Modeling Promotional Pasta Sales with Hierarchical Poisson Autoregression

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

Contents

1 Introduction

1.1 Background

1.2 Data Description

We have identified a dataset of pasta sales from an Italian grocery store with observations made between January 2014 and December 2018. This dataset contains the quantity of sales for 118 unique items across 1,798 equally-spaced time points (days). Along with this quantity of sales, we are provided with a brand identifier and a binary label of whether a given item was promoted on that day as well. This time series data is thus hierarchical, with four brands and up to 45 unique items in each brand.

```
data("data_set_tidy")
glimpse(data_set_tidy)
```

```
## Rows: 212,164
## Columns: 5
## $ DATE  <date> 2014-01-02, 2014-01-02, 2014-01-02, 2014-01-02, 2014-01-02, 201~
## $ brand <chr> "B1", "B1", "B1", "B1", "B1", "B1", "B1", "B1", "B1", "B1", "B1"~
## $ item  <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12", "~
## $ QTY   <dbl> 7, 3, 0, 2, 3, 1, 0, 4, 0, 0, 0, 0, 7, 2, 5, 2, 0, 7, 1, 0, 0, 4~
## $ PROMO <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

Each row corresponds to daily sales for a specific item, along with brand and promotion indicators.

1.3 Research Questions

The goal of this report is to investigate the following:

- How do temporal trends and promotion-sales relationships vary across the four pasta brands?
- Can we improve forecasting of sales by accounting for the hierarchical structure in the data compared to simply modeling each brand (and item) independently?

We compare likelihood-based models (EM, Newton-Raphson), Bayesian inference (MCMC), and machine learning benchmarks (Random Forests, Hidden Markov Models), using out-of-sample mean squared error (MSE) as the main evaluation metric.

1.4 Exploratory Data Analysis

2 Methods

2.1 Motivation for autoregressive modeling

To address our research goals, we consider several variations of time series models. The key commonality behind these model is that they all use information from the previous history of sales as a predictor of the current sales for a given item. These types of models are known as *autoregressive (AR) models*, which have been widely-used across disciplines for time series forecasting [Hamilton, 2020]. Additionally, our models all include the current promotion status for a given item as a predictive of the current sales for that item.

One issue that arises from this dataset is that the sales outcomes are discrete counts. Classical AR-based approaches, such as the Autoregressive Moving Average (ARMA), make normality assumptions about the current outcomes given past history. However, given the discrete data, a more appropriate assumption in our case is that the sales are distributed according to a Poisson distribution. We can then incorporate the autoregressive structure into the Poisson mean parameter, leading to a Poisson autoregressive model (PAR), formulated below. To deal with the inherent correlations across items within each brand, we also formulate a variation of this PAR model that can jointly model sales across all items, after accounting for similarities by brand.

2.2 Poisson Autoregression for single item modeling

We first consider modeling each item independently. For a given item, denote the sales at timepoint t by y_t, where $t = 1, \ldots, T$. Also let $\mathbf{x}_t$ be a control vector including promotion indicator at time t (and a 1 element for an intercept).

Based on the work by Brandt and Williams [2001], we formulate the single-item Poisson autoregression model (PAR) as

$$y_t \sim \text{Poisson}(m_t)$$

$$m_t = \sum_{l=1}^{q} \beta_l y_{t-l} + \left(1 - \sum_{l=1}^{q} \beta_l\right) \cdot \exp(\mathbf{x}_t' \boldsymbol{\gamma})$$

where m_t is the mean Poisson parameter at time t, β_l is the autoregressive coefficient for lag l ($l = 1, \ldots, q$), and $\boldsymbol{\gamma}$ represents the effect of current covariates (including promotion) on sales. The time-varying Poisson mean can be interpreted as a weighted combination of previous observations of sales and the covariates, where the exponential term comes from the log link used for Poisson regression. We impose the constraints $\beta_l \geq 0$ and $\sum_{l=1}^{q} \beta_l < 1$ to ensure that the AR process is stationary (i.e. not diverging over time).

2.3 Likelihood Function for PAR

Let $\boldsymbol{\theta}$ denote the full set of model parameters. The full likelihood across all items and times is:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{t=q+1}^{T} \frac{m_t^{y_t} e^{-m_t}}{y_t!}$$

The corresponding log-likelihood, used for optimization and posterior inference, is:

$$\log \mathcal{L}(\boldsymbol{\theta}) = \sum_{t=q+1}^{T} \left[y_t \log(m_t) - m_t - \log(y_t!) \right]$$

Where,

$$m_t = \underbrace{\sum_{l=1}^{q} \beta_l y_{t-l}}_{\text{AR part}} + \underbrace{\left(1 - \sum_{l=1}^{q} \beta_l\right)}_{\text{mixing weight}} \cdot \underbrace{\exp(\mathbf{x}_t' \boldsymbol{\gamma})}_{\text{covariate part}}$$

2.4 Gradient (Score Function) for PAR

Define the following:

- $a_t = \sum_{l=q}^{q} \beta_t y_{t-l}$
- $c_t = \exp(x_t^T \boldsymbol{\gamma})$
- $w = 1 - \sum \beta_t$
- $m_t = a_t + w \cdot c_t$

Then, the derivative of the log-likelihood with respect to γ_j is:

$$\frac{\partial \log \mathcal{L}}{\partial \gamma_j} = \sum_{t=q+l}^{T} \left[\frac{y_t}{m_t} - 1 \right] \cdot w \cdot c_t \cdot x_{tj}$$

And the derivative of the log-likelihood with respect to β_k is:

$$\frac{\partial \log \mathcal{L}}{\partial \beta_k} = \sum_{t=q+l}^{T} \left[\frac{y_t}{m_t} - 1 \right] [y_{t-k} - c_t]$$

2.5 Poisson Vector Autoregression (PVAR) for multi-item modeling

Next, we consider modeling more than one item at a time. Suppose we have n items, T timepoints and B brands. Let $g_i \in \{1, \ldots, B\}$ denote the brand (group) to which item i belongs.

We propose a hierarchical Bayesian extension of the PAR model, which we refer to as the Poisson Vector Autoregression (PVAR) model. The sales y_{it} for item i and timepoint t are modeled similarly to the PAR case as

$$y_{it}|m_{it} \sim \text{Poisson}(m_{it})$$
$$m_{it} = \sum_{l=1}^{q} \beta_{i,l} y_{i,t-l} + \left(1 - \sum_{l=1}^{q} \beta_{i,l}\right) \exp(\mathbf{x}_{it}' \boldsymbol{\gamma}_i)$$

where each item is assumed to have its own autoregressive coefficients $\beta_{i,l}$ and covariate effects $\boldsymbol{\gamma}_i$. We use the following hierarchical priors on model parameters

$$\boldsymbol{\gamma}_i \sim N(\boldsymbol{\mu}_{g_i}, \Sigma_{g_i})$$
$$\boldsymbol{\mu}_{g_i} \sim N(\boldsymbol{\mu}_0, \Sigma_0),\ \Sigma_{g_i} \sim \text{Inv-Wishart}(\nu, \Psi)$$
$$\tilde{\boldsymbol{\beta}}_i|\tau_{g_i} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{g_i}),\ \beta_{i,l} = \tau_{g_i}\tilde{\beta}_{i,l}$$
$$\boldsymbol{\alpha}_{g_i} \sim \text{Dirichlet}(\alpha_0, \ldots, \alpha_0),\ \tau_{g_i} \sim \text{Beta}(a_\tau, b_\tau)$$

where $\boldsymbol{\mu}_0$, Σ_0, ν, Ψ, α_0, a_τ, and b_τ are fixed hyperparameters. Essentially, these priors assume that the item-level for a given brand come from some common distribution. Effectively, this will pool information from within each brand when estimating the effects of previous sales and promotion for each item. We use Dirichlet and Beta distributions for the AR coefficients as a way to impose the constraint that $\sum_{l=1}^{q} \beta_{i,l} < 1$, which leads to stationarity in the Poisson AR processes for each item.

3 Estimation Methods

3.1 EM Algorithm

- E-step: Applicable if we add new random intercepts (either item or brand level)
- M-step: Maximize the expected log-likelihood with respect to model parameters γ, β, and τ

3.2 Bayesian Inference (MCMC)

- Fit the independent Poisson Autoregression (PAR) model for each item with MCMC
- Place priors directly on model parameters
- τ: mixing weight for AR component, $\beta = \tau\tilde{\boldsymbol{\beta}}$

Priors:

- $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma_\gamma)$
- $\Sigma_\gamma \sim \text{Inv-Wishart}(\nu, \Psi)$ — (approximated as fixed in Stan)
- $\tilde{\boldsymbol{\beta}} \sim \text{Dirichlet}(\alpha)$
- $\tau \sim \text{Beta}(a, b)$

4 Model Evaluation and Predictive Performance

We evaluate models using holdout-based prediction: the final 10–20% of each item's time series is reserved for testing.

```
# Example: results <- evaluate_holdout(...)
# knitr::kable(results)
```

** Evaluation Metrics ** - Root Mean Squared Error (RMSE) - Mean Absolute Error (MAE) - Log Predictive Density (for Bayesian models) - Posterior predictive interval coverage (for MCMC)

5 Machine Learning Benchmarks

5.1 Random Forest

```
# rf_fit <- fit_random_forest(...)
```

6 Summary and Discussion

- Which methods give the most accurate predictions for sparse, autocorrelated data?
- What is the trade-off between hierarchical shrinkage and item-specific flexibility?
- Are machine learning models (e.g., random forests) better suited for forecasting than structured generative models?

References

Patrick T Brandt and John T Williams. A linear poisson autoregressive model: The poisson ar (p) model. *Political Analysis*, 9(2):164–184, 2001.

James D Hamilton. *Time series analysis*. Princeton university press, 2020.