structure in the data, compared with modeling each brand and item independently? What about models that account for sparsity (for items with many days of no sales) or periodicity?
- How do temporal trends, and promotion-sales relationships, vary across pasta brands? Which points in time show the largest deviations? Is there evidence of periodic (seasonal) trends, and how do these relate across brands and items?

2 Methods

Notation

We have $i = 1, 2, \ldots, n$ items, $h = 1, \ldots, m$ brands, and $t = 1, 2, \ldots, T$ time points. Let $y_{hi,t}$ be the number of sales of item i in brand h at time t and $x_{hi,t}$ the (binary) promotion.

Models

To develop a forecasting model, we will adapt various classical approaches for time series modeling. For instance, the autoregressive model (AR) (see, e.g. Box et al. (2015)) assumes that the current outcome depends linearly on the previously observed outcomes (up to a certain lag q), covariates, and a noise term. This can be expressed as

$$y_t = \alpha_0 + \sum_{l=1}^{q} \beta_l y_{t-l} + \gamma x_t + \epsilon_t \tag{1}$$

where y_t is the outcome at time point t, α_0 is the population intercept, the β_l terms capture the dependence between y_t and previous time points, γ captures the effect of (time-varying) covariate x_t on y_t, and ϵ_t is the noise term, assumed to be mean zero Gaussian. As a first step, we plan to fit these models for each brand and item independently. The R package *astsa* utilizes model fitting methods such as Burg's Algorithm and Yule-Walker equations. However, we plan to also implement a MLE-based approach. Chapter 2 of the book (Funatogawa et al. (2018)) details the likelihood and steps for MLE for AR models.

In order to capture shared patterns among items within each brand, we propose a mixed effects AR model (see Nicholls and Quinn (2012) for example) of the form

$$y_{hi,t} = (\alpha_0 + a_h) + \sum_{l=1}^{q} (\beta_l + b_{h,l}) y_{hi,t-l} + (\gamma + g_h) x_{hi,t} + \epsilon_t \tag{2}$$
$$\alpha_h \sim N(0, \sigma^2_\alpha),\ b_{h,l} \sim N(0, \sigma^2_\beta),\ g_h \sim N(0, \sigma^2_\gamma),\ \epsilon_t \sim N(0, \sigma^2_\epsilon)$$

where we introduce random intercept a_h, and random slopes $b_{h,l}$ and g_h for the effects of previous timepoints and covariates, respectively. We assume that each term comes from a common distribution specific to each brand, in effect pooling information across items within each brand while still allowing for deviations. We will also consider simplifications and possible extensions of the above model, including modeling cross-brand dependencies. For model fitting, we will consider a Bayesian variation of Model 2, with suitable priors for model coefficients and residual noise. We will implement a Markov Chain Monte Carlo (MCMC) algorithm to sample the posterior of model parameters, which in turn allows for both prediction and uncertainty quantification (e.g. credible intervals).

Other competing approaches we plan to try include extensions of AR, such as ARMA and SARIMA (which can account for periodic trends), and alternatives like Hidden Markov Models.

Moving Average Model

$$Y_{i,t} = \alpha + \epsilon_{i,t} + \sum_{j=1}^{q} \theta_i \epsilon_{i,t-j}$$

ARMA Model

$$y_t = \alpha + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \epsilon_t + \epsilon_{t-1}\theta_q + \ldots + \epsilon_{t-q}\theta_q$$

Analysis Plan

We plan to fit the AR model described above and its variations using both a MLE-approach (using EM or NR algorithm) and with MCMC. For each model we fit, we will consider both non-hierarchical models, fit to each item independent, and a hierarchical variation which includes brand-specific model parameters (i.e. mixed effects model). Lastly, we will fit a machine learning approach such as Random Forest or Hidden Markov Models with off-the-shelf packages.

To evaluate predictive performance, we will implement a holdout prediction scheme. The first K timepoints will be used for model training, and we will assess prediction (with RMSE, e.g.) on the remaining timepoints. We will vary the size of K to see how fast prediction improves as the amount of training data increases for the various models.

To address our set of questions, we will explore and interpret the different coefficient estimates for various models, after fitting to the entire dataset. In particular, the MCMC approach allows naturally quantify uncertain for both model terms and the predicted sales outcome, allowing to track where in time the different brands have evidence of significant differences. We will assess periodic trends with exploratory data analysis and utilizes specific AR-based models such as SARIMA.

References

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Funatogawa, I., Funatogawa, T., et al. (2018). *Longitudinal data analysis: autoregressive linear mixed effects models*. Springer.

Nicholls, D. F. and Quinn, B. G. (2012). *Random Coefficient Autoregressive Models: An Introduction: An Introduction*, volume 11. Springer Science & Business Media.