Bayesian Dynamic Modeling and Forecasting of Count Time Series

by

Lindsay R. Berry

Department of Statistical Science
Duke University

Date: ____________________________
Approved:

____
Mike West, Supervisor

____
Mine Cetinkaya-Rundel

____
Andrew Cron

____
Jason Xu

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2019

Abstract

Bayesian Dynamic Modeling and Forecasting of Count Time Series

by

Lindsay R. Berry

Department of Statistical Science
Duke University

Date: ____________________

Approved:

____
Mike West, Supervisor

____
Mine Cetinkaya-Rundel

____
Andrew Cron

____
Jason Xu

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2019

Abstract

Problems of forecasting related time series of counts arise in a diverse array of applications such as consumer sales, epidemiology, ecology, law enforcement, and tourism. Characteristics of high-frequency count data including many zeros, high variation, extreme values, and varying means make the application of traditional time series methods inappropriate. In many settings, an additional challenge is producing online, multi-step forecasts for thousands of individual series in an efficient and flexible manner. This dissertation introduces novel classes of models to address efficiency, efficacy and scalability of dynamic models based on the concept of decouple/recouple applied to multiple series that are individually represented via novel univariate state-space models. The novel dynamic count mixture model involves dynamic generalized linear models for binary and conditionally Poisson time series, with dynamic random effects for overdispersion, and the use of dynamic covariates in both binary and non-zero components. New multivariate models then enable information sharing in contexts where data at a more highly aggregated level provide more incisive inference on shared patterns such as trends and seasonality. This novel decouple/recouple strategy incorporates cross-series linkages while insulating parallel estimation of univariate models. We extend these models to a general framework appropriate for settings in which count data arises through a compound process. The motivating application is in consumer sales contexts where variability in high-frequency sales data arises from the compounding effects of the number of transactions and the number of sales-per-transactions. This framework involves adapting the dynamic count mixture model to forecast transactions, coupled with a binary cascade concept using a sequence of Bayesian models to predict the number of units per transaction. The motivation behind the binary cascade is that the appropriate way to model rare events is through a sequence of conditional probabilities of increasingly rare outcomes. Several case studies in many-item, multi-step ahead supermarket sales forecasting demonstrate improved forecasting performance using the proposed models, with discussion of forecast accuracy metrics and the benefits of probabilistic forecast accuracy assessment.

Dedicated to my family.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

$Po(\lambda)$	Poisson distribution with rate parameter λ.
$Ber(\pi)$	Bernoulli distribution with success probability π.
$Ga(\alpha, \beta)$	Gamma distribution with shape parameter α and rate parameter β.
$Be(\alpha, \beta)$	Beta distribution with shape parameters α and β.
$Nb(k, p)$	Negative binomial distribution with number of failures k and success probability p.
$N(\mu, \nu)$	Normal distribution with mean μ and variance ν.
$BBin(n, \alpha, \beta)$	Beta-binomial distribution with number of trials n and shape parameters α and β.
$\mathbb{1}(\cdot)$	Indicator function.
$(\mathbf{m}, \mathbf{C})$	Distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$.
$\mathcal{D}_t$	All data observed up to time t.
$\mathcal{I}_t$	All additional information, beyond observed data, available at time t.
$s{:}t$	Indices $s, s+1, \ldots, t-1, t$.
$\mathrm{E}(X)$	The expected value of random variable X.
$\mathrm{V}(X)$	The variance of random variable X.

Abbreviations

ACP	Autoregressive Conditional Poisson.

ARMA	Autoregressive Moving Average.
c.d.f.	Cumulative Distribution Funciton.
DAG	Directed Acyclic Graph.
DBCM	Dynamic Binary Cascade Model.
DCMM	Dynamic Count Mixture Model.
DGLM	Dynamic Generalized Linear Model.
DLM	Dynamic Linear Model.
FFBS	Forward Filtering Backward Sampling/Smoothing.
GLARMA	Generalized Linear ARMA.
HPD	Highest Posterior/Predictive Density.
INAR	INteger-valued AutoRegressive.
KL	Kullback-Liebler.
MAD	Mean Absolute Deviation.
MCMC	Markov Chain Monte Carlo.
MRPS	Mean Ranked Probability Score.
p.d.f.	Probability Density Function.
PIT	Probabilistic Integral Transform.
p.m.f.	Probability Mass Function.
RMSE	Root Mean Squared Error.
SKU	Stock Keeping Unit.
sMSE	Scaled Mean Squared Error.
UPC	Universal Product Code.

Acknowledgements

First, I would like to express my appreciation to my advisor, Mike West, for being a constant source of knowledge, guidance, and encouragement. As an advisor, Mike was deeply invested not only in my research, but also in my personal development and success. I could not have asked for a better mentor. Many thanks to my defense committee members, Andrew Cron, Jason Xu, and Mine Çetinkaya-Rundel, for their time and support. I would like to further thank Andrew Cron for his valuable contributions as a research collaborator and for generously sharing his insights and advice with me. Thanks also to my prelim committee members, David Banks and Beka Steorts, for their support during my graduate career.

This research was due, in large part, to the contributions and guidance of Paul Helman and his team at 84.51°. Special thanks to Xiaojie Zhou, Natalia Connolly, Hank Vaccaro, Mark Lee, and Cindy Li for their suggestions, help, and kindness. Thanks are also due to 84.51° and the Duke Graduate School for the financial support of this research, and to ISBA and ISBIS for conference travel funding.

My graduate school experience would not have been the same without my classmates and friends in the Statistics department. As I look back on my time at Duke, one thing I am most grateful for is getting to spend each day with my friends in Old Chem 222. I am especially grateful to my four-year desk buddy, Jake, my fellow yogi, Liz, and my self-designated fashion advisor, Abbas.

Finally, I would like to thank all of the friends and family that have supported me unconditionally through every adventure. To my best friend, Oksana, thank you for sticking with me for so many years. To my parents—thank you for being my biggest cheerleaders, role models, and supporters. Your love and encouragement means the world to me. To John—thank you for making coming home to you and Mav the best part of my days. I would not have made it here without you.

1

Introduction

Modeling and forecasting of multivariate time series of non-negative counts are common interests among many companies and research groups. One key area that motivates our work is that of product sales/demand forecasting, exemplified by forecasting sales based on historical data and concomitant information in commercial outlets including supermarkets and e-commerce sites. Forecasts for inventory management, production planning, pricing, and marketing decisions are at the heart of business analytics in such environments. Historically, product sales forecasting has focused on aggregate sales of entire categories/stores or weekly/monthly sales of items due to the lack of relevant data at the daily level for items. However, the increasing availability of rich point-of-sale data on transactions of retail items has created interest in forecasting daily sales of individual products (Boone et al., 2019).

For large retailers this is a high-dimensional problem as forecasts are required for multiple time granularities for many individual products across multiple outlets. Seaman (2018) explains that physical stores can sell hundred of thousands of individual items while online retailers may sell millions. In order to function, large retailers may need to produce up to a billion individual forecasts each day. To be effective in such settings, models must run efficiently in an on-line manner as new data are collected, and do so automatically as a routine while having the ability to flag exceptions and call for intervention as needed. The challenge is to define a flexible class of product-level models that can be customized to individual products within a general framework. Then, forward/sequential analysis and multi-step ahead forecasting must be effective and efficient computationally, and enable integration across potentially many products to share information while maintaining scalability to increasingly large-scale problems.

Our research to address these challenges begins with definition and development of a novel class of univariate models for time series of non-negative counts. Anchored in our case-study context of forecasting daily sales of products at a large supermarket chain, key questions include accounting for various levels of seasonality (weekly, monthly, yearly), holiday effects, price/promotion information, and unpredictable drifts in levels and variability of sales. High-frequency time series like daily sales are often characterized by high variability and extreme values, and levels of demand across products can vary drastically, with some products selling dozens of units per day, and others having many days with zero sales. Time series at the fine-scale resolution of individual item sales typically contain many zeros and low counts, so that traditional time series models and methods– such as exponential smoothing (Hyndman et al., 2008), ARIMA models (Box et al., 2008), and conditionally Gaussian/linear state-space models (West and Harrison, 1997; Prado and West, 2010)– are not appropriate.

Due to the high-dimension, demand forecasters rely on simple, univariate, extrapolative models (Ma et al., 2016; Seaman, 2018). Common examples include the naive forecast of the last observation, moving averages, simple exponential smoothing, ARIMA models, and exponential smoothing state space models. Generally, these models do not incorporate covariates such as price and promotion, and if estimated at all, the covariate effects are static over time. Ali et al. (2009) found that simple methods perform well in periods without promotions, but that more advanced models improve forecasting when promotion features are incorporated. In common usage, these simple models produce only point forecasts and uncertainty is ignored. Exponential smoothing models can produce interval predictions, but these are either based on assumptions of normality inappropriate for low-counts, or empirical approaches like bootstrapping that may not adapt to dynamic changes in the distribution (Hyndman et al., 2005; Taylor, 2007). Zero-inflation and overdispersion are difficult to handle, and are usually addressed on a case-by-case basis through specific alternative models. In this high-dimensional setting, it is infeasible to individually customize models for each item and to monitor the suitability of the chosen model over time. Instead, we desire an automatically flexible model that can handle common characteristics of count data like zeros and overdispersion. While these simple time series models have proven useful in aggregate sales forecasting, they are inappropriate for forecasting low-valued counts.

A number of approaches to forecasting count-valued time series have, of course, been developed. The issues of intermittent demand (many zeros in sales) and low counts have been a main concern (e.g. Croston, 1972), as has over-dispersion relative to Poisson structures. A range of modified Poisson, negative binomial, so-called "hurdle shifted" Poisson and jump-process models have been explored with specific applications (e.g. Chen et al., 2016; Chen and Lee, 2017; Snyder et al., 2012; McCabe and Martin, 2005). Quantile regression and forecasting have been applied for count data settings such as inventory management where safety stock is a major concern (Taylor, 2007; Trapero et al., 2019). An intuitive approach for reducing intermittency of data is temporal aggregation into non-interlapping intervals, e.g. aggregating daily to weekly data, such as ADIDA (Nikolopoulos et al., 2011) and MAPA (Kourentzes et al., 2014). Challenges associated with temporal aggregation include selecting the appropriate level of aggregation and disaggregation scheme, excessive smoothing of the data, and loss of information due to having fewer observations.

Comparative analyses in Yelland (2009) highlight the utility of state-space approaches including the canonical Gamma-Poisson "local level" model (West and Harrison, 1997 section 10.8; Prado and West, 2010 section 4.3.7). This and other standard Bayesian state-space models have proven utility in a range of discrete-time series contexts including dynamic network studies (e.g. Chen et al., 2018) where short-term forecasting and local smoothing are primary goals. With a view towards improved predictive ability and– critically– multi-step forecasting, it is perhaps somewhat surprising that more elaborate and predictive Bayesian state-space models have not yet become central to the area, especially in the context of some of the key genesis developments in Bayesian forecasting in commercial settings (e.g. West and Harrison, 1997, chapter 1, and references therein) and their exploitation over several decades.

Chapter 2 provides important background on Bayesian state-space models and, in particular, standard univariate dynamic generalized linear models (DGLMs: West et al., 1985; West and Harrison, 1997 chapter 15; Prado and West, 2010 section 4.4). We detail standard sequential learning, forecasting, and component discounting in DGLMs, and provide key examples of DGLMs which we will build upon in future chapters. Section 2.5 describes three variational Bayes approaches to conjugate prior specification in DGLMs, and a practical comparison of these methods. Finally, Section 2.6 concludes the chapter with a discussion of common point forecast metrics and probabilistic forecast evaluation for count data. The consumer forecasting field has tended to focus on very specific point-forecast metrics, and part of our work here is to broaden the perspective on forecast evaluation in response to, and enabled by, the availability of fully specific probability forecast distributions.

Chapter 3 defines and develops a new class of dynamic count mixture models (DCMMs), coupling Bayesian state-space models for binary time series with conditional count models. We build on standard univariate DGLMs to define a general and flexible class of basic dynamic models that are customizable to individual series. A critical aspect of DCMMs is that they inherit the sequential learning and forecasting of Bayesian state-space models, allowing fast, parallel processing of decoupled univariate models for individual series. These models allow for the incorporation of series-specific predictors for zero-count prediction as well as for forecasting levels of non-zero counts, and– critically for many applications– dynamic random effects extensions for over-dispersion relative to conditionally Poisson models.

The motivating application in consumer sales forecasting is introduced in Section 3.4. The case study context is supermarket sales of many individual items, and several examples of item-level sales highlight the uses of DCMMs and the flexibility of the new model class to adapt to substantially differing features of count time series. As part of this, we discuss results for a range of metrics for forecast assessment, including standard point-forecast measures, probabilistic calibration and coverage.

Chapter 4 addresses the interest in multivariate cross-series linkages and borrowing of information on shared characteristics and patterns. Our main focus here is on the potential for multivariate models to improve multi-step ahead forecasts at the level of individual series, while maintaining efficiency of the forward/sequential analysis and, critically, enabling scaling to many series. Among multivariate approaches, a number of authors have explored models for counts or proportions (e.g. Quintana and West, 1988; West and Harrison, 1997 chapter 16; Da-Silva and Migon, 2016). However, there do not exist general classes of models addressing our key desiderata of flexibility at the single-series level, analytic tractability and capacity to scale to higher dimensions. Most existing models tend to be specific to applications and not easily amenable to integrating covariate information at the individual series level. Further, many require intense computations such as Markov chain Monte Carlo or sequential particle methods (e.g. Cargnoni et al., 1997; Aktekin et al., 2018), which is antithetical to our concern for fast, sequential analysis and cross-series integration with many series. Our work here builds on the flexible class of univariate DCMMs and defines a novel multi-scale approach to integrating cross-series information about common patterns, exemplified in terms of time-varying seasonality where a seasonal pattern is evident across series but with series-specific random effects. Critically, our new decouple/recouple approach enables information sharing while avoiding intense computations typical of random-effects/hierarchical models. The basic idea is of using aggregate level data to inform on micro-level series is one example of a decouple/recouple strategy that maximally exploits series-specific customization while enabling integration in multivariate models; see Chen et al. (2019, 2018); Gruber and West (2016, 2017) for models that exploit this strategy in very different contexts. Section 4.5 revisits application in the consumer sales case study, illustrating the use and impact of the multi-scale framework across several products in the context of sales forecasting of multiple related items.

Chapter 5 compares the univariate and multi-scale DCMM to other existing models that could be used in the context of product sales forecasting. The comparison models each fall within the framework of observation driven models with an autoregressive dependence component. We present results for common point forecast metrics as well as probabilistic forecast evaluation. Although the performance for these comparison approaches is reasonable for point forecast metrics, the DCMM is preferable in our application due to the dynamic covariate effects, efficient multi-scale framework, and automatic handling of overdispersion and excess zeros.

Chapter 6 defines the new class of dynamic binary cascade models (DBCMs) incorporating a novel concept of binary cascades. This begins with flexible DCMMs to assess and forecast daily item-level transactions. Coupled with this, development of our dynamic binary cascade concept involves a class of Bayesian nonparametric models to predict the number of items sold per transaction (or “basket”). This is a new approach involving novel Bayesian dynamic models that are customizable to diverse levels of sales from sporadic/intermittent to persistent. Section 6.2 concerns the integration of cross-series information using the novel multi-scale approach introduced in Section 4. We adapt this to forecasting transactions rather than sales; this enables relevant data shared in forecasting item-level demand, which is then coupled with the new binary cascade approach for sales per transaction. This decouple/recouple framework maximally exploits analytic tractability for sequential learning and forecasting for each individual item and enables information sharing across items while maintaining computational scalability; the resulting computational burden remains linear in the number of items. Section 6.3 develops and showcases a series of examples of the application of this new model class in analysis and forecasting of supermarket sales with a number of items evidencing substantially different features in sales levels and variation over time.

Summary comments and a discussion of potential future research areas in Section 7 conclude the dissertation.

2

Bayesian State-Space Models and Count Forecasting

2.1 Bayesian State-Space Modeling

The models presented in this dissertation fall within the framework of Bayesian state-space models. This chapter provides relevant information about dynamic generalized linear models (DGLMs), a type of Bayesian state-space model, which serve as a building block for the reminder of the dissertation. Bayesian state-space models are defined by an observation equation, evolution equation, and initial prior information. The observation equation represents the stochastic relationship between some output and an underlying state vector. The evolution equation specifies the structure of the random evolution of the state vector as a function of the previous value of the state vector.

There are many benefits to Bayesian state-space models in the context of product sales forecasting. Data are modeled on their natural scale rather than being transformed. Bayesian analysis naturally implements sequential learning and forecasting through state evolutions and prior-posterior updates at each time t. Bayesian forecasting utilizes full predictive distributions rather than just point forecasts, and simulation at any time point provides access to summary forecasts on arbitrary functions of the future data over multiple steps ahead. This allows forecasters to study quantities like the total sales over the next k days, the probability of exceeding some number of sales in the next k days, etc. Components of state-space models, such as levels, trends, seasonality, and regression components, are easily interpretable by non-statisticians. Time-varying model components allow non-stationarities to be captured and for models to adapt to unpredictable changes. Inference on time-varying model components can help understand how relationships between predictors and the outcome may vary through time. A key question in product sales forecasting is how changes in price and promotion affect sales and whether this effect varies over time. In addition to producing forecasts, Bayesian state-space models can provide insight into questions of this nature. Finally, the Bayesian framework allows incorporation of expert information or interventions into the model at any time point via modifications of "current" priors over state parameters. For example, store managers may want to intervene on a model if they have outside knowledge what may affect sales such as an upcoming hurricane/blizzard or other local information.

2.2 Sequential Learning in DGLMs

General notation follows that of standard Bayesian dynamic linear models (West and Harrison, 1997). Denote by y_t a univariate time series observed at discrete, equally-space times $t = 1, \dots, T$. At any time t, having observed $y_{1:t}$, the available information is denoted—and sequentially updated—by $\mathcal{D}_t = \{y_t, \mathcal{D}_{t-1}, \mathcal{I}_{t-1}\}$, where $\mathcal{I}_{t-1}$ represents any additional relevant information besides past data becoming available at time $t-1$ (such as covariate values and information used to define interventions in the model). For any vector of time indices $t+1{:}t+k$ for $k > 0$, forecasting $y_{t+1:t+k}$ at time t is based on the information set $\{\mathcal{D}_t, \mathcal{I}_t\}$. The full class of DGLMs defines Bayesian state-space modeling of data conditionally arising from distributions with exponential family form.

In a specified DGLM, the observation model has p.d.f. of exponential family form

$$p(y_t \mid \eta_t, \phi) = b(y_t, \phi) \exp\left[\phi\{y_t \eta_t - a(\eta_t)\}\right] \tag{2.1}$$

where η_t is the natural parameter that maps to linear predictor $\lambda_t = g(\eta_t)$ via link function $g(\cdot)$. The known scale factor ϕ is set to 1 in Bernoulli and Poisson models, and $b(\cdot,\cdot)$ is a known function specific to the chosen sampling distribution. Often, $g(\cdot)$ is set to the identity, and the natural parameter is the linear predictor. Dynamic regression is defined by the state-space form

$$\lambda_t = \mathbf{F}_t'\boldsymbol{\theta}_t \quad \text{where} \quad \boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t \quad \text{and} \quad \boldsymbol{\omega}_t \sim (\mathbf{0}, \mathbf{W}_t) \tag{2.2}$$

with the following elements:

- $\boldsymbol{\theta}_t$ is the latent, time-varying state vector, and $\mathbf{F}_t$ is a known vector of constants or realized values of predictor variables (a.k.a. regressors).
- The evolution equation in (2.2) specifies a conditionally linear Markov process for the state vector through time: $\mathbf{G}_t$ is a known state matrix specifying structural evolution of the state vector, and $\boldsymbol{\omega}_t$ is a stochastic innovation vector (or evolution "noise").
- The notation $\boldsymbol{\omega}_t \sim (\mathbf{0}, \mathbf{W}_t)$ indicates that $\mathrm{E}(\boldsymbol{\omega}_t|\mathcal{D}_{t-1}, \mathcal{I}_{t-1}) = \mathbf{0}$ and $\mathrm{V}(\boldsymbol{\omega}_t|\mathcal{D}_{t-1}, \mathcal{I}_{t-1}) = \mathbf{W}_t$, the latter variance matrix being known at time $t-1$. The distribution of $\boldsymbol{\omega}_t$ is otherwise not specified, but is generally expected to be unimodal and symmetric about zero.
- The $\boldsymbol{\omega}_t$ are independent over time and, at time $t-1$, $\boldsymbol{\omega}_t$ and $\boldsymbol{\theta}_{t-1}$ are conditionally independent given $\{\mathcal{D}_{t-1}, \mathcal{I}_{t-1}\}$.

The standard DGLM analysis (West et al., 1985; West and Harrison, 1997 chapter 15; Prado and West, 2010 section 4.4) has the following features.

1. At any time $t-1$, the current information is summarized in the posterior for the current state vector via the mean vector and variance matrix, namely $(\boldsymbol{\theta}_{t-1} \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1}) \sim (\mathbf{m}_{t-1}, \mathbf{C}_{t-1})$.
2. Through the evolution equation this induces 1−step ahead prior moments on the state vector of the form $(\boldsymbol{\theta}_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1}) \sim (\mathbf{a}_t, \mathbf{R}_t)$ with $\mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1}$ and $\mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t$.
3. A version of the so-called variational Bayes concept then applies to choose a conjugate prior for η_t, denoted by $(\eta_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1}) \sim \mathrm{CP}(\alpha_t, \beta_t)$ with form $p(\eta_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1}) = c(\alpha_t, \beta_t) \exp\{\alpha_t\eta_t - \beta_t a(\eta_t)\}$. Here $c(\cdot, \cdot)$ is a function of the hyper-parameters of known form depending on the specific exponential family model. See Section 2.5 for further discussion on the variational Bayes methods.
4. The hyper-parameters α_t and β_t are evaluated so that the conjugate prior satisfies the prior moment constraints

$$\mathrm{E}(\lambda_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1}) = f_t = \mathbf{F}_t'\mathbf{a}_t \quad \text{and} \quad \mathrm{V}(\lambda_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1}) = q_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t.$$

5. Forecasting y_t 1−step ahead uses the conjugacy-induced predictive distribution with p.d.f. $p(y_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1}) = b(y_t, \phi)c(\alpha_t, \beta_t)/c(\alpha_t + \phi y_t, \beta_t + \phi)$.
6. On observing y_t, the posterior for η_t has the conjugate form of $(\eta_t \mid \mathcal{D}_t) \sim CP(\alpha_t + \phi y_t, \beta_t + \phi)$.
7. Under this posterior, mapping back to the linear predictor $\lambda_t = g(\eta_t)$ implies posterior mean and variance $g_t = \mathrm{E}(\lambda_t \mid \mathcal{D}_t)$ and $p_t = \mathrm{V}(\lambda_t \mid \mathcal{D}_t)$.
8. Finally, linear Bayes updating implies posterior mean vector and variance matrix of the state vector as $(\boldsymbol{\theta}_t \mid \mathcal{D}_t) \sim (\mathbf{m}_t, \mathbf{C}_t)$ given by

$$\mathbf{m}_t = \mathbf{a}_t + \mathbf{R}_t\mathbf{F}_t(g_t - f_t)/q_t \quad \text{and} \quad \mathbf{C}_t = \mathbf{R}_t - \mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t'(1 - p_t/q_t)/q_t.$$

This completes the time $t-1$-to-t evolve-predict-update cycle.

This sequential learning scheme, detailed in (West et al., 1985), utilizes conjugate analysis and a variational Bayes concept for the state vector evolution, and linear Bayes theory for the state vector updating. The result is a computationally simple update similar to the Kalman filter which produces coherent forecast distributions in a sequential manner. Triantafyllopoulos (2009) compared the forecasting performance of this variational/linear Bayes approach to two alternative on-line DGLM estimation methods. The first method follows Fahrmeir (1992) in using standard normal approximations based on numerical prior-to-posterior mode updates and Taylor series expansion of the log posterior at each time point; this is one version of so-called extended Kalman filtering (West and Harrison, 1997 chapter 13; West, 1981). The second approach is a sequential Monte Carlo (SMC, or particle filtering) method. The linear/variational Bayes approach is shown to improve forecasting performance compared to the extended Kalman filter method, and to have a much lower computational cost and fewer implementation barriers than the sequential Monte Carlo method. Particle filters have several well-documented issues including weight degeneracy, sample impoverishment, the choice of appropriate importance densities, and on-line parameter learning (Arulampalam et al., 2002; Triantafyllopoulos, 2009). Although specialized approaches attempt to address each of these concerns, the implementation of particle filtering is not straightforward in a setting with thousands of series. Additionally, as addressed in the discussion following Lopes et al. (2011), weight degeneracy is unavoidable as the length of the data increases unless the corresponding number of particles also increases exponentially. As the length of our data increases, this quickly makes particle filtering a computationally infeasible approach in our high-dimensional setting of high-frequency data requiring on-line inference.

Some of the structure and computations implied require comment and are highlighted in the key cases of interest for count data. In each case, the link function $g(\cdot)$ is the identity so that $\eta_t = \lambda_t$.

2.3 Key Examples of DGLMs

2.3.1 Bernoulli Logistic DGLM

Here the series y_t is relabelled as $z_t = 0/1$ with $z_t \sim Ber(\pi_t)$ and $\eta_t = \text{logit}(\pi_t)$. In the exponential family p.d.f. form the terms are $\phi = 1$, $b(y_t, \phi) = 1$ and $a(\eta_t) = \log(1 + \exp(\eta_t))$. The conjugate prior in part 4 above is Beta, $\pi_t \sim Be(\alpha_t, \beta_t)$, with the hyper-parameters defining $f_t = \psi(\alpha_t) - \psi(\beta_t)$ and $q_t = \psi'(\alpha_t) + \psi'(\beta_t)$, where $\psi(\cdot)$ and $\psi'(\cdot)$ are the digamma and trigamma functions, respectively. The values (α_t, β_t) can be trivially computed from (f_t, q_t) via iterative numerical solution based on standard Newton-Raphson. The form of the $1-$step ahead forecast is Beta-Bernoulli with $(z_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1}) \sim BBer(1, \alpha_t, \beta_t)$ defined simply by $\Pr(z_t = 1 | \mathcal{D}_{t-1}, \mathcal{I}_{t-1}) = \alpha_t/(\alpha_t + \beta_t)$. The updated moments of the linear predictor in part 7 above are then trivially computed via the equations $g_t = \psi(\alpha_t + z_t) - \psi(\beta_t + 1 - z_t)$ and $p_t = \psi'(\alpha_t + z_t) + \psi'(\beta_t + 1 - z_t)$.

2.3.2 Poisson Loglinear DGLM

Here $y_t \sim Po(\mu_t)$ with $\eta_t = \log(\mu_t)$. In the exponential family p.d.f. form the terms are $\phi = 1$, $b(y_t, \phi) = 1/y_t!$ and $a(\eta_t) = \exp(\eta_t)$. The conjugate prior in part 4 above is Gamma, $\mu_t \sim Ga(\alpha_t, \beta_t)$, with the hyper-parameters defining $f_t = \psi(\alpha_t) - \log(\beta_t)$ and $q_t = \psi'(\alpha_t)$. The values (α_t, β_t) can be trivially computed from (f_t, q_t) via iterative numerical solution based on standard Newton-Raphson. The $1-$step ahead forecast is negative binomial, $(y_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1}) \sim Nb(\alpha_t, \beta_t/(1 + \beta_t))$. The updated moments of the linear predictor in part 7 above are trivially computed via the equations $g_t = \psi(\alpha_t + y_t) - \log(\beta_t + 1)$ and $p_t = \psi'(\alpha_t + y_t)$.

2.3.3 Normal DLM

We also note the special case of normal models when the DGLM reduces to a conditionally normal DLM. This is of relevance to count time series in case of large counts where a log transform—for example—of the count series can often be well-modeled using a normal DLM as an approximation. This also allows for inclusion of volatility via a time-varying conditional variance. With y_t the logged values of the original count series, a normal model has $y_t \sim N(\mu_t, v_t)$ with $\eta_t = \mu_t$. In the exponential family p.d.f. form the term ϕ becomes, generally, a time-dependent precision, $\phi_t = 1/v_t$, while $b(y_t, \phi_t) = (\phi_t/2\pi)^{1/2} \exp(-\phi_t y_t^2/2)$ and $a(\eta_t) = \eta_t^2/2$.

The conjugate prior in part 4 above is normal, $\mu_t \sim N(a_t, A_t v_t)$ which matches the general conjugate form with $\alpha_t = a_t/A_t$ and $\beta_t = 1/A_t$. Prior to posterior updating in part 8 reduces to a standard Kalman filter update. When embedded in the DLM, the additional assumption that the evolution noise terms $\boldsymbol{\omega}_t$ in eqn. (2.2) are also normal implies that DGLM evolution/updating equations are exact in this special case. However, for most practical applications it is relevant to also estimate the conditional variances $v_t = 1/\phi_t$. The simplest and most widely-used extension is that based on a standard Beta-Gamma stochastic volatility model for ϕ_t which, is analytically tractable. The resulting theory is then based on normal/inverse gamma prior and posterior distributions for (μ_t, v_t). Details of the resulting modifications to forward filtering and forecasting analysis are very standard (West and Harrison, 1997 chapter 4 and section 10.8; Prado and West, 2010 section 4.3).

2.4 Traditional Component Discounting

Specification of the required evolution variance matrices $\mathbf{W}_t$ in eqn. (2.2) uses the standard, parsimonious and effective discount method based on component discounting (West and Harrison, 1997, chapter 6). In most practical models the state vector is naturally partioned into components representing different explanatory effects, such as trends (e.g., local level, local gradient), seasonality (time-varying seasonal factors or Fourier coefficients) and effects of independent predictor variables. That is, for some integer J we have $\boldsymbol{\theta}_t' = (\boldsymbol{\theta}_{t1}', \dots, \boldsymbol{\theta}_{tJ}')$. It is natural to define $\mathbf{W}_t$ to represent potentially differing degrees of stochastic variation in these components and this is enabled using separate discount factors $\delta_1, \ldots, \delta_J$, where each $\delta_j \in (0, 1]$. A high discount factor implies a low level of stochastic change in the corresponding elements of the state vector, and vice-versa (with $\delta_j = 1$ implying no stochastic noise at all—obviously desirable but rarely practically relevant). The definition of $\mathbf{W}_t$ is as follows.

From Section 2.2 part 2 above, the time $t-1$ prior variance matrix of $\mathbf{G}_t\boldsymbol{\theta}_{t-1}$ is $\mathbf{P}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t'$; this represents information levels about the state vector following the deterministic evolution via $\mathbf{G}_t$ but before the impact of the evolution noise that then simply adds $\mathbf{W}_t$. Write $\mathbf{P}_{tj}$ for the diagonal block of $\mathbf{P}_t$ corresponding to state subvector $\boldsymbol{\theta}_{tj}$ and set

$$\mathbf{W}_t = \text{block diag}[\mathbf{P}_{t1}(1-\delta_1)/\delta_1, \ldots, \mathbf{P}_{tJ}(1-\delta_J)/\delta_J].$$

Then the implied prior variance matrix of $\boldsymbol{\theta}_t$ following the evolution has corresponding diagonal block elements $\mathbf{R}_{tj} = \mathbf{P}_{tj}/\delta_j$ while maintaining off-diagonal blocks from $\mathbf{P}_t$. Thus, the stochastic part of the evolution increases uncertainties about state vector elements in each subvector j by $100(1-\delta_j)/\delta_j\%$, maintains the correlations in $\mathbf{P}_{tj}$ for state elements within the subvector j, while reduces cross-correlations between state vector elements in differing subvectors. In practice, high values of the δ_j are desirable and typical applications use values in the range $0.97 - 0.999$ with, generally, robustness in terms of forecasting performance with respect to values in the range. Evaluation of forecast metrics on training data using different choices of discount factors is a basic strategy in model building and tuning.

2.5 Variational Bayes Techniques for Conjugate Prior Specification

In Section 2.2, we discussed a variational Bayes concept in step 3 of the standard DGLM analysis that has been in use in ranges of applications since the early 1980s since West et al. (1985). In this section, we provide additional details about this method and introduce two alternative approaches. In Section 2.2 part 2 above, we have 1-step ahead prior mean and variance $(\mathbf{a}_t, \mathbf{R}_t)$ on the state vector $\boldsymbol{\theta}_t$. Given the relationship $\lambda_t = \mathbf{F}_t'\boldsymbol{\theta}_t$ in eqn. (2.2), the prior moments of $\boldsymbol{\theta}_t$ imply prior moments of λ_t. Although this defines prior moments for λ_t, there is not a fully specified prior distribution on λ_t since we do not specify the distribution of $\boldsymbol{\theta}_t$. In order to define a closed form predictive distribution for y_t, we must fully specify the prior for λ_t. We could conceivably choose any prior distributions, and we specify conjugate priors for several reasons. Since our likelihood is of exponential family form, a conjugate prior will alway exist. Furthermore, with a conjugate prior, the resulting 1-step predictive distribution of y_t can be written in closed form as a function of the exponential family normalizing constant. Finally, after y_t is observed, the conjugate prior-to-posterior update is mathematically and computationally simple.

We denote the conjugate prior on the linear predictor as $\lambda_t \sim CP(\alpha_t, \beta_t)$. The specific form of the conjugate prior depends on the specified likelihood distribution. Given our choice of conjugate prior, there are several ways we can select values of hyperparameters α_t and β_t. Ideally, we would like to select (α_t, β_t) so that the conjugate prior is, in some sense, "close" to the prior moments on $\boldsymbol{\theta}_t$. We describe three methods for choosing hyperparameters in the following sections, and compare the applied results of these three methods in Section 2.5.3.

2.5.1 Moment Matching

In this approach, as in the original methodology of DGLMs (West et al., 1985), and following Section 2.2 part 4 above, we choose the conjugate hyperparameters so that the mean and variance of the conjugate prior are the same as the mean and variance implied on η_t by $(\boldsymbol{\theta}_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1})$. The moment matching approach for Bernoulli and Poisson distributions was briefly described in Sections 2.3.1 and 2.3.2, and we expand upon the details for the Poisson distribution here.

The conjugate prior for the Poisson distribution with mean μ_t is $\mu_t \sim Ga(\alpha_t, \beta_t)$. The linear predictor λ_t is equal to the natural parameter of the Poisson distribution, $\eta_t = \log(\mu_t)$. Given that μ_t follows a Gamma distribution and $\log(\mu_t)$ is a sufficient statistic of the Gamma distribution, we can use the moment generating function of the sufficient statistic to find the mean and variance of η_t. Using this method, we find that $\mathrm{E}(\eta_t) = \mathrm{E}(\log(\mu_t)) = \psi(\alpha_t) - \log(\beta_t)$ and $\mathrm{V}(\eta_t) = \mathrm{V}(\log(\mu_t)) = \psi'(\alpha_t)$.

The moment matching method selects (α_t, β_t) so that the mean of $\log(\mu_t)$ under the Gamma prior is equal to f_t, and the variance of $\log(\mu_t)$ under the Gamma prior is equal to q_t. Plugging in the previous formulas for the mean and variance of $\log(\mu_t)$, we get the following system of equations: $f_t = \psi(\alpha_t) - \log(\beta_t)$ and $q_t = \psi'(\alpha_t)$. Given values of f_t and q_t, we can use an iterative method such as Newton's method to find the values of (α_t, β_t) that satisfy these equations.

2.5.2 KL Divergence Minimization

Through the relationship $\lambda_t = \mathbf{F}_t'\boldsymbol{\theta}_t$, the state vector prior, $(\boldsymbol{\theta}_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1})$, induces prior moments on η_t such that $\eta_t \sim (f_t, q_t)$. We denote this implied prior as $p(\eta_t)$. As discussed above, we specify a conjugate prior form on η_t in this setting. The conjugate prior density has the form

$$q(\eta_t) = p(\eta_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1}) = c(\alpha_t, \beta_t) \exp\{\alpha_t \eta_t - \beta_t a(\eta_t)\} \tag{2.3}$$

for some parameters (α_t, β_t) to be chosen. In this approach, our objective is to choose (α_t, β_t) so that $q(\cdot)$ is "close" to $p(\cdot)$. We measure the "distance" between two probability distributions with the Kullback-Liebler (KL) divergence. We select (α_t, β_t) in order to minimize the KL divergence between priors $p(\cdot)$ and $q(\cdot)$. Since KL divergence is not symmetric, we consider both the forward and reverse KL divergence. Minimizing the forward and reverse KL divergence will result in two unique Gamma priors, so we consider these different methods for the remainder of this section.

Forward KL Divergence

The forward KL divergence of $q(\cdot)$ from $p(\cdot)$ is

$$\begin{aligned} K(p \mid\mid q) = K_f &= \int \log\left(\frac{p(\eta_t)}{q(\eta_t)}\right) p(\eta_t) d\eta_t \\ &= \int \log(p(\eta_t))p(\eta_t)d\eta_t - \int \log(q(\eta_t))p(\eta_t)d\eta_t. \qquad (2.4) \end{aligned}$$

Note that the first term is constant with respect to q. In order to minimize K_f, we must maximize the second term. Using eqn. (2.3), we can write $\log(q(\eta_t))$ as

$$\log(q(\eta_t)) = \log(c(\alpha_t, \beta_t)) + \alpha_t\eta_t - \beta_t a(\eta_t).$$

Plugging this into the second term of eqn. (2.4), we see that, as a function of (α_t, β_t),

$$K_f = C - \log(c(\alpha_t, \beta_t)) + \int \{\alpha_t\eta_t - \beta_t a(\eta_t)\} p(\eta_t) d\eta_t$$

where C is a constant. To find the minimum of K_f, we take the partial derivative of K_f with respect to α_t and β_t and set it equal to zero. That is,

$$\begin{aligned} -\frac{\partial K_f}{\partial \alpha_t} = 0 &= \frac{\partial c(\alpha_t, \beta_t)/\partial \alpha_t}{c(\alpha_t, \beta_t)} + \int \eta_t p(\eta_t) d\eta_t \\ \Rightarrow \mathrm{E}_p(\eta_t) &= -\frac{\partial c(\alpha_t, \beta_t)/\partial \alpha_t}{c(\alpha_t, \beta_t)} \qquad (2.5) \end{aligned}$$

and

$$\begin{aligned} -\frac{\partial K_f}{\partial \beta_t} = 0 &= \frac{\partial c(\alpha_t, \beta_t)/\partial \beta_t}{c(\alpha_t, \beta_t)} - \int a(\eta_t) p(\eta_t) d\eta_t \\ \Rightarrow \mathrm{E}_p(a(\eta_t)) &= \frac{\partial c(\alpha_t, \beta_t)/\partial \beta_t}{c(\alpha_t, \beta_t)}. \qquad (2.6) \end{aligned}$$

The notation $E_p(\cdot)$ specifies that we are taking the expectation with respect to $p(\cdot)$ rather than $q(\cdot)$. Given the values of $\mathrm{E}_p(\eta_t)$ and $\mathrm{E}_p(a(\eta_t))$, we can solve for the

(α_t, β_t) that minimize K_f. Note that $c(\alpha_t, \beta_t)$ is a constant from the conjugate prior for the exponential family when written in terms of the natural parameter. For common conjugate priors (ex: Beta), we must reparameterize the exponential family in terms of the natural parameter. These result applies generally to all natural exponential family distributions. We present the specific results for Poisson and Bernoulli sampling distributions below.

Poisson Forward KL Divergence

For details on the Poisson DGLM, see Section 2.3.2. Through a change of variables, the conjugate prior $\mu_t \sim Ga(\alpha_t, \beta_t)$ can be written in terms of $\eta_t = \log(\mu_t)$ as

$$q(\eta_t) = \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \exp\{\alpha_t \eta_t - \beta_t e^{\eta_t}\}$$

where $c(\alpha_t, \beta_t) = \beta_t^{\alpha_t}/\Gamma(\alpha_t)$ from eqn. (2.3). Before applying eqns. (2.5) and (2.6), we compute the partial derivatives of $c(\alpha_t, \beta_t)$ and divide by $c(\alpha_t, \beta_t)$ to obtain

$$\frac{\partial c(\alpha_t, \beta_t)/\partial \alpha_t}{c(\alpha_t, \beta_t)} = \log(\beta_t) - \psi(\alpha_t)$$

and

$$\frac{\partial c(\alpha_t, \beta_t)/\partial \beta_t}{c(\alpha_t, \beta_t)} = \frac{\alpha_t}{\beta_t}.$$

Plugging these partial derivatives into eqns. (2.5) and (2.6), and solving for α_t and β_t, we obtain the following result:

$$\beta_t = \exp\{\psi(\alpha_t) - \mathrm{E}_p(\log(\mu_t))\}, \tag{2.7}$$

$$\alpha_t = \beta_t \mathrm{E}_p(\mu_t) = \mathrm{E}_p(\mu_t) \exp\{\psi(\alpha_t) - \mathrm{E}_p(\log(\mu_t))\}. \tag{2.8}$$

To solve for (α_t, β_t), we must know the expectation of μ_t and $\log(\mu_t)$ with respect to $p(\cdot)$. Under $p(\cdot)$, we know that $\mathrm{E}(\log(\mu_t)) = f_t$, but the mean of μ_t is unknown unless we make additional assumptions about $p(\cdot)$. If we assume that $p(\cdot)$ follows a normal distribution ($\log(\mu_t) \sim N(f_t, q_t)$), then the expectation of μ_t is $\exp\{f_t + q_t/2\}$. Given f_t and q_t, we can solve for (α_t, β_t) using an iterative solver like Newton's method.

Bernoulli Forward KL Divergence

For details on the Bernoulli DGLM, see Section 2.3.1. The conjugate prior $\pi_t \sim Be(a_t, b_t)$ can be written in terms of the natural parameter $\eta_t = \text{logit}(\pi_t)$ as

$$q(\eta_t) = \frac{\Gamma(a_t + b_t)}{\Gamma(a_t)\Gamma(b_t)} \exp\{a_t\eta_t - (a_t + b_t)\log(e^{\eta_t} + 1)\}.$$

To write this prior in the same notation as eqn. (2.3), we substitute $\alpha_t = a_t$ and $\beta_t = a_t + b_t$ so that

$$q(\eta_t) = \frac{\Gamma(\beta_t)}{\Gamma(\alpha_t)\Gamma(\beta_t - \alpha_t)} \exp\{\alpha_t\eta_t - \beta_t\log(e^{\eta_t} + 1)\}.$$

Written in this form, we identify $c(\alpha_t, \beta_t) = \Gamma(\beta_t)/(\Gamma(\alpha_t)\Gamma(\beta_t - \alpha_t))$. We now compute the partial derivatives of $c(\alpha_t, \beta_t)$ and divide by $c(\alpha_t, \beta_t)$ to obtain

$$\begin{aligned}\frac{\partial c(\alpha_t, \beta_t)/\partial\alpha_t}{c(\alpha_t, \beta_t)} &= \psi(\beta_t - \alpha_t) - \psi(\alpha_t)\\ &= \psi(b_t) - \psi(a_t),\end{aligned}$$

and

$$\begin{aligned}\frac{\partial c(\alpha_t, \beta_t)/\partial\beta_t}{c(\alpha_t, \beta_t)} &= \psi(\beta_t) - \psi(\beta_t - \alpha_t)\\ &= \psi(a_t + b_t) - \psi(b_t).\end{aligned}$$

We now solve for the initial Beta prior hyper parameters a_t and b_t by plugging these partial derivatives into eqns. (2.5) and (2.6). This results in

$$\begin{aligned}\text{E}_p(\text{logit}(\pi_t)) &= \psi(a_t) - \psi(b_t),\\ \text{E}_p(\log(1 - \pi_t)) &= \psi(b_t) - \psi(a_t + b_t).\end{aligned}$$

Under $p(\cdot)$, we know that the expectation of $\text{logit}(\pi_t)$ is f_t. However, we do not know the expectation of $\log(1-\pi_t)$ unless we make further assumptions about $p(\cdot)$. In order to solve these equations, we can assume that $p(\cdot)$ follows a normal distribution, so that $\text{logit}(\pi_t) \sim N(f_t, q_t)$. Under this assumption, there is no analytical solution for the expectation of $\log(1-\pi_t)$, but we can find this expectation using numerical integration.

Reverse KL Divergence

The reverse KL divergence of $p(\cdot)$ from $q(\cdot)$ is

$$\begin{aligned} K(q \mid\mid p) = K_r &= \int \log\left(\frac{q(\eta_t)}{p(\eta_t)}\right) q(\eta_t) d\eta_t \\ &= \underbrace{\int \log(q(\eta_t)) q(\eta_t) d\eta_t}_{A} - \underbrace{\int \log(p(\eta_t)) q(\eta_t) d\eta_t}_{B}. \end{aligned} \tag{2.9}$$

First, we focus on term A, and plug in $\log(q(\eta_t))$ using eqn. (2.3). This term simplifies to

$$A = \log(c(\alpha_t, \beta_t)) + \alpha_t \text{E}_q(\eta_t) - \beta_t \text{E}_q(a(\eta_t)). \tag{2.10}$$

Since the second term includes $q(\eta_t)$, we must assume a distribution for $p(\eta_t)$ in order to maximize the second integral. We assume that $p(\cdot)$ is a normal distribution, so that $\eta_t \sim N(f_t, q_t)$. Plugging in $\log(p(\eta_t))$, this term simplifies to

$$B = -\frac{1}{2q}\text{E}_q(\eta_t^2) + \frac{f}{q}\text{E}_q(\eta_t). \tag{2.11}$$

Note that we are now taking expectations with respect to $q(\cdot)$, so the expectations will depend on (α_t, β_t). In order to take derivatives of these terms, we must write the expectations in terms of α_t and β_t. We present these details below for both Poisson and Bernoulli distributions.

Poisson Reverse KL Divergence

In the Poisson setting, the conjugate prior $q(\cdot)$ implies that $\mu_t \sim Ga(\alpha_t, \beta_t)$. For a Gamma random variable, $X \sim Ga(a, b)$, the following are true:

$$\begin{aligned} \mathrm{E}(\log(X)) &= \psi(a) - \log(b), \\ \mathrm{V}(\log(X)) &= \psi'(a). \end{aligned}$$

We can simplify eqns. (2.10) and (2.11) by plugging in the expectations with respect to $q(\cdot)$. Under $q(\cdot)$, these terms become

$$\begin{aligned} A &= \log(c(\alpha_t, \beta_t)) + \alpha_t(\psi(\alpha_t) - \log(\beta_t)) - \beta_t(\alpha_t/\beta_t) \\ &= \log(c(\alpha_t, \beta_t)) + \alpha_t(\psi(\alpha_t) - \log(\beta_t)) - \alpha_t, \end{aligned}$$

and

$$B = -\frac{1}{2q}\left(\psi'(\alpha_t) + \psi(\alpha_t)^2 - 2\log(\beta_t)\psi(\alpha_t) + \log(\beta_t)^2\right) + \frac{f}{q}\left(\psi(\alpha_t) - \log(\beta_t)\right).$$

The expectation of η_t^2 can be found by using the fact that $\mathrm{E}_q(\eta_t^2) = \mathrm{V}_q(\eta_t) + \mathrm{E}_q(\eta_t)^2$. Plugging A and B back into eqn. (2.9), and taking derivatives with respect to α_t and β_t, we can minimize the reverse KL divergence. After minimizing, we can solve the following system of equation for α_t and β_t:

$$\begin{aligned} f_t &= \psi(\alpha_t) - \log(\beta_t), \\ q_t\alpha_t\psi'(\alpha_t) &= q_t - \frac{\psi''(\alpha_t)}{2}. \end{aligned}$$

We can use an iterative solver, such as the Newton-Raphson method, to solve for α_t and β_t given values of f_t and q_t.

Bernoulli Reverse KL Divergence

In the Bernoulli setting, the conjugate prior $q(\cdot)$ implies that $\pi_t \sim Be(a_t, b_t)$. For a beta distributed random variable, $X \sim Be(a, b)$, the following are true:

$$\begin{aligned}
\mathrm{E}(\log(X)) &= \psi(a) - \psi(a+b), \\
\mathrm{E}(\log(1-X)) &= \psi(b) - \psi(a+b), \\
\mathrm{E}(\mathrm{logit}(X)) &= \mathrm{E}\left(\log\left(\frac{X}{1-X}\right)\right) = \mathrm{E}(\log(X)) - \mathrm{E}(\log(1-X)) \\
&= \psi(a) - \psi(b), \\
\mathrm{V}(\mathrm{logit}(X)) &= \psi'(a) + \psi'(b).
\end{aligned}$$

Using these formulas, we can simplify eqns. (2.10) and (2.11) by plugging in the expectations with respect to $q(\cdot)$. Using the expectation and variance of $\mathrm{logit}(\pi_t)$, we can solve for the expectation of $\mathrm{logit}(\pi_t)^2$. Under this Bernoulli setting, $a(\eta_t) = \log(1 + e^{\eta_t})$ implies that $a(\eta_t) = -\log(1 - \pi_t)$, and we can use the previous facts about Beta random variables to find the expectation of $a(\eta_t)$. Under $q(\cdot)$, A and B simplify to:

$$\begin{aligned}
A &= \log(c(a_t, b_t)) + a_t(\psi(a_t) - \psi(b_t)) - (a_t + b_t)(\psi(a_t + b_t) - \psi(b_t)), \\
B &= -\frac{1}{2q}\left(\psi'(a_t) + \psi'(b_t) + \psi(a_t)^2 - 2\psi(a_t)\psi(b_t) + \psi(b_t)^2\right) + \frac{f}{q}(\psi(a_t) - \psi(b_t)).
\end{aligned}$$

Given f_t and q_t, we can use a numerical optimization method to select (a_t, b_t) which maximize K_r.

2.5.3 Applied Comparison of Variational Methods

In this section, we present an example comparing the conjugate priors using the moment matching and forward/reverse KL divergence minimization procedures. We assume a Poisson likelihood, $y_t \sim Po(\mu_t)$, and conjugate Gamma prior, $\mu_t \sim Ga(\alpha_t, \beta_t)$. For the two KL divergence procedures, we assume that distribution $p(\eta_t) \sim N(f_t, q_t)$.

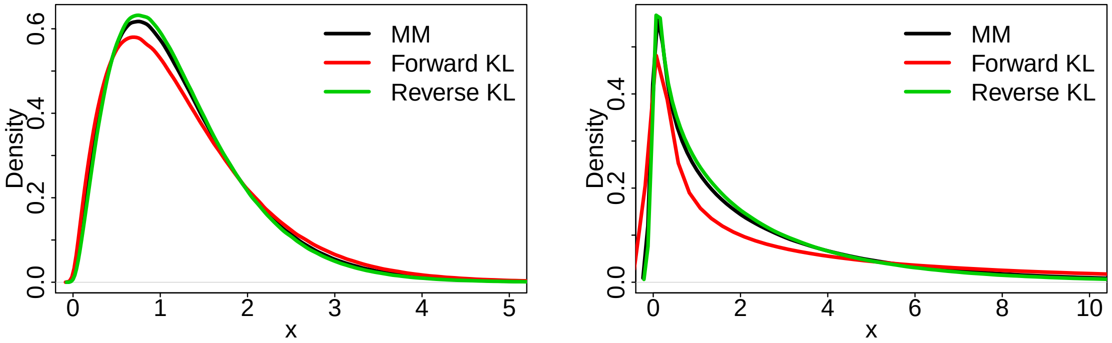


Figure 2.1: Comparison of conjugate Gamma priors under the moment matching (MM) method (black), forward KL divergence method (red), and reverse KL divergence method (green) when $f_t = 0$, and $q_t = 0.5$ (left) and 3 (right).

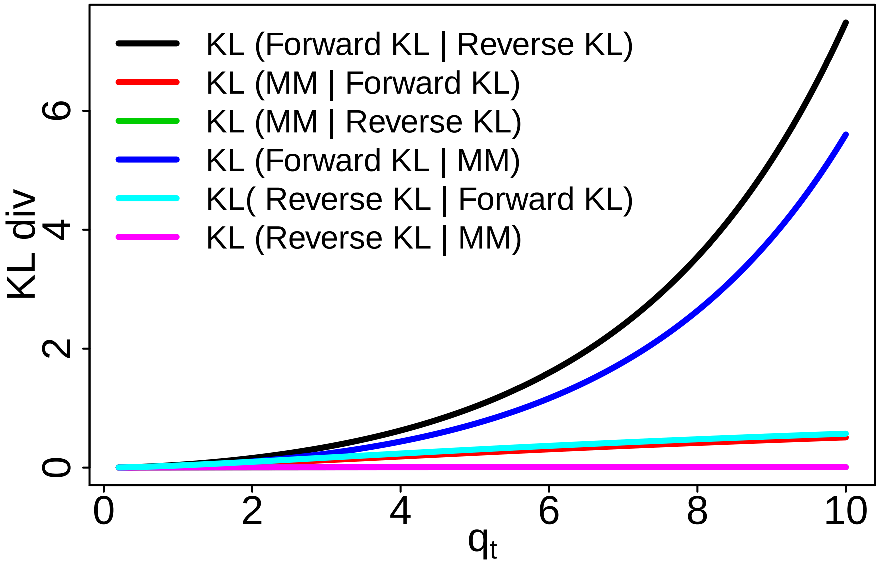


Figure 2.2: KL divergence of conjugate Gamma priors versus q_t. We compare conjugate priors found using the moment matching (MM) method, forward KL divergence method, and reverse KL divergence method.

The moment matching method does not make any assumptions about $p(\cdot)$ besides the mean and variance. We assume $f_t = 0$, and consider values of q_t between 0.2 and 10. We use the KL divergence to assess the "distance" between the resulting Gamma priors. The KL divergence of "approximating" $Ga(a_q, b_q)$ from "true" $Ga(a_p, b_p)$ is:

$$(a_q - a_p)\psi(\alpha_p) - \log\psi(a_p) + \log\psi(a_q) + a_q(\log(b_p) - \log(b_q)) + a_p(b_q - b_p)/b_p.$$

Figure 2.1 plots the conjugate Gamma priors using the moment matching and forward/reverse KL divergence methods with $q_t = 0.5$ (left) and 3 (right). The black line is the moment matching prior, the red line is the forward KL divergence prior, and the green line is the reverse KL divergence prior. Note that the x-axes on these plots do not show the full support of the priors, and that we only plot the partial support to illustrate the differences between the priors.

When $q_t = 0.5$, all three priors are similar in terms of location and shape. The moment matching method results in a $Ga(2.46, 1.98)$ prior, the forward KL procedure results in $Ga(2.15, 1.68)$, and the reverse KL divergence procedure returns $Ga(2.59, 2.11)$. The moment matching and reverse KL priors are most similar, although the moment matching prior has a slightly longer tail than the reverse KL prior. The forward KL prior has a slightly longer tail than the moment matching prior. When $q_t = 3$, the density of all three priors is concentrated near zero. The moment matching prior is $Ga(0.68, 0.27)$, the forward KL prior is $Ga(0.43, 0.10)$, and the reverse KL prior is $Ga(0.74, 0.33)$. Again, the moment matching and reverse KL priors are very similar, with the tail of the moment matching prior very slightly longer. The forward KL prior has a longer tail than the moment matching and reverse KL priors.

Figure 2.2 plots the KL divergence between the conjugate Gamma priors versus q_t under the moment matching, forward KL, and reverse KL methods. Since the KL divergence is not symmetric, we consider the forward and reverse KL divergence between each combination of the three priors. From this plot, we can see that KL divergence increases as q_t increases. As we noticed in the previous plots, the moment matching and reverse KL divergence priors are most similar and have the smallest KL divergence across q_t. Note that the KL divergence from the moment matching to the reverse KL divergence prior (green line) is covered by the pink line. The largest KL divergences occur when comparing the forward KL divergence prior to the reverse KL divergence prior and the moment matching prior.

From this comparison, we can conclude that the moment matching and reverse KL divergence methods result in very similar conjugate prior specifications. The moment matching prior has a slightly longer tail than the reverse KL divergence prior. The forward KL divergence prior results in the conjugate prior with the longest tail. This result can be explained by the appearance of a $\log(q(\eta_t)/p(\eta_t))$ term in the forward KL divergence, and $\log(p(\eta_t)/q(\eta_t))$ in the reverse KL divergence. The forward KL divergence will be large when $q(\eta_t)$ is close to zero where $p(\eta_t)$ is nonzero. This behavior will result in an optimal q that is always greater than zero when $p(\cdot)$ is greater than zero. This explains the longer tail and higher variance of the forward KL conjugate priors. The reverse KL divergence will be large when $p(\cdot)$ is close to zero where $q(\cdot)$ is nonzero. This results in an optimal $q(\cdot)$ that is zero when $p(\cdot)$ is zero. Therefore, the reverse KL divergence approach will result in a conjugate prior with lower variance and a shorter tail than the forward KL divergence prior.

2.6 Forecast Evaluation for Count Data

The consumer sales/demand forecasting and other literatures cover a range of metrics for forecast evaluation, with variants of traditional loss functions for point forecasts customized to count data with concern, particularly, for cases of low counts (e.g., among recent contributions, see Kolassa, 2016; Snyder et al., 2012; Hyndman and Koehler, 2006; Fildes and Goodwin, 2007; Yelland, 2009; Gneiting, 2011; Morlidge, 2015; Czado et al., 2009). As noted in Hyndman and Koehler (2006), simply adopting common measures of forecast accuracy can produce "misleading results" when applied to low valued count data. According to Kolassa (2016), retailers have many considerations regarding store replenishment including logistical constraints (pack sizes, delivery schedules, truck loads), complex cost functionals, and different aggregation levels (store-level replenishment, ordering from manufacturer, promotion planning). Given that consumer sales forecasts feed into so many different decisions, they argue against focusing on single functionals like the mean, median, or quantiles in favor of focusing on a correct predictive distribution.

Our view is that "a forecast" is the full predictive distribution rather than one or more point summaries. For actionable decisions, understanding the potential implications of uncertainty as reflected in the full distribution can be key, while also adding significantly to evaluation and comparisons of forecast accuracy. Any point forecast selected should be rationalized and understood as a decision made on the basis of utility/loss considerations in the forecasting context, with implicit or explicit derivation from a decision analysis perspective. The predictive mean is optimal under squared error loss, the median for absolute error loss, and the mode for the (typically not substantively relevant) 0-1 loss. If the loss function is an asymmetric piecewise linear function, $L(f, y) = (\mathbb{1}(y < f) - \alpha)(f - y)$ for outcome y with point forecast f, then the α-quantile of the predictive distribution is the optimal point forecast. Modifications of the absolute percentage error (APE) loss function $L(f, y) = |1 - f/y|$ are commonly used in consumer demand forecasting of strictly positive counts y. Under this loss, the optimal f is the (-1)-median, i.e., the median of the p.d.f. $g(y) \propto p(y)/y$ when $p(y)$ is the forecast p.d.f., although typical application is based on sub-optimal choices of f as full forecast distributions are rarely developed. In Section 2.6.1, we provide additional details about APE and the effects of using non-optimal point forecasts on the expected loss. It is also easy to explore novel modifications and extensions of loss functions from a decision analysis perspective. For example, APE does not allow for zero outcomes, while practical extensions—such as ZAPE, with $L(f, y) = |1 - f/y|\mathbb{1}(y > 0) + l(f)\mathbb{1}(y = 0)$ for some increasing function $l(f) > 0$—are amenable to simple optimization to define relevant point forecasts if desired. We explore ZAPE further in Section 2.6.2. Specific loss functions should be chosen in the context of resulting decisions to be made. In inventory control, there are costs associated with missed sales due to stock-outs, as well as the cost of overstocking items. The forecaster may be interested in a quantile of the distribution to reflect these utilities.

2.6.1 Absolute Percentage Error

Consider a forecasting scenario where random variable $y > 0$ has p.d.f. $p(y)$ and c.d.f. $P(y)$. We present a proof that the optimal point forecast under APE loss is the (-1)-median. Gneiting (2011) presents the more general result that all loss functions of the form $L(y, f) = y^\beta |y - f|$ have the β-median as the optimal point forecast (the median of the density proportional to $y^\beta p(y)$.) An alternative representation for APE is the piecewise expression

$$L(y, f) = \frac{|y - f|}{y} = \begin{cases} 1 - \frac{f}{y}, & \text{if } y \geqslant f, \\ \frac{f}{y} - 1, & \text{if } y < f. \end{cases} \tag{2.12}$$

Define the p.d.f. $g(y)$ such that $g(y) = cp(y)/y$ where c is a normalizing constant, and let G be the corresponding c.d.f. By plugging in APE as written in eqn. (2.12), we can express the implied risk function $R(f)$ in the following manner:

$$
\begin{aligned}
R(f) &= \int_0^\infty L(y,f)dP(y) = \int_0^\infty L(y,f)p(y)dy \\
&= \int_0^f L(y,f)p(y)dy + \int_f^\infty L(y,f)p(y)dy \\
&= \int_0^f \left(\frac{f}{y} - 1\right) p(y)dy + \int_f^\infty \left(1 - \frac{f}{y}\right) p(y)dy \\
&= \int_0^f \frac{f}{c}\frac{cp(y)}{y}dy - \int_0^f p(y)dy + \int_f^\infty p(y)dy - \int_f^\infty \frac{f}{c}\frac{cp(y)}{y}dy \\
&= \frac{f}{c}G(f) - P(f) + (1 - P(f)) - \frac{f}{c}(1 - G(f)) \\
&= \frac{2f}{c}G(f) - 2P(f) + 1 - \frac{f}{c}.
\end{aligned}
$$

Then, the APE optimal point forecast is the minimizer of $R(f)$, defined by

$$
\begin{aligned}
\frac{\partial R(f)}{\partial f} = 0 &= \frac{2}{c}\left(G(f) + cp(f)\right) - 2p(f) - \frac{1}{c} \\
&= \frac{2}{c}G(f) - \frac{1}{c} \\
&\Rightarrow G(f) = \frac{1}{2}.
\end{aligned}
$$

The optimal point forecast f under APE loss is the median of $g(y)$, or the so-called (-1)-median of $p(y)$. If $p(y)$ is a discrete distribution, then the $1/y$ term in $g(y)$ downweights mass on $y > 1$, and, as a result, the median of $g(y)$ is less than or equal to the median of $p(y)$. For some distributions $p(y)$, there exist closed form expressions for the (-1)-median. For example, the lognormal distribution has a closed form (-1)-median, and we discuss this specific example further below. In other cases, closed form expressions do not exist, but sampling from $p(y)$ allows estimation of the (-1)-median using the following importance sampling scheme.

Importance Sampling for the (-1)-median of $p(y)$:

1. Assume that we have n independent samples from $p(y)$, and denote the ordered samples as $y_1, \ldots, y_n$.
2. Compute the importance weights, $w_i^* = p(y_i)/y_i$
3. Normalize weights to sum to one, i.e. define $w_i = w_i^*/\sum_i w_i^*$.
4. Define a to be the maximum index such that $\sum_{i=1}^{a} w_i \leqslant 0.5$, and b the minimum index such that $\sum_{i=1}^{b} w_i \geqslant 0.5$.
5. The (-1)-median lies in the range (y_a, y_b). If $p(\cdot)$ is continuous, then the y_i will be distinct, $a = b$, and this will produce a unique median of $g(\cdot)$. If $p(\cdot)$ is discrete, then $a \neq b$, and the range (y_a, y_b) may contain multiple values.

Since $g(\cdot)$ has more mass towards zero and a lighter upper tail than $p(\cdot)$, we can expect the importance sampling weights to be well-behaved. The effective sample size (ESS) of this importance sampling scheme is ESS $= 100(n \sum_{i=1}^{n} w_i^2)^{-1}$.

Example: Lognormal case

To illustrate the (-1)-median and the value of APE under different point forecasts, we assume that $p(\cdot)$ is a lognormal p.d.f. Let y be a lognormal random variable such that $y \sim p(\cdot) \equiv LN(m, v)$. In other words, $x = \log(y)$ implies that $x \sim N(m, v)$. The corresponding c.d.f. is $P(y) = \Phi\left(v^{-1/2}(\log y - m)\right)$ where Φ is the standard normal c.d.f. Given this specification of $p(\cdot)$, then $g(y) = cp(y)/y$ is a lognormal p.d.f. such that $g(\cdot) \equiv LN(m - v, v)$. For random variable $y \sim p(\cdot)$, the median is $\exp(m)$, the expectation is $\exp(m + v/2)$, and the mode is $\exp(m - v)$. For random variable $z \sim g(\cdot)$, the median is $\exp(m - v)$, and the expectation is $\exp(m - v/2)$. The (-1)-median of lognormal y is equivalent to the mode of y, and has value $\exp(m - v)$. For low values of v, $p(y)$ is less skewed, and the (-1)-median is close to the median of

$\exp(m)$. As v increases, $p(y)$ becomes more positively skewed, and the (-1)-median approaches zero. Since the (-1)-median is the optimal point forecast under APE, it will result in the lowest expected APE loss. Next, we compare the expected loss under optimal and non-optimal point forecasts f. Plugging in eqn. (2.12), we can simplify $R(f)$ as

$$\begin{aligned}
R(f) &= \int_0^\infty L(y,f)p(y)dy \\
&= \int_0^f \frac{f-y}{y}p(y)dy + \int_f^\infty \frac{y-f}{y}p(y)dy \\
&= \int_0^f \left(\frac{f}{y}-1\right)p(y)dy + \int_f^\infty \left(1-\frac{f}{y}\right)p(y)dy \\
&= f\int_0^\infty \frac{p(y)}{y}dy - \int_0^f p(y)dy + \int_f^\infty p(y)dy - f\int_f^\infty \frac{p(y)}{y}dy \\
&= \frac{f}{c}\int_0^\infty g(y)dy - \int_0^f p(y)dy + \left(1-\int_0^f p(y)dy\right) - \frac{f}{c}\left(1-\int_0^f g(y)dy\right) \\
&= \frac{2f}{c}\int_0^f g(y)dy - 2\int_0^f p(y)dy + 1 - \frac{f}{c}.
\end{aligned}$$

In this expression, c is the normalizing constant for $g(\cdot)$ which is equal to $\exp(m-v/2)$ in this lognormal example. We now simplify $R(f)$ under the assumption of the lognormal p.d.f. $p(y)$ and specific choices of f. We consider values of f equal to the optimal (-1)-median point forecast and the non-optimal mean, median, and (-1)-mean point forecasts. The (-1)-mean of y is defined as the mean of $g(y)$.

First, we find the expected APE loss for the optimal (-1)-median point forecast $f=\exp(m-v)$.

This is

$$\begin{aligned}
\mathrm{E}(L(y,f)) &= 2\exp\left(-\frac{v}{2}\right)\int\limits_{0}^{\exp(m-v)} g(y)dy - 2\int\limits_{0}^{\exp(m-v)} p(y)dy + 1 - \exp\left(-\frac{v}{2}\right) \\
&= 2\exp\left(-\frac{v}{2}\right)\Phi\left(0\right) - 2\Phi\left(-v^{1/2}\right) + 1 - \exp\left(-\frac{v}{2}\right) \\
&= 1 - 2\Phi(-v^{1/2}).
\end{aligned}$$

For the optimal f, the expected loss is between zero and one, and depends only on v. As v gets close to zero, the expected loss approaches zero. As v increases, the expected loss approaches one. Next, we find the expected APE loss when the point forecast is the median of y. This is

$$\begin{aligned}
\mathrm{E}(L(y,\exp(m))) &= 2\exp\left(\frac{v}{2}\right)\int\limits_{0}^{\exp(m)} g(y)dy - 2\int\limits_{0}^{\exp(m)} p(y)dy + 1 - \exp\left(\frac{v}{2}\right) \\
&= 2\exp\left(\frac{v}{2}\right)\Phi(v^{1/2}) - 2\Phi(0) + 1 - \exp\left(\frac{v}{2}\right) \\
&= \exp\left(\frac{v}{2}\right)\left(1 - 2\Phi(-v^{1/2})\right) \\
&= \exp\left(\frac{v}{2}\right)\mathrm{E}(L(y,f)).
\end{aligned}$$

Using the non-optimal median of y as a point forecast, the expected loss varies between zero and infinity. The expected loss under the median is higher than the expected loss under the (-1)-median, and this difference increases as v increases. Next, we find the expected loss with the mean of y as the point forecast. This is

$$\begin{aligned}
\mathrm{E}(L(y,\exp(m+v/2))) &= 2\exp(v)\Phi\left(\frac{3}{2}v^{1/2}\right) - 2\Phi\left(\frac{1}{2}v^{1/2}\right) + 1 - \exp(v) \\
&= \exp(v)\left(1 - 2\Phi\left(-\frac{3}{2}v^{1/2}\right)\right) + \left(1 - 2\Phi\left(\frac{1}{2}v^{1/2}\right)\right).
\end{aligned}$$

The expected APE loss under the mean varies between zero and infinity. Due to the $\exp(v)$ term, this expected loss will increase much faster with v than the expected loss under the previous point forecasts. Finally, we find the expected loss when our forecast is the (-1)-mean of y, i.e. the mean of $g(y)$. This is

$$\mathrm{E}(L(y, \exp(m - v/2))) = 2\left(1 - 2\Phi\left(-\frac{1}{2}v^{1/2}\right)\right).$$

Under the (-1)-mean, the expected loss varies between zero and two. As v goes to zero, the expected loss will approach zero. As v increases, the expected loss will approach two. Table 2.1 displays the expected APE loss for each point forecast when $v = 0.1, 1, 2, 5, 10$. When $v = 0.1$, the expected loss is very similar for each of the point forecasts. However, as v increases, it is clear that the (-1)-median and (-1)-mean have lower expected loss than the mean and median. From the expression for each forecast f, we can see that APE increases as the forecast f increases. The APE highly penalizes large point forecasts f. This example shows that improper usage of point forecasts like the mean and median under APE can result in large forecast errors.

In the previous derivations, we have used the standard decision analysis definition of an optimal point forecast as the one which minimizes the expected loss. However, the loss has an entire distribution, and we may be interested in other quantities from this loss distribution like the median or an upper quantile. For example, we can derive the point forecast f which minimizes the probability that the loss exceeds some quantile q. In this section, we derive the point forecast which minimizes $Q(f) = Pr(L(y, f) > q)$ for the APE loss function. An alternative expression for $Q(f)$ is $Q(f) = 1 - Pr(L(y, f) \leqslant q)$. Plugging in the APE loss, this simplifies to

$$\begin{aligned} Pr\left(\left|\frac{y-f}{y}\right| \leqslant q\right) &= Pr\left(-q \leqslant \frac{y-f}{y} \leqslant q\right) \\ &= Pr\left(-q - 1 \leqslant -f/y \leqslant q - 1\right) \\ &= Pr\left(1 - q \leqslant f/y \leqslant q + 1\right). \end{aligned}$$

Now, we use the example with $y \sim LN(m, v)$ to further simplify $Q(f)$. Recall that the lognormal distribution on y implies that if $x = \log(y)$, then $x \sim N(m, v)$. Using the fact that linear transformations of normal random variables are normally distributed, we can show that if $z = \log(1/y) = -\log(y)$, then $z \sim N(-m, v)$. Additionally, if $w = \log(ay) = \log(a) + \log(y)$ for some constant a, then $w \sim N(\log(a) + m, v)$. These characteristics imply that $f/y \sim LN(\log(f) - m, v)$. Then, we can simplify $Q(f)$ using the lognormal c.d.f. as

$$Pr\left(\left|\frac{y-f}{y}\right| \leqslant q\right) = \Phi\left(\frac{\log(q+1) - \log f + m}{v^{1/2}}\right) - \Phi\left(\frac{\log(1-q) - \log f + m}{v^{1/2}}\right).$$

Consider $q \geqslant 1$, and notice that if $f = 0$, then the APE will be one with probability one. Therefore, if $q \geqslant 1$, then $f = 0$ will always minimize $Q(f)$. Given this behavior, we now find the f which minimizes $Q(f)$ for $q < 1$. In the following derivation, let ϕ denote the standard normal p.d.f., i.e. $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$. We have

$$\begin{aligned}
\frac{\partial Q(f)}{\partial f} = 0 = \phi\left(\frac{\log(q+1) - \log f + m}{v^{1/2}}\right)\left(-\frac{1}{fv^{1/2}}\right) - \\
\phi\left(\frac{\log(1-q) - \log f + m}{v^{1/2}}\right)\left(-\frac{1}{fv^{1/2}}\right) \\
&= \exp\left\{-(1/2v)(\log(q+1) - \log f + m)^2\right\} - \\
&\quad \exp\left\{-(1/2v)(\log(1-q) - \log f + m)^2\right\} \\
&\Rightarrow (\log(q+1) - \log f + m)^2 = (\log(1-q) - \log f + m)^2 \\
&\Rightarrow \log(q+1) - \log f + m = -\log(1-q) + \log f - m \\
&\Rightarrow f = \exp\left[m + \left(\frac{\log(1-q^2)}{2}\right)\right].
\end{aligned}$$

The point forecast that minimizes $Q(f)$ depends on q and m, but not on v. Depending on q, this point forecast will be between zero and $\exp(m)$, the median of $p(y)$. As q approaches one, the optimal point forecast goes to zero. When q is zero, the optimal point forecast is the median. Table 2.2 displays the $Q(f)$-minimizing point forecasts for $q = 0.01, 0.1, 0.5, 0.9, 0.99$ and several values of m and $\exp(m)$. We see that when

Table 2.1: Expected APE loss for different values of v and point forecasts f from $p(y)$.

f	Expression	$v = 0.1$	$v = 1$	$v = 2$	$v = 5$	$v = 10$
(-1)-median	$\exp(m - v)$	0.25	0.68	0.84	0.97	1.00
(-1)-mean	$\exp(m - v/2)$	0.25	0.76	1.04	1.47	1.77
median	$\exp(m)$	0.26	1.13	2.29	11.87	148.18
mean	$\exp(m + v/2)$	0.28	1.97	6.62	147.56	22025.53

Table 2.2: Optimal point forecasts that minimize the probability of APE exceeding q for different values of m.

m	$\exp(m)$	$q = 0.01$	$q = 0.1$	$q = 0.5$	$q = 0.9$	$q = 0.99$
-2	0.14	0.14	0.13	0.12	0.06	0.02
-1	0.37	0.37	0.37	0.32	0.16	0.05
0	1.00	1.00	0.99	0.87	0.44	0.14
1	2.72	2.72	2.70	2.35	1.18	0.38
2	7.39	7.39	7.35	6.40	3.22	1.04

q is low, the point forecast is very close to the median. As q approaches one, the point forecast decreases. For higher values of m, the optimal point forecast decreases more slowly to zero than for lower values of m. Using this table, we can see that the optimal point forecast for minimizing expected loss versus minimizing $Q(f)$ may vary substantially for low values of q. For example, let $m = 2$, $v = 1$, and $q = 0.1$. The (-1)-median of 2.72 minimizes the expected loss, but $f = 7.35$ minimizes $Q(f)$ for $q = 0.1$. As v increases, the (-1)-median will decrease to zero while the $f = 7.35$ will still minimize $Q(f)$.

In this section, we have assumed that $y > 0$ since the APE is undefined when $y = 0$. However, in our application of forecasting daily sales, zeros are a common occurrence. In the next section, we discuss a simple extension of APE to accommodate zero-valued observations.

2.6.2 Zero-Adjusted Absolute Percentage Error

In this section, we study a practical extension of APE that allows for zero outcomes. We define the ZAPE as $L(y, f) = |1 - f/y|\mathbb{1}(y > 0) + l(f)\mathbb{1}(y = 0)$ where $l(f)$ is a non-negative and non-decreasing function. Examples of $l(f)$ include $l(f) = kf$ for $k > 0$ and $l(f) = \min(f, 1)$. We consider the context where y is a non-negative integer, and $p(y)$ is discrete probability distribution. Let $\pi_0 = p(0)$, and define $g(y) = cy^{-1}p(y)\mathbb{1}(y \geqslant 1)$ with the corresponding c.d.f. $G(y)$ and normalizing constant $c \geqslant 1$. Plugging in the expression for ZAPE, we can simplify the risk $R(f)$ as

$$R(f) = \sum_{y=0}^{\infty} L(y, f)p(y) = l(f)\pi_0 + \sum_{y=1}^{\infty} |y - f|y^{-1}p(y).$$

The risk is more dominated by the first term for items with higher probabilities of observing a zero. As π_0 increases, high point forecasts are increasingly penalized. Consider $l(f) = kf$ as a first case to explore. Larger values of k increasingly penalize large forecasts when a zero is observed. Using direct calculus, we can show that, for this choice of $l(f)$, $R(f)$ is minimized at f given by

$$f = \begin{cases} 0, & \text{if } kc\pi_0 \geqslant 1, \\ G^{-1}((1 - kc\pi_0)/2), & \text{if } kc\pi_0 < 1. \end{cases}$$

Since k, c, and π_0 are positive, the optimal point forecast under ZAPE is less than or equal to the median of $g(y)$. When π_0 is close to zero, then f is chose to the median of $g(y)$, i.e. the (-1)-median of $p(y)$. For large values of π_0, the value c increases as f becomes smaller and eventually hits zero. In these cases, the optimal point forecast is exactly zero. Next, we consider $l(f) = \min(1, f)$, and derive the optimal point forecast f given by

$$f = \begin{cases} 0, & \text{if } c\pi_0 \geqslant 1, \\ G^{-1}(0.5), & \text{if } c\pi_0 < 1. \end{cases}$$

As π_0 increases, the value of c increases as f decreases to zero. As π_0 decreases, the optimal point forecast f is the (-1)-median which is the same as the optimal point forecast under APE.

Both of these ZAPE optimal point forecasts depend on the value of c, the normalizing constant of $g(y)$. Given c, we can use importance sampling from $p(\cdot)$ to approximate the required quantiles of $G(\cdot)$. In cases where the closed form expression of $p(y)$ is known, we can estimate c through numerical integration. In other cases, we can sample from $p(\cdot)$, but there is no closed form expression for the distribution. In these cases, we can use direct or accept/reject Monte Carlo estimation to estimate c.

Direct Monte Carlo

Given a random sample from $p(\cdot)$, we denote the non-zero samples as $y_1, \ldots, y_n$, and define sample frequencies, $\hat{p}(y)$ on each value of y. A direct Monte Carlo approximation to the probability $g(y)$ at any y value is $\hat{g}(y) = c\hat{p}(y)/y$. Define $\theta = 1/c$, and observe that normalizing implies the estimated value $\hat{\theta} = 1/\hat{c} = \sum_{y=1}^{\infty} \hat{p}(y)/y$. An equivalent expression for this Monte Carlo estimate is $\hat{\theta} = (1/n)\sum_{i=1}^{n} 1/y_i$ and the variance is $\mathrm{V}(\hat{\theta}) = (1/n)\mathrm{V}(1/y)$. Let $x = 1/y$, and note that $y \geqslant 1$ implies $x \in (0, 1]$. It follows directly that $\mathrm{E}(x) \leqslant 1$ and $\mathrm{E}(x^2) \leqslant \mathrm{E}(x)$. We can bound the variance of x as follows

$$\mathrm{V}(x) = \mathrm{E}(x^2) - \mathrm{E}(x)^2 \leqslant \mathrm{E}(x) - \mathrm{E}(x)^2 = \mathrm{E}(x)(1 - \mathrm{E}(x)) \leqslant 0.25.$$

Using this bound, we see that the variance of the Monte Carlo estimate of θ is finite and bounded by $1/4n$. Given that $\hat{c}$ is a function of $\hat{\theta}$, we can apply the delta method to approximate the variance of $\hat{c}$ as n goes to infinity. First we note that $\hat{\theta}$ is an unbiased, finite variance Monte Carlo estimator of θ and that $\sqrt{n}(\hat{\theta} - \theta) \to N(0, \sigma^2)$ in distribution for some finite variance $\sigma^2 \leqslant 1/4n$. Given this assumption and the fact that $\hat{c} = f(\hat{\theta}) = 1/\hat{\theta}$, the delta method implies that

$$\sqrt{n}(\hat{c} - c) \to N\left(0, \frac{\sigma^2}{\theta^4}\right).$$

Since $\theta \in (0, 1]$, the variance of $\hat{c}$ will be larger than that of $\hat{\theta}$. For small values of θ, the variance of $\hat{c}$ could be large. The value of θ will be small either when $p(0)$ is large, or when $p(y)$ has most of its mass on large integers rather than low integers.

Accept/Reject Monte Carlo

Alternatively, standard acceptance sampling applies to this setting as follows.

1. Since $y \geqslant 1$, the quantity $\pi(y) = g(y)/(cp(y)) = 1/y$ lies in $(0, 1]$.

2. Generate a single Monte Carlo draw $y \sim p(\cdot)$.

3. Evaluate the probability $\pi(y) = 1/y$ and then accept y with this probability. If accepted, then record y as a draw from $g(y)$; otherwise reject it and redraw.

4. Equivalently, generate $u \sim U(0, 1)$ independently of y. Then, accept y as a draw from $g(\cdot)$ if $u \leqslant \pi(y)$.

The key of the underlying theory is that the distribution of y conditional on having been accepted is the correct distribution with p.d.f. $g(y)$. By Bayes' theorem, the p.d.f. of an accepted y is

$$\frac{p(y)Pr(u \leqslant \pi(y) \mid y)}{Pr(u \leqslant \pi(y))}$$

where

- $Pr(u \leqslant \pi(y) \mid y) = \pi(y)$ since $u \sim U(0, 1)$, and

- $Pr(u \leqslant \pi(y)) = \sum_y Pr(u \leqslant \pi(y) \mid y)p(y) = \sum_y \pi(y)p(y) = \sum_y g(y)/c = 1/c.$

Thus, the resulting posterior p.d.f. of y conditional on acceptance is $cp(y)\pi(y) = g(y)$ as required. If accepted, the realized value comes from the target distribution $g(y)$. Repeat this for many independent trials to generate a Monte Carlo sample of independent draws from $g(\cdot)$ given by the accepted values.

2.6.3 Decision Analysis for Multiple Series

We have discussed details of decision analysis in forecasting a univariate y. However, in the context of product sales forecasting, we may be interested in forecasting multivariate $\mathbf{y}$ where each element represents a series or a different forecast horizon. Let $\mathbf{y} = y_{1:n} = (y_1, \ldots, y_n)'$ for n individual observations, and $p(\mathbf{y})$ denote the multivariate forecast p.d.f. The corresponding point forecasts are $\mathbf{f} = f_{1:n} = (f_1, \ldots, f_n)'$. For any series i, the outcomes and forecasts of all other $n-1$ series are denoted by $\mathbf{y}_{-i}$ and $\mathbf{f}_{-i}$.

Mean Absolute Percentage Error

We can extend the APE loss function for a set of n series via

$$L(\mathbf{y}, \mathbf{f}) = \frac{1}{n} \sum_{i=1:n} \frac{|y_i - f_i|}{y_i}.$$

As we noted for APE, the MAPE is inappropriate for contexts with frequent zeros since the loss is undefined if any $y_i = 0$. Even if MAPE is defined, when we aggregate across series, forecast errors for low values of y_i are over-penalized compared to the same errors for high values of y_i. For example, a forecast error of one results in an APE of 1 when $y_i = 1$ and an APE of 0.1 when $y_i = 10$. In a context of forecasting n individual items, MAPE will be dominated by low-selling items rather than high-selling items. This quality is generally undesirable since high selling items are generally more important than low selling items in an inventory control context. The implied risk function for MAPE is

$$R(\mathbf{f}) = \sum_{\mathbf{y}} L(\mathbf{y}, \mathbf{f}) p(\mathbf{y}) = \sum_{i=1:n} R_i(f_i)$$

where

$$R_i(f_i) = \sum_{y_i=1}^{\infty} y_i^{-1} |y_i - f_i| p_i(y_i).$$

For each series i, let $p_i(y_i)$ represent the implied univariate margin for y_i. The MAPE risk is the sum of n terms, and each term is the APE risk for the given series y_i. The MAPE risk is minimized by individually minimizing the APE risk for each series. Thus, the MAPE optimal forecast is the vector of (-1)-medians of each marginal $p_i(y_i)$.

Weighted Absolute Percentage Error

Weighted absolute percentage error (WAPE) is a modification of MAPE aimed at resolving the issues of division by zero. The denominator of y_i in MAPE is replaced by the average of all of the y_i values in WAPE. WAPE is defined for the specific set of n series via

$$L(\mathbf{y},\mathbf{f}) = a(\mathbf{y}) \sum_{i=1:n} |y_i - f_i| \quad \text{with} \quad a(\mathbf{y})^{-1} = \sum_{i=1:n} y_i.$$

This definition assumes that at least one $y_i > 0$ so that $a(\mathbf{y})$ is defined and thus $0 < a(\mathbf{y}) \leqslant 1$. Often, WAPE is defined as the ratio of the MAD to the mean of the y_i. The intuition underlying WAPE is that the error for each series is weighed relative to the sum of the y_i. Conditional on the sum of y_i, the impact of a forecast error of one on WAPE is the same if $y_i = 1$ or $y_i = 10$. The implied risk function is

$$R(\mathbf{f}) \equiv \sum_{\mathbf{y}} L(\mathbf{y},\mathbf{f}) p(\mathbf{y}) d\mathbf{y} = \sum_{i=1:n} R_i(f_i)$$

where

$$R_i(f_i) = \sum_{y_i=0}^{\infty} |y_i - f_i| h_i(y_i) \quad \text{with} \quad h_i(y_i) = \sum_{\mathbf{y}_{-i}} a(\mathbf{y}) p(\mathbf{y}).$$

This is always defined and non-negative. The risk function is the sum of n non-negative terms and, in terms of dependence on $\mathbf{f}$, the term for each item i depends only on f_i. As a result, the minimum overall risk is achieved by separately minimizing each of the terms in the sum with respect to that item-specific forecast. Define the joint p.d.f. $g(\mathbf{y}) = ca(\mathbf{y})p(\mathbf{y})$ where c is the implied normalizing constant. Note that c depends on n and the selected items, as well as $p(\cdot)$, but not on $\mathbf{f}$. For each series i, write $g_i(y_i)$ for the implied univariate margin for y_i.

The WAPE-optimal forecast vector $\mathbf{f}$ is the vector of medians of the n marginal p.d.f.s under the joint p.d.f. $g(\mathbf{y}) = ca(\mathbf{y})p(\mathbf{y})$ where $a(\mathbf{y})^{-1} = \sum_{i=1:n} y_i$. The special case of APE when $n = 1$ and $p(0) = 0$ gives the well-known optimal f_1 as the (-1)-median, i.e., the median of $g(y_1) \propto p(y_1)/y_1$. In other cases, this is an extension of the (-1)-median concept to a joint distribution. Note that this result applies for any weighting function $a(\mathbf{y})$ that depends only on $\mathbf{y}$ and not $\mathbf{f}$.

Weighted Absolute Forecast Error

The weighted absolute forecast error (WAFE) is defined for n series via

$$L(\mathbf{y}, \mathbf{f}) = a(\mathbf{y}, \mathbf{f}) \sum_{i=1:n} |y_i - f_i| \quad \text{with} \quad a(\mathbf{y}, \mathbf{f})^{-1} = \sum_{i=1:n} (y_i + f_i)/2.$$

The implied risk function again has the form of a sum of n terms, with the term for series i given by

$$R_i(f_i) = \sum_{y_i=0}^{\infty} |y_i - f_i| h_i(y_i \mid \mathbf{f}) \quad \text{with} \quad h_i(y_i \mid \mathbf{f}) = \sum_{\mathbf{y}_{-i}} a(\mathbf{y}, \mathbf{f}) p(\mathbf{y}).$$

This is not amenable to the same analysis as WAPE since the full forecast vector $\mathbf{f}$ appears in the function $h_i(y_i \mid \mathbf{f})$.

2.6.4 Evaluating Predictive Count Distributions

In addition to point forecasts, it is also important to report predictive uncertainty intervals. Let $p(\cdot)$ denote a forecast p.d.f., and $P(\cdot)$ the corresponding c.d.f. For continuous $P(\cdot)$, the quantile function is defined as $P^{-1}(u) = \{x : P(x) = u\}$. The

$100(1-\alpha)\%$ central interval is defined as the interval (a, b) such that $a = P^{-1}(\alpha/2)$ and $b = P^{-1}(1-\alpha/2)$. The resulting interval has exactly $100\alpha/2\%$ of the density below a and above b. For discrete $P(\cdot)$, the quantile function is defined as $P^{-}(u) = \min\{x : P(x) \geqslant u\}$. For discrete distributions, it is not always possible to construct central intervals with exactly $100(1-\alpha)\%$ density. The discrete $100(1-\alpha)\%$ central interval is defined as (a, b) where $a = P^{-}(\alpha/2)$ and $b = P^{-}(1-\alpha/2)$. The resulting interval (a, b) will include greater than or equal to $100(1-\alpha)\%$ of the predictive density.

An alternative uncertainty region is the highest posterior/predictive density (HPD) region. The $100(1-\alpha)\%$ HPD region is defined as the set of values that contain $100(1-\alpha)\%$ of the density such that the density of values inside the region is higher than all values outside of the region. For discrete distributions, we define the $100(1-\alpha)\%$ HPD interval to be the values that contain at least $100(1-\alpha)\%$ of the density, where values within the region have higher mass than those excluded. For unimodal and symmetric distributions, the central interval and HPD intervals are the same. For bimodal distributions, the HPD region will often include multiple disjoint regions whereas the central interval is, by definition, a single interval. Except for distributions of high counts, count distributions are often neither symmetric nor unimodal. Being bounded by zero and the frequency of extreme high values often cause positive skew in the distribution. Additionally, the common zero inflation of count data can lead to multiple modes in forecast distributions. As a result, we recommend using HPD predictive regions rather than central predictive intervals. Given simulations from a discrete distribution, we can easily approximate a $100(1-\alpha)\%$ HPD region as the smallest region corresponding to at least $100(1-\alpha)\%$ of the sorted empirical p.m.f. values.

Figure 2.3 displays a simulated count distribution and the corresponding 50% central interval and 50% HPD region. Due to the discrete nature of the data, the central interval includes 55% of the density, and the HPD region includes 56% of the density. Despite both being 50% regions, these two regions provide very different summaries of the underlying distribution. The central interval omits the two values with the highest density, and includes values 3–5 with relatively low density. The HPD region is composed of two separate intervals: 0–1 and 8–11. The HPD region includes fewer values than the central interval, and yet it includes all of the values with the largest predictive density.

In this dissertation, we summarize predictive uncertainty through HPD regions rather than central predictive intervals. With a single outcome, we only know whether the observed value falls into the corresponding 50% HPD region, but this does not indicate whether our 50% predictive region was accurate. However, with repeat outcomes, we expect about 50% of the observed values to be included in the corresponding 50% HPD regions. The observed coverage of our $100(1-\alpha)\%$ predictive regions over time is the percentage of observed values that are included within the $100(1-\alpha)\%$ HPD regions. Nominal coverage indicates that our intervals are, on average, well calibrated to the observed values. Under coverage indicates that, on average, our predictive intervals are too narrow. Over coverage indicates that, on average, our predictive intervals are too wide. We recommend assessing the coverage of many widths of intervals ranging from $0 : 100\%$ probability.

Probability Integral Transform

An additional tool for evaluating entire forecast distributions is the probability integral transform (PIT) (Rosenblatt, 1952). Assume that at time t, we forecast based on a p.d.f. $\widehat{p}_t$ with corresponding c.d.f. $\widehat{P}_t$, and that the data actually follows a true c.d.f. P_t. After observing the outcome y_t, we define the PIT as

$$u_t = \widehat{P}_t(y_t) = \int_{-\infty}^{y_t} \widehat{p}_t(y)dy.$$

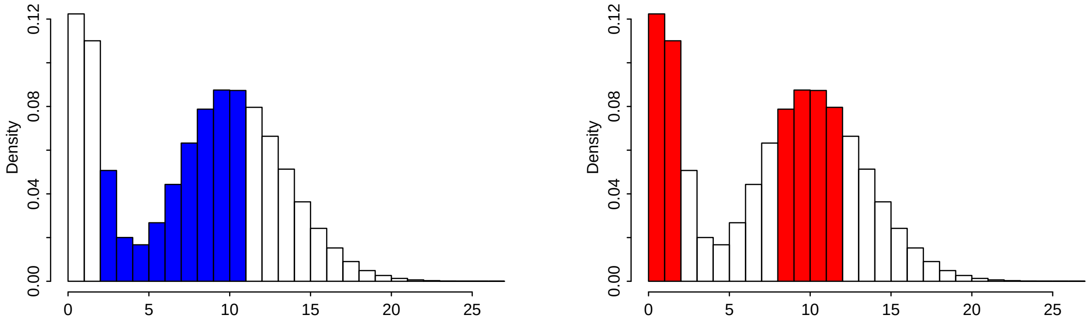


FIGURE 2.3: Histograms of a simulated count distribution and the shaded 50% central interval (left) and 50% highest predictive density region (right).

Suppose first that all distributions are continuous. Then, if $\widehat{P}_t = P_t$, the $u_t \sim U(0,1)$ since $Pr(u_t \leqslant u) = Pr(P_t(y_t) \leqslant u) = Pr(y_t \leqslant P_t^{-1}(u)) = P(P^{-1}(u)) = u$. Given u_t values defined over a period of time, we can assess the uniformity of the PIT values using histograms or probability plots. Any deviations from uniformity can diagnose misspecification of the forecast distributions. However, the previously defined u_t values are not uniform when P_t is discrete since the c.d.f. and quantile function take on discrete values. For cases of discrete predictive distributions, we apply the randomized PIT described in Kolassa (2016). For discrete $\widehat{P}_t$, we draw the randomized PIT values as

$$u_t \sim U(\widehat{P}_t(y_t - 1), \widehat{P}_t(y_t))$$

where we define $\widehat{P}_t(-1) = 0$. If $\widehat{P}_t = P_t$, then the u_t are again uniformly distributed on $(0,1)$. Figure 2.4 displays the PIT and rPIT plots for simulated Negative Binomial observations and a Poisson predictive distribution. The left plot displays the PIT values for the count data, and we see that the PIT values are discrete and clearly not uniformly distributed. The right plot displays the randomized PIT values for count data, and there is a slight S-shape to the rPIT values. This shape indicates that the assumed Poisson predictive distribution is narrower than the true Negative Binomial distribution of the observations.

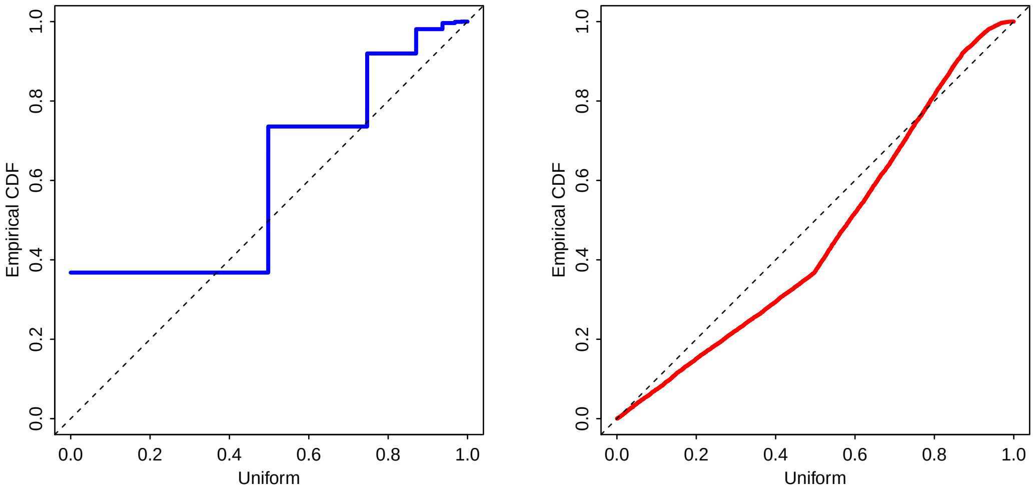


FIGURE 2.4: Probability plots of probability integral transform values versus uniform quantiles for simulated count data. (a): Left, application of the standard PIT to simulated Negative Binomial data. (b): Right, the rPIT applied to simulated Negative Binomial observations with assumed Poisson forecast distribution.

Discrete Ranked Probability Score

The PIT q-q plots are useful tools when evaluating probabilistic forecasts for individual series, but it is difficult to use PIT plots to assess forecasting across many series. In those contexts, it is useful to have a numerical summary/metric for evaluating probabilistic forecasts. The discrete ranked probability score (DRPS) is a proper scoring rule to assesses the location and width of probability distributions (Kolassa, 2016; Snyder et al., 2012). For observation y_t and forecast c.d.f. $\widehat{P}_t$, the discrete ranked probability score is defined as

$$DRPS(\widehat{P}_t, y_t) = \sum_{k=0}^{\infty} (\widehat{P}_t(k) - \mathbb{1}(y \leqslant k))^2.$$

The DRPS is ideal in our DGLM framework since, unlike other scoring rules, it can be calculated when forecasts are based on simulation. Long-term forecasting performance can be summarized by averaging DRPS over time. Comparing DRPS of multiple models can models can tell us which has better performance, however it does not give any information on where the two models differ. In contrast, the rPIT values can shed light on differences in predictive distributions.

Binary Calibration

We are also concerned with the long-run performance of forecasts on binary outcomes (specifically zero/non-zero sales). Ideal calibration of binary forecasts means that, of the days binary outcomes are forecast with probability near p, the binary outcomes will occur on approximately $100p\%$ of days. We assess the accuracy of binary predictions over time using binary calibration plots (or reliability diagrams). Reliability diagrams are frequently used in evaluating long-term weather forecasts such as precipitation forecasts (Weisheimer and Palmer, 2014; Hartmann et al., 2002). In a binary calibration plot, we bin the forecast probabilities either into equal width bins or into bins of equal sample size. Then, for the days within each bin, we evaluate the realized frequency of binary outcomes. We use shading to display the width of each bin, and approximate 95% binomial confidence intervals to indicate the uncertainty around the observed proportion in each bin. The binary calibration plot has the predicted probability on the x-axis, and the observed frequency on the y-axis. Reliable or well-calibrated predictions would lead the observed frequency or confidence interval in each bin to fall within the shaded region of probabilities.

3

Dynamic Count Mixture Models

3.1 Flexible Mixtures of DGLMs: Dynamic Count Mixture Models

A DCMM combines binary and conditionally Poisson DGLMs in a format similar to various existing models for time series of non-negative counts. It is often practically imperative to treat zero versus non-zero outcomes separately from forecasting the integer outcomes conditional on them being non-zero. The novelty here is to use the flexible classes of DGLMs for the two components in an overall model, with dynamic predictor components in each that can be customized to context. With non-negative count time series y_t, define the binary series $z_t = \mathbb{1}(y_t > 0)$ where $\mathbb{1}(\cdot)$ is the indicator function. A DCMM for outcomes y_t is defined by observation distributions in which

$$z_t \sim Ber(\pi_t) \quad \text{and} \quad y_t \mid z_t = \begin{cases} 0, & \text{if } z_t = 0, \\ 1 + x_t, \quad x_t \sim Po(\mu_t), & \text{if } z_t = 1, \end{cases}$$

over all time t. The parameters π_t and μ_t are time-varying according to binary and Poisson DGLMs, respectively, i.e.,

$$\text{logit}(\pi_t) = \mathbf{F}_t^{0\prime}\boldsymbol{\xi}_t \quad \text{and} \quad \log(\mu_t) = \mathbf{F}_t^{+\prime}\boldsymbol{\theta}_t \tag{3.1}$$

with latent state vectors $\boldsymbol{\xi}_t$ and $\boldsymbol{\theta}_t$ and known dynamic regression vectors $\mathbf{F}_t^0$ and $\mathbf{F}_t^+$, in an obvious notation. The DCMM independently models π_t, the probability of a non-zero count, and μ_t, the expected size of the non-zero count. The regression vectors can include distinct model components if we expect different factors to impact π_t and μ_t. The conditional model for $(y_t \mid z_t = 1)$ is a shifted Poisson DGLM. In sequential learning, the positive count model component will be updated only when $z_t = 1$. When a zero count is observed, the positive count value is implicitly treated as missing. This combination of DGLMs allows for a range of applications with a substantial probability of zeros over time. If, on the other hand, a time series has few or no zeros, the binary model will play a relatively limited role in forecasting. In our motivating application, this flexibility is essential due to the varying frequency of zero sales across items (low versus high sellers) and over time (e.g. seasonal products).

The forward filtering and forecast analysis evolves and updates prior moments of the state vectors in each of the binary and shifted Poisson models at each step separately. Then, each forecasts one or more steps ahead by evolving state vector moments into the future, and applying the variational Bayes constraint to conditional conjugacy. Thus, the marginal predictive distribution for y_{t+k} at any $k > 0$ steps ahead from time t is the implied mixture of a Bernoulli and shifted Poisson, with the conjugate gamma prior predictive for the Poisson rate μ_{t+k} defining a conditional shifted negative binomial forecast distribution for that component. In most applications, however, we are interested in full joint forecasts of paths $y_{t+1:t+k}$ over a sequence of future times $1{:}k$ from the current time t. Looking at these joint predictive distributions provides information on dependencies between time points, and allows for calculation of other forecast quantities. For example, in forecasting daily sales of a supermarket item over each of the next $k = 14$ days we may also be interested in quantities such as such the cumulative sales up to each day in that period, the number of those 14 days with zero sales, the probability that cumulative sales exceed some specified level, and so forth. To adequately (or at all) address such broader practical forecasting questions, we forward-simulate the predictive distributions. That is, generate large Monte Carlo samples of the full predictive distribution $p(y_{t+1:t+k}|\mathcal{D}_t, \mathcal{I}_t)$ from the current time t. This is easily implemented as noted and detailed in Section 3.3, and such Monte Carlo samples can be trivially manipulated and interrogated to quantify forecast distributions for any function of the series of future outcomes of interest.

3.2 DCMM Random Effects Extension

Due to the binary DGLM component, DCMMs can flexibly model time series of counts with many or few zero counts. Another common characteristic of non-negative count data– especially at higher levels of counts– is over-dispersion relative to conditional Poisson models. The primary contextual development of the shifted Poisson DGLM will aim to customize the choice of $\mathbf{F}_t^+$ and associated evolution equation to best predict non-zero sales. While resulting forecast distributions may be generally accurate in terms of location, they may still turn out to under-estimate uncertainties and, in particular, fail to adequately capture infrequent extremes (typically higher, though sometimes also lower values of y_t). Various approaches to this appear in the literature, but all essentially come down to adding a representation of this excess and purely unpredictable variation. This is best addressed directly via random effects, and this is easily done in the DCMM using a novel random effects extension in the Poisson DGLM component.

Start with the shifted Poisson DGLM with regression vector, $\mathbf{F}_{t,0}$, state vector, $\boldsymbol{\theta}_{t,0}$, and linear predictor $\mathbf{F}_{t,0}'\boldsymbol{\theta}_{t,0}$. Call this the baseline model, i.e., the DGLM with no random effects. The random effects extension generalizes to the linear predictor $\mathbf{F}_{t,0}'\boldsymbol{\theta}_{t,0} + \zeta_t$ where the ζ_t are time t-specific, independent, zero-mean random effects. This is trivially implemented as an extended DGLM. That is, redefine the state vector as $\boldsymbol{\theta}_t = (\zeta_t, \boldsymbol{\theta}_{t,0}')'$ and the corresponding dynamic regression vector as $\mathbf{F}_t = (1, \mathbf{F}_{t,0}')'$, so that the new model has $\log(\mu_t) = \mathbf{F}_t'\boldsymbol{\theta}_t = \mathbf{F}_{t,0}'\boldsymbol{\theta}_{t,0} + \zeta_t$ as required. This defines a different, more general DGLM that admits time t individual and unpredictable variation over and above the baseline. The state evolution equation will be modified to add a first row and column to the state evolution matrix with zero elements representing lack of dependence of random effects time-to-time as well as independence of other state vector elements. It remains to specify levels of expected contributions of random effects, and this is done using a random effects discount factor ρ, building on the standard use of discount factors for DGLM evolution variance matrices as a routine (Section 2.4 here; see also West and Harrison, 1997 chapter 6). In particular, at each time $t-1$ suppose that prior uncertainty about the core state vector elements $\boldsymbol{\theta}_{t,0}$ at time t is reflected in the prior variance matrix $\mathbf{R}_{t,0}$, so that the uncertainty about the baseline linear predictor is represented by $q_{t,0} \equiv \mathrm{V}[\mathbf{F}_{t,0}'\boldsymbol{\theta}_{t,0}|\mathcal{D}_{t-1}, \mathcal{I}_{t-1}] = \mathbf{F}_{t,0}'\mathbf{R}_{t,0}\mathbf{F}_{t,0}$. Then, a random effects discount factor $\rho \in (0,1]$ defines the conditional variance of ζ_t by $v_t \equiv \mathrm{V}[\zeta_t|\mathcal{D}_{t-1}, \mathcal{I}_{t-1}] = q_{t,0}(1-\rho)/\rho$. If ρ is set to one, then this model is simply the Poisson DGLM without the random effect. As ρ gets closer to zero, the variance of the random effect increases. Here ρ becomes a model hyper-parameter to be explored along with others. For $\rho < 1$, the additional variance injected into the time t prior is now relative to the variance of the underlying baseline models, so we have access to interpretation of ρ as defining a relative or percentage contribution to predictive uncertainty, as with standard discounting in state-space models. The impact is seen in increased dispersion of forecast distributions; some aspects of this on increased variance of the predictive negative binomials for future non-zero counts are highlighted in further technical details in Section 3.2.1.

3.2.1 Discount Factor Specifications for Random Effect

We use the random effects extension of Section 3.2 for the shifted Poisson case in evaluating forecasts of non-zero count series. As detailed in that section, this is enabled by extension of the state vector to include time-specific zero mean elements defining these random effects. Practically, this is defined using a random effects discount factor $\rho \in (0, 1]$ whose net impact on the DGLM analysis summarized in Section 2.2 is simply to inflate the prior variance of the linear predictor $\lambda_t = \log(\mu_t)$. That is, $q_t = \mathrm{V}[\lambda_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1}]$ in Section 2.2 part 4 is modified to $q_t + v_t$ where $v_t = q_t(1 - \rho)/\rho$, resulting in q_t/ρ. As with the standard discounting on state vectors above, evaluation of forecast metrics on training data using different choices of ρ is a basic strategy for choosing values, and these will be specific to each time series.

More theoretical insight can be gained by considering the impact on the implied 1−step forecast distributions. In the standard DGLM with no random effects, recall that the conditional prior for the Poisson mean μ_t is $Ga(\alpha_t, \beta_t)$ where the parameters are chosen to be consistent with the prior mean f_t and variance q_t of $\lambda_t = \log(\mu_t)$, namely $f_t = \psi(\alpha_t) - \log(\beta_t)$ and $q_t = \psi'(\alpha_t)$. Suppose we have a relatively precise prior– with q_t modestly high– so that the approximations $\psi(\alpha_t) \approx \log(\alpha_t)$ and $\psi'(\alpha_t) \approx 1/\alpha_t$ are valid (these are in fact very accurate approximations in many applications). Then $f_t \approx \log(\alpha_t/\beta_t)$ and $q_t \approx 1/\alpha_t$, resulting in $\alpha_t = 1/q_t$ and $\beta_t = \alpha_t \exp(-f_t)$. The implied 1−step forecast distribution for y_t is negative binomial with mean $\alpha_t/\beta_t = \exp(f_t)$ and variance $\alpha_t/\beta_t + \alpha_t/\beta_t^2 = \exp(f_t)(1 + \exp(f_t)q_t)$.

3.2.2 Discount Factor Effect on Forecast Variance

Now consider the impact of the random effects model extension. As noted above, the practical impact of the discount factor ρ is that q_t is inflated to q_t/ρ. The resulting negative binomial forecast distribution then has the same mean $\exp(f_t)$– not impacted at all by ρ– but now has variance $\exp(f_t)(1 + \exp(f_t)q_t/\rho)$. This has the same base component $\exp(f_t)$ (the "Poisson" component) but the second term (considered the "extra-Poisson" variation in the negative binomial) increases by a factor of $1/\rho$. Note that the impact of the random effects extension is then to increase forecast variances more at higher levels of the series (higher f_t), consistent with the aim of improving forecasts for infrequent higher events.

3.3 Forecasting in DCMMs

The compositional nature of the DCMM yields access to a full predictive distribution at a future time point that is a mixture of the forecast distributions implied for each of the independent Bernoulli and shifted Poisson DGLMs. Forecasting $k-$steps ahead from time t, the forward evolution of DGLM state vectors over times $t+1{:}t+k$ and variational Bayes' constraint to conjugate forms for implied Bernoulli and Poisson parameters yields analytic tractability and trivial computation. That is, at time t, the $k-$step ahead forecast distribution has a p.d.f. of the compositional form

$$p(y_{t+k} \mid \mathcal{D}_t, \mathcal{I}_t, \pi_{t+k}) = (1-\pi_{t+k})\delta_0(y_{t+k}) + \pi_{t+k} h_{t,t+k}(y_{t+k})$$

where:

- $(\pi_{t+k} \mid \mathcal{D}_t, \mathcal{I}_t) \sim Be(\alpha_t^0(k), \beta_t^0(k))$ and $\delta_0(y)$ is the Dirac delta function at zero.
- $h_{t,t+k}(y_{t+k})$ is the density function of $y_{t+k} = 1 + x_{t+k}$ where x_{t+k} has the negative binomial distribution

$$(x_{t+k} \mid \mathcal{D}_t, \mathcal{I}_t) \sim Nb\Big(\alpha_t^+(k), \frac{\beta_0^+(k)}{1+\beta_t^+(k)}\Big).$$

- The defining parameters $\alpha_t^0(k)$, $\beta_t^0(k)$, $\alpha_t^+(k)$, $\beta_t^+(k)$ are computed from the binary and positive count DGLMs, respectively.

That is, the mixture places probability $1-\pi_{t+k}$ on $y_{t+k}=0$, and probability π_{t+k} on the implied shifted negative binomial distribution.

Depending on the forecasting context, we may be interested in marginal or joint forecasting. The above details define the relevant marginal forecast distributions at any time. However, we are generally much more interested joint forecasts for $y_{t+1}, \ldots, y_{t+k}$ and, as discussed in Section 3.1, in forecasting paths with an opportunity to explore dependencies in outcomes over time as well as to predict functions of them. This is trivially– and computationally efficiently– enabled using forward simulation, as follows.

- At time t, propagate the posterior $(\boldsymbol{\theta}_t \mid \mathcal{D}_t)$ to the implied $1-$step ahead prior $(\boldsymbol{\theta}_{t+1} \mid \mathcal{D}_t, \mathcal{I}_t)$.
- Apply the variational Bayes' constraint to the implied conjugate prior on η_{t+1}, so that the implied $1-$step ahead forecast distribution is of the form given in part 5 of Section 2.2; that is, Beta-Bernoulli or negative binomial depending on the chosen DGLM.
- Simulate an outcome y^*_{t+1} from this $1-$step forecast distribution.
- Treating this synthetic outcome as "data", perform the time $t+1$ update step to revise the prior to posterior for the state vector, now based on modified information $\mathcal{D}_{t+1}$ in which the unknown y_{t+1} is substituted by its synthetic value y^*_{t+1}.
- Evolve to time $t+2$ and repeat the process to simulate a synthetic y^*_{t+2} and then update based the model.
- Continue this process over lead-times $t+3, \ldots, t+k$ to the chosen forecast horizon.

This results in one synthetic path $y^*_{t+1}, \ldots, y^*_{t+k}$ defining a single Monte Carlo sample from the joint distribution of $y_{t+1}, \ldots, y_{t+k}$ conditional on $\{\mathcal{D}_t, \mathcal{I}_t\}$. Repeat this process many times to generate a large Monte Carlo sample from this joint predictive distribution. At each marginal time point $t+j$, the corresponding samples give a

Monte Carlo representation of the predictive distribution at that lead-time, while the full sample provides the required opportunities for inference on paths, dependencies of outcomes between time points, and functions (cumulative outcomes, exceedance over some specific level, etc) of the path into the future.

3.4 Aspects of Product Sales Forecasting with DCMMs

3.4.1 Context and Data

Our case study concerns multi-step forecasting of many individual items in each of a large number of US supermarkets (or stores). An item is defined as a unique stock keeping unit (SKU) and sales forecasts for the following 1:14 days are updated daily. For our examples here, we extract a very small number of items at one chosen store to provide illustrations and insights into the utility of DCMMs. The selected data set records daily sales of 179 SKUs in one particular store over the 2,192 days from July 1st 2009 to July 1st 2015. Each of these products is in the pasta category, in one of 14 subtypes of pasta. By percentage of unit sales, the primary pasta types are spaghetti (25%), macaroni (13%), wholewheat (10%), and penne (9%). The products comprise 20 brands, and the majority of unit sales (44%) come from the supermarket's in-house brand. Additional information includes the price paid per transaction, and whether or not each SKU was on promotion on the date of the transaction. In the contexts of analyses to follow, the data for several items are summarized in Table 3.1 and shown in Figures 3.1 and 4.2; these provide some insights into heterogeneity of daily sales patterns.

Table 3.1: Some summaries of daily pasta sales data by item

Item	Mean	Median	Variance	% 0 sales
A	1.0	0	1.8	51.8
B	9.9	9	29.4	1.3
C	4.7	4	15.6	5.9
D	3.4	2	10.6	14.4

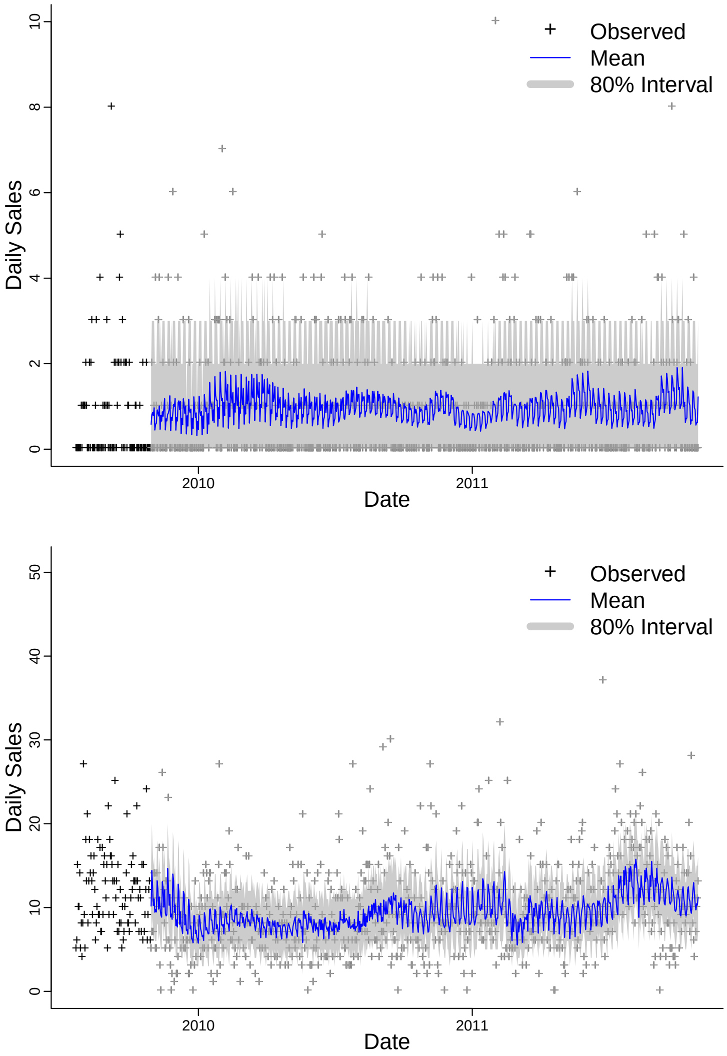


FIGURE 3.1: Data and aspects of 1−step ahead forecast distributions for items A (upper) and B (lower). Shading: 80% predictive credible intervals; full line: predictive mean.

3.4.2 Example Univariate DCMMs

As example series, data in Figure 3.1 show daily sales of two selected spaghetti items in one store over the two year period from July 2009 to October 2011. Item A has relatively low daily sales with a mean of 1.0, a median of 0, and zero sales occurring on 52% of days. Based on the prevalence of days with zero sales, the binary component of the DCMM will play an important role in forecasting item A. Item B is a high selling spaghetti product in this store, with a median of 9 sales per day and zero sales occurring on only 1.3% of days. The daily sales of item B appear to have high variance, and on some days we see sales spike to above 30 units.

The same form of DCMM is applied to each item. Each of the Bernoulli and shifted Poisson components has a local level, a regression component with log price as predictor, and a full Fourier seasonal component of period 7 to reflect the day-of-week effects. All components are dynamic, allowing for variation over time in local level, regression coefficient, and Fourier coefficients. Thus each of the binary and Poisson DGLMs have regression vector and evolution matrix defined by

$$\mathbf{F}_t' = \big(1,\, \log(\text{price}_t),\, 1,0,\, 1,0,\, 1,0\big) \qquad \text{and} \qquad \mathbf{G}_t = \text{blockdiag}[1,\, 1,\, \mathbf{H}_1,\, \mathbf{H}_2,\, \mathbf{H}_3]$$

with

$$\mathbf{H}_j = \begin{pmatrix} \cos(2\pi j/7) & \sin(2\pi j/7) \\ -\sin(2\pi j/7) & \cos(2\pi j/7) \end{pmatrix}, \quad j = 1{:}3.$$

Exploratory analysis of an initial three weeks of data was used to specify prior moments on the state vector at $t = 1$ representing day 22 of the full data set. For the Poisson component, this used standard reference Bayesian analysis assuming no time variation in parameters to define "ballpark" initial priors. For the Bernoulli component, we specify the prior mean of the level to be logit(p), where p is the proportion of the first 21 days on which positive sales occurred. Thus, p represents an empirical estimate of π_t, the probability of non-zero sales occurring on day t. All other prior means are set to zero, and the prior variance matrix set as the identity. Fixed discount factors of 0.999 (Bernoulli), 0.99 (item B Poisson), and 0.995 (item A Poisson) are used on each of the level, regression and seasonal components; these values were chosen based on the results of previous analyses of daily sales data. The lower discount factor for item B reflects the need for more adaptability to the non-stationarities present in Figure 3.1. The DCMM analyses were run through the first

100 days of data before forecasting. Starting at time $t = 101$, full forecast distributions for 1:14 days ahead were computed at each day, and updated recursively over the next two years. These define Monte Carlo samples of size 5,000 of synthetic future sales over rolling 14 day periods. Some illustration of these forecasts appears in Figure 3.1.

3.4.3 Point Forecasts and Metrics

Section 2.6 discusses relevant aspects of forecast evaluation for count data within consumer sales forecasting. Based on this discussion, our main focus in this application is on evaluating the entire forecast distribution rather than specific choices of loss functions. However, to connect with common practice and recent literature, we explore various point forecast metrics in Section 4.5. First, we broaden perspective to evaluation of the full forecast distributions using the practically relevant issues of coverage and calibration of predictive distributions.

3.4.4 Probabilistic Forecast Evaluation

Figure 3.1 gives the overall impression that the DCMM forecasts relatively well in the short-term for both items A and B, clearly picking-up the seasonal patterns and responding to changes over time. Then, the infrequent higher sales levels are better forecast for item A than apparently for item B, the latter being a higher-selling item. To explore forecast performance in more detail, Figures 3.2, 3.3 and 3.4 summarize aspects of the full predictive distributions generated by the DCMMs for item A (left frames) and B (right frames).

Figure 3.2 displays coverage of forecast distributions for each of 1, 7, and 14$-$days ahead. These graph the empirical coverage obtained over the full year of forecasting for predictive credible (highest predictive density– HPD) intervals in each case. An ideal model would lead to coverage plots close to the 45$-$degree line. For item A,

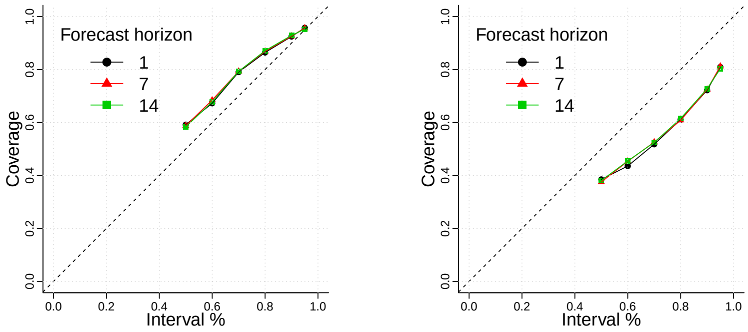


FIGURE 3.2: Coverage plots for items A (left) and B (right) from 1−, 7− and 14−day ahead forecasts.

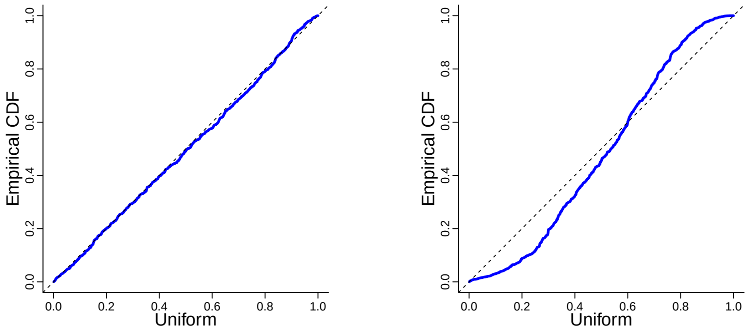


FIGURE 3.3: Randomized PIT plots from 1−day ahead forecasting of items A (left) and B (right). Full line: ordered randomized PIT values; dashed: 45−degree line.

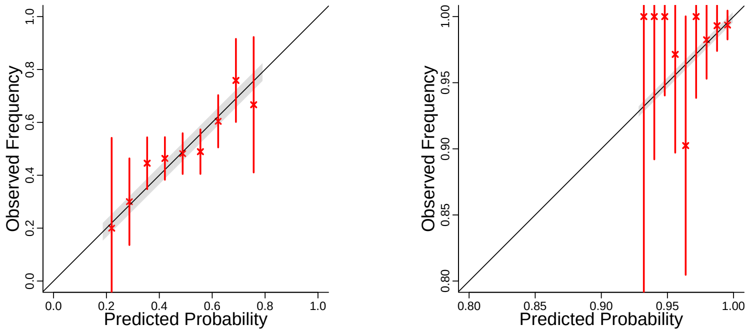


FIGURE 3.4: Binary calibration plots from 1−day ahead forecasting of non-zero sales of items A (left) and B (right). Crosses mark observed frequencies in each bin, horizontal grey shading indicates variation of forecasts in each bin, and vertical bars indicate binomial variation based on the number of days in each bin.

forecast distributions have slight over-coverage. For example, empirical coverage of the $1-$day ahead 50% predictive intervals is about 57%. In contrast, for item B we see evidence of under-coverage at all horizons, related mainly to the apparent inability of the model to adequately forecast the infrequent higher sales values.

A second probabilistic evaluation uses the probabilistic integral transform (PIT), i.e., a general residual plot based on the predictive c.d.f. for each outcome. Since predictive distributions are discrete, this involves the randomized PIT (Kolassa, 2016). If sales counts y are forecast with predictive c.d.f. $P(\cdot)$, define $P(-1) = 0$ and draw a random quantity $p_y \sim U(P(y-1), P(y))$ given the observed value of y. Over the sequence of repeat forecasting events an ideal model would generate realizations p_y that are approximately uniform. For each item, Figure 3.3 plots the ordered randomized PIT values from the $1-$day ahead forecasts distributions versus uniform quantiles. The concordance of the outcomes with uniformity is apparently strong for item A. For item B, however, we see significant non-uniformity and a shape consistent with forecast distributions that are just too light-tailed; that is, the outcome sales data on item B exhibit higher levels of variation than the DCMM predictive distributions capture.

The third probabilistic evaluation focuses on frequency calibration properties of forecasts of zero/non-zero sales. Ideal calibration means that, of the days non-zero sales are forecast with probability near p, approximately $100p\%$ actually have non-zero sales. In practice, we bin the probability scale according to variability in forecast probabilities of non-zero sales across the year, and evaluate the realized frequency of non-zero sales on the days within each bin. Figure 3.4 displays the results for $1-$step ahead binary predictions. For item A, the predicted probabilities of non-zero sales range in 30:80%; these are allocated into ten bins of equal width. The figure displays the observed frequency of non-zero sales within each bin and an approximate 95% binomial confidence interval based on the number of days within each bin. Horizontal shading displays the width of the predicted probability bins. Ideal predictions would lead the observed frequency in each bin to fall within the shaded area, while the vertical bars indicate limits based on sample size in each bin. For item A, the performance is apparently very good indeed. For item B, a relatively high-selling item, predicted probabilities of non-zero daily sales range in 90:100% over time. As with item A, the vertical calibration intervals intersect the 45−degree line and the horizontal shading, indicating that there are no obvious issues with forecasting the zero/non-zero outcome. In terms of the full DCMM, the binary component is less important for item B than for item A. Then, as noted above, the under-dispersion of forecasts of item B is a clear negative; this is addressed with the random effects extension below.

3.4.5 Random Effects Extensions

We illustrate the potential of the random effects extension of DCMMs introduced in Section 3.2 to adapt to the over-dispersion issues in the basic analysis of item B in the last section. The main details of the analysis remain the same, but now the model is modified to include day-specific random effects. The summarized analysis uses a random effects discount factor $\rho = 0.2$, chosen following exploration of the impact of different values over the first 100 days of data. We considered values between 0.1 and 1, and selected ρ based on a combination of metrics including the 1-step negative binomial predictive log-likelihood, rPIT uniformity, and coverage of 1-step predictions in the Poisson DGLM.

Figure 3.5 displays the updated 1−step forecast means and 80% credible intervals over time, to be compared to Figure 3.1. Note the wider forecast intervals that are to be expected. Figure 3.6 displays the resulting coverage of the 1, 7, and 14−day ahead forecast credible intervals over time, and the calibration plot of the randomized PIT values from 1−day ahead forecasts. Coverage has increased and substantially improved to conform with the 45−degree line, while PIT values are in much closer concordance with uniformity. Overall, the addition of the random effect has accounted for some of the under-dispersion of forecast distributions in the baseline DCMM, and these aspects of forecast evaluation indicate clear and practical improvements as well as overall accuracy.

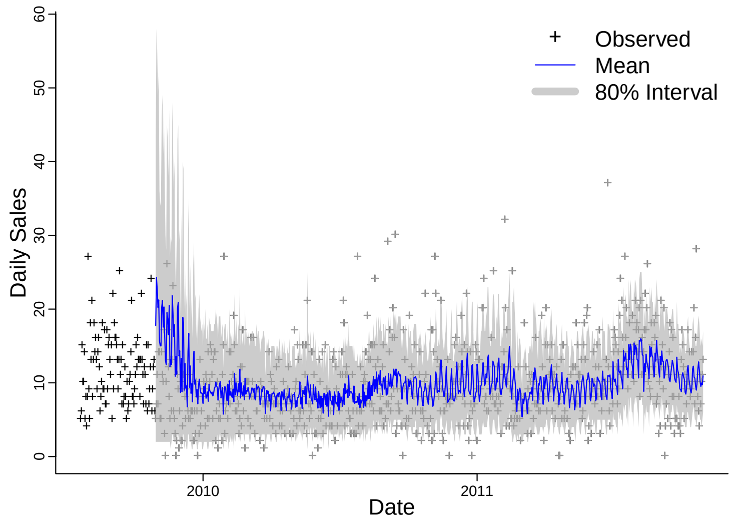


FIGURE 3.5: Daily sales of item B, and the 1−day ahead forecast from the DCMM with random effects. The blue line indicates the forecast mean, and the gray shading indicates the forecast distribution 80% credible interval.

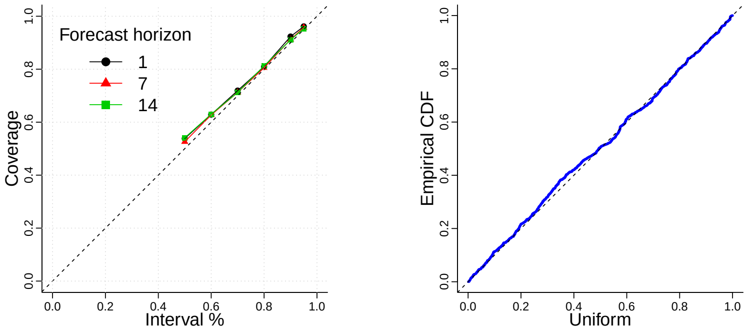


FIGURE 3.6: Empirical coverage plot (left) and randomized PIT plot (right) for 1−day ahead forecasting of item B using the DCMM with random effects extension. Compare with the results under the basic DCMM in the right-hand frames of Figures 3.2 and 3.3.

4

Multi-scale DCMM

4.1 Decouple/Recouple and Shared Features Across Series

We now turn to recoupling sets of univariate DCMMs in contexts where there may be opportunity to improve multi-step forecasts via more accurate assessment of effects and patterns shared across the series. That is, we aim to exploit traditional ideas of hierarchical random effects models where common features over time are "seen differently" by each univariate series, and where both series-specific effects and noise obscure the patterns at the individual series level. This is particularly relevant in sales and demand forecasting when items are sporadic, i.e., in our DCMMs in cases of non-negligible zero sales and otherwise low count levels. Products can often be grouped hierarchically based on characteristics like product family, brand, and store location. Sales and demand patterns within such groups may have similar trends or seasonal patterns due to external factors such as marketing, economy, weather, and so forth. Our main example here focuses on the daily seasonal pattern over each week, a pattern that is heavily driven by customer traffic through the stores related to weekly behavioral factors. This general pattern is naturally shared by many individual series, but at low levels of sales it is substantially obscured by inherent noise so that using information from other series through a multivariate approach is of interest. If products are grouped, then we may expect estimates of seasonality at the aggregate level to be more accurate and less noisy than at the individual level. Using aggregated seasonality in a model rather than individual item seasonality may then improve forecasting for individual items. On the other hand, seasonality exhibited by items that sell at higher levels may be much more evident and the potential to gain forecast accuracy there less obvious. Part of our interest is to explore these potential gains across a range of items and demand levels.

Our desiderata include maintaining flexibility in customizing DCMMs to individual time series along with the ability to run fast, sequential Bayesian analysis of inherently decoupled univariate analyses of many series. In product demand forecasting, retailers are generally interested in many thousands of items simultaneously. Due to time and computational constraints, the conventional approach is to rely on univariate methods which forecast independently across SKUs, and, therefore, this is a central consideration. We avoid large-scale and complex multivariate modeling that would otherwise necessitate the use of intense MCMC (or other) computations that would obviate the use of efficient sequential analysis while most seriously limiting the ability to scale in the number of time series. To do this, we maximally exploit the univariate framework of Section 6.1 with series decoupled conditional on factors shared in common across series, and then we recouple by utilizing a separate model to analyze and forecast these common factors.

4.1.1 Examples of Learning Shared Feature

In this section, we present an example based on West and Harrison, 1997, example 16.1. Consider a company selling N products, where y_i is the daily sales of the ith product, $i = 1, \ldots, N$. Suppose y_i can be expressed as

$$y_i = f_i + \beta_i \phi + \epsilon_i, \quad \epsilon_i \sim N(0, v_\epsilon)$$

where f_i is the known expected value of y_i, and ϕ and ϵ_i are zero-mean, uncorrelated random variables. Here, ϕ represents a common latent factor affecting each item y_i, and ϵ_i represents item-specific variation. Each item can "adjust" the shared latent factor ϕ through the item-specific coefficient β_i. Our interest in this example is in inferring the latent factor ϕ, and whether conditioning on aggregate data can lead to more precise estimates of ϕ. We set a prior of $\phi \sim N(0, v_\phi)$. Let $T = \sum_i y_i$ be the total sales of all products. We can write T as

$$T = \sum_{i=1}^{N} (f_i + \beta_i \phi + \epsilon_i) = f + \beta \phi + \epsilon,$$

where

$$\epsilon = \sum_{i=1}^{N} \epsilon_i, \quad \beta = \sum_{i=1}^{N} \beta_i, \quad f = \sum_{i=1}^{N} f_i.$$

The joint distribution of (y_i, ϕ) is

$$\begin{pmatrix} y_i \\ \phi \end{pmatrix} \sim N\left(\begin{pmatrix} f_i \\ 0 \end{pmatrix}, \begin{pmatrix} \beta_i^2 v_\phi + v_\epsilon & \beta_i v_\phi \\ \beta_i v_\phi & v_\phi \end{pmatrix} \right).$$

Conditional on y_i, the posterior distribution of ϕ becomes

$$\phi \mid y_i \sim N(m_{\phi|y}, v_{\phi|y})$$

where

$$m_{\phi|y} = \frac{\beta_i v_\phi}{\beta_i^2 v_\phi + v_\epsilon}(y_i - f_i) \quad \text{and} \quad v_{\phi|y} = v_\phi - \frac{\beta_i v_\phi^2}{\beta_i^2 v_\phi + v_\epsilon}.$$

The joint distribution of (T, ϕ) is

$$\begin{pmatrix} T \\ \phi \end{pmatrix} \sim N\left(\begin{pmatrix} f \\ 0 \end{pmatrix}, \begin{pmatrix} \beta^2 v_\phi + N v_\epsilon & \beta v_\phi \\ \beta v_\phi & v_\phi \end{pmatrix} \right).$$

Conditional on T, the posterior for ϕ is

$$\phi \mid T \sim N(m_{\phi|T}, v_{\phi|T})$$

where

$$m_{\phi|T} = \frac{\beta v_\phi}{\beta^2 v_\phi + N v_\epsilon}(T - f) \quad \text{and} \quad v_{\phi|T} = v_\phi - \frac{\beta^2 v_\phi^2}{\beta^2 v_\phi + N v_\epsilon}.$$

The values of $v_{\phi|y}$ and $v_{\phi|T}$ summarize posterior uncertainty about ϕ conditional on one series, y_i, and the aggregate series T. For simplicity, if we assume that each $\beta_i = 1$, then these terms simplify to:

$$v_{\phi|y} = v_\phi \left(1 - \frac{v_\phi}{v_\phi + v_\epsilon}\right),$$

$$v_{\phi|T} = v_\phi \left(1 - \frac{N v_\phi}{N v_\phi + v_\epsilon}\right).$$

If we assume that $N = 1{,}000$, $v_\phi = 1$, and $v_\epsilon = 99$, then these terms become

$$v_{\phi|y} = 0.99 \quad \text{and} \quad v_{\phi|T} = 0.09.$$

Under these assumptions, the posterior uncertainty about ϕ is much lower if we condition on the total sales T rather than individual sales y_i. This supports the hypothesis that learning latent factors in aggregate models can lead to more precise estimates.

The previous example studies the posterior variance of a common factor that acts additively on an item's sales. However, in some models, such as a Poisson DGLM, a common factor ϕ may act additively on the log mean, so the common factor $\theta = e^\phi$ acts multiplicatively on the mean. Here we discuss an additional example for a multiplicative common factor. Again, assume that a company sells N products, and that we can represent the item-level sales y_i as

$$y_i \sim Po(\mu_t) \quad \text{where} \quad \log(\mu_i) = m + \phi,$$

and where the log Poisson mean is the linear predictor, m is the known mean of all items, and ϕ is a latent factor common to all items. This model implies that

$$\mu_i = e^m e^\phi = e^m \theta$$

where θ is a latent factor that acts multiplicatively on the mean μ_i. Again, our goal is on inferring the latent factor θ. We assume a conjugate prior on the latent factor, $\theta \sim Ga(a, b)$. Since $\mu_i = e^m \theta$, this prior induces a conjugate prior on μ_i such that $\mu_i \sim Ga(a, b/e^m)$. If we observe y_i, computing the posterior of μ_i involves simple conjugate analysis and thus

$$(\mu_i \mid y_i) \sim Ga(a + y_i, \frac{b}{e^m} + 1)$$

$$\Rightarrow (\theta \mid y_i) \sim Ga(a + y_i, b + e^m).$$

Now, let $T = \sum_i y_i$ be the total sales of all products. Due to the additivity of independent Poissons, we can represent T as $T \sim Po(\mu)$ with

$$\mu = \sum_i \mu_i = N e^m e^\phi = N e^m \theta$$

where

$$\log \mu = \log N + m + \phi,$$

so the common factor θ again acts multiplicatively on the aggregated Poisson mean μ. Now, our prior on θ again implies a prior on μ such that $\mu \sim Ga(a, b/(Ne^m))$. Conditional on T, the conjugate posterior for μ and the implied posterior on θ are

$$(\mu \mid T) \sim Ga(a + T, \frac{b}{Ne^m} + 1)$$

$$\Rightarrow (\theta \mid T) \sim Ga(a + T, b + ne^m).$$

The values of $\mathrm{V}(\theta \mid y_i)$ and $\mathrm{V}(\theta \mid T)$ summarize posterior uncertainty about multiplicative latent factor θ conditional on the sales of a single series versus the total sales of all products. These variances are

$$\mathrm{V}(\theta \mid y_i) = \frac{a + y_i}{(b + e^m)^2}$$

and

$$\mathrm{V}(\theta \mid T) = \frac{a + T}{(b + Ne^m)^2}.$$

If we assume a prior with $a = b = 0.1$, and values $N = 100$, $m = 0.5$, and that $y_i = 1$ and $T = 50$, then these values become

$$\mathrm{V}(\theta \mid y_i) = 0.36 \quad \text{and} \quad \mathrm{V}(\theta \mid T) = 0.002.$$

In most cases, the posterior uncertainty about θ will be lower when we condition on T rather than y_i. These two examples support the idea that if we believe there is a common factor acting multiplicatively on N items, it is worthwhile to learn this common factor using an aggregate model rather than individually at the item level.

4.2 Multi-Scale Models and Top-Down Recoupling

The essential structure is that of a set of decoupled dynamic latent factor models as relates to traditional Bayesian multivariate approaches for conditionally normal data (e.g. Aguilar et al., 1999; Aguilar and West, 2000; Carvalho and West, 2007), but with the major novelty of inferring latent factors externally. The latter draws conceptually on prior work in multi-scale models in time series (e.g. Ferreira et al., 2003, 2006) and on "bottom-up/top-down" ideas in Bayesian forecasting (West and Harrison, 1997, section 16.3). The basic idea of top-down forecasting is to forecasting patterns at a highly aggregate level and then somehow disaggregate down to the individual series. Practical methods using point forecasts of aggregate sales over a set of items typically use some kind of weighting to disaggregate. While our framework is general, we focus on the key example of seasonal patterns, where our work contributes to an existing literature. For example, the method of group seasonal indices (GSI: Withycombe, 1989; Chen and Boylan, 2007) aggregates series in a defined group and then estimates seasonal factors from the aggregate series. We follow this concept, but with an holistic class of Bayesian state-space models– with time-varying patterns and full probabilistic specifications– rather than direct data adjustments assuming constant seasonal patterns. The full Bayesian approach was rooted in conditional linear, normal examples in (West and Harrison, 1997, section 16.3), and the developments here extend that to a full class of multivariate DCMMs for count series. Importantly, we use full simulation of posterior distributions of common factor patterns to send to the set of univariate DCMMs, so that all relevant uncertainties are quantified and reflected in the resulting forecast distributions of the univariate series.

4.3 Model Structure and Notation

A set of N series $y_{i,t}$, $(i = 1{:}N)$, follow individual DCMMs sharing some common factors of interest. Denote these individual models by $\mathcal{M}_i$ for $i = 1{:}N$. In each of the binary and shifted Poisson DGLM components, $\mathcal{M}_i$ has both series-specific and common state-space components with latent factor predictors shared across the N series. In $\mathcal{M}_i$, the time t state and regression vectors for each DGLM component have partitioned forms with individual and common predictors. For example, the shifted Poisson component has state and regression vectors defined by

$$\mathcal{M}_i : \qquad \boldsymbol{\theta}_{i,t} = \begin{pmatrix} \boldsymbol{\gamma}_{i,t} \\ \boldsymbol{\beta}_{i,t} \end{pmatrix}, \quad \mathbf{F}^{+}_{i,t} = \begin{pmatrix} \mathbf{f}_{i,t} \\ \boldsymbol{\phi}_t \end{pmatrix}, \qquad i = 1{:}N, \tag{4.1}$$

with subvectors of conformable dimensions; the linear predictor is then

$$\lambda_{i,t} = \boldsymbol{\gamma}_{i,t}'\mathbf{f}_{i,t} + \boldsymbol{\beta}_{i,t}'\boldsymbol{\phi}_t.$$

Here $\mathbf{f}_{i,t}$ contains constants and series-specific predictors– such as item-specific prices and promotions in the sales forecasting context. The latent factor vector $\boldsymbol{\phi}_t$ is common to all series– such as seasonal or brand effects in the sales forecasting context. Each series has its own state component $\boldsymbol{\beta}_{i,t}$ so that the impacts of common factors are series-specific as well as time-varying.

In parallel, a separate model of some form– in our case, a DLM– also depends on $\boldsymbol{\phi}_t$ and possibly other factors. Denote this model by $\mathcal{M}_0$. Forward sequential analysis of data relevant to $\mathcal{M}_0$ defines posterior distributions for $\boldsymbol{\phi}_t$ at any time t that can be used to infer and forecast the $\boldsymbol{\phi}_t$ process as desired. These inferences on the common factors are then forwarded to each model $\mathcal{M}_i$ to use in forecasting the individual series.

Generally, in standard updating and forecasting of Bayesian state space models, the regression vector $\mathbf{F}_t$ is known at each time point t. However, in this multi-scale context, our knowledge about $\boldsymbol{\phi}_t$ is summarized at each time point by a posterior distribution from an external model. Therefore, we propose a procedure for forecasting and updating the DCMM which accounts for the fact that there is now uncertainty about $\mathbf{F}_t$. The following steps, based on simulating from the posterior of $\boldsymbol{\phi}_t$, enable forecasting and updating while accounting for all uncertainty about $\boldsymbol{\phi}_t$.

4.4 Summary Analysis

Consider now multi-step forecasting and then $1-$step updates and evolution from times $t-1$ to t. Extend the notation for information sets $\mathcal{D}_t, \mathcal{I}_t$ to be model specific with model indices $i = 0{:}N$.

Forecasting: This is a direct and full probabilistic extension– enabled by simulation– of the simple theoretical examples of "conditioning on external forecast information" in conditionally linear, normal models in examples of West and Harrison, 1997, (section 16.3).

1. At $t-1$ we have conditionally independent prior summaries for the series-specific states i, namely $(\boldsymbol{\theta}_{i,t}|\mathcal{D}_{i,t-1}, \mathcal{I}_{i,t-1})$ for each $i = 1{:}N$, having evolved independently from the $\boldsymbol{\theta}_{i,t-1}$.
2. Independently, model $\mathcal{M}_0$ simulates the trajectory of the latent factor process into the future, i.e., generates independent samples $\boldsymbol{\phi}^s_{t:t+k} \sim (\boldsymbol{\phi}_{t:t+k}|\mathcal{D}_{0,t-1}, \mathcal{I}_{0,t-1})$ at time $t-1$ and for any $k \geqslant 0$, where $s = 1{:}S$ indexes S Monte Carlo samples.
3. Send these synthetic latent factors to each individual model. In $\mathcal{M}_i$, use S parallel DCMM analyses to forecast over times $t{:}t+k$. Each analysis conditions on one sampled $\boldsymbol{\phi}^s_{t:t+k}$. One Monte Carlo draw from the implied predictive distribution of $y_{i,t:t+k}|\boldsymbol{\phi}^s_{t:t+k}, \mathcal{D}_{i,t-1}, \mathcal{I}_{i,t-1}$ yields a sampled trajectory $y^s_{i,t:t+k}$, so we create a Monte Carlo sample of size S accounting for the inferences on, and uncertainty about, the latent factor process as defined under $\mathcal{M}_0$.

The last step builds on the fact that the DCMM analysis detailed earlier applies to each series– independently across series– conditional on a value of $\boldsymbol{\phi}_t$. Note also the use of parallelization.

Updating: Observing the $y_{i,t}$ we now update in each model $\mathcal{M}_i$ separately.

4. For each s, compute the value of the 1$-$step ahead predictive p.d.f $p(y_{i,t}|\boldsymbol{\phi}^s_t, \mathcal{D}_{i,t-1}, \mathcal{I}_{i,t-1})$ from the conditionally conjugate DGLM analysis. Use these as marginal likelihoods to evaluate implied posterior probabilities over the $s = 1{:}S$ latent factor values relative to uniform $(1/S)$ prior probabilities.
5. Apply the standard DGLM updating to compute Monte Carlo sample $s-$specific posterior mean vectors and variance matrices for $\boldsymbol{\theta}_{i,t}|\boldsymbol{\phi}^s_t, y_{i,t}, \mathcal{D}_{i,t-1}, \mathcal{I}_{i,t-1}$.

Marginalize over the $\boldsymbol{\phi}_t^s$ with respect to the probabilities from 4 above to deduce implied Monte Carlo approximations to the posterior mean vector and variance matrix of $\boldsymbol{\theta}_{i,t}|\mathcal{D}_{i,t}$, where $\mathcal{D}_{i,t}$ now implicitly includes information from $\mathcal{D}_0$ as well as $\{y_{i,t}, \mathcal{D}_{i,t-1}, \mathcal{I}_{i,t-1}\}$.

Evolution to time $t+1$ now completes the cycle.

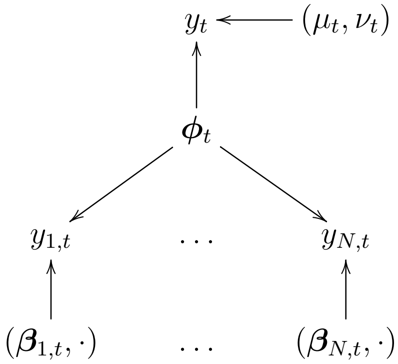


FIGURE 4.1: Directed graph representing multiscale framework

Figure 4.1 shows a representation of the multi-scale framework as a directed acyclic graph (DAG). In this DAG, the latent factor $\boldsymbol{\phi}_t$ affects each individual series $y_{i,t}$ as well as a separate series y_t. For this DAG, we assume that the model $\mathcal{M}_0$ is a DLM fit to data y_t. One example of a relevant y_t would be a series of aggregated daily sales, such as $y_t = \sum_{i=1}^N y_{i,t}$. Beyond $\boldsymbol{\phi}_t$, each $y_{i,t}$ also depends on coefficient $\boldsymbol{\beta}_{i,t}$ and other factors. In this representation, y_t also depends on μ_t and ν_t, the observation mean and variance in a DLM. A key assumption in the multi-scale framework is that, conditional on all data up to time $t-1$, all of the information about $\boldsymbol{\phi}_t$ is contained in y_t. We are adopting the view that—although there is information about $\boldsymbol{\phi}_t$ in each $y_{i,t}$—it is to be ignored as $\mathcal{M}_0$ already captures most of the relevant information. This point of view is supported by the conclusion of examples in Section 4.1.1 that $\boldsymbol{\phi}_t$ is more precisely estimated conditional on the total sales than a single series $y_{i,t}$. This view allows us to maintain our parallel analysis by independently updating each

$\mathcal{M}_i$ conditional on $\boldsymbol{\phi}_t$ simulated from $\mathcal{M}_0$.

The model provides opportunity to explore differences across series, and over time, in effects of latent factors. If the effects on series $y_{i,t}$ are similar to those under $\mathcal{M}_0$, then $\boldsymbol{\beta}_{i,t}$ will be close to one. Inferred trajectories of $\boldsymbol{\beta}_{i,t}$ over time will capture relevant deviations and allow comparisons across series in how strongly they relate to the latent factors. Further discussion and an applied exampled of inference on multi-scale latent factors is given in Section 4.6.

4.5 An Example in Multi-Scale Sales Forecasting

4.5.1 Context, Data and Models

Our case study involves a number of multi-scale features that offer potential for improved forecasting using the multivariate/multi-scale strategy. Items are related in product type categories, by brand, and by consumer behavior as evidenced in several ways including, simply, the day-of-week seasonality related to "store traffic". We give one example here focused on this latter feature, comparing forecasting performance of a multi-scale model– based on one specific common pattern model $\mathcal{M}_0$– to a set of individual DCMMs. In one selected store, we identify $N = 17$ spaghetti items for the example. Figure 4.2 plots the daily sales of four of these items over the period July 2009 to October 2011, giving some indication of the ranges of demand levels and patterns over time. We take models $\mathcal{M}_i$ in which the individual DCMMs of eqn. (4.1) have $\mathbf{f}'_{i,t} = (1, \log(\text{price}_{i,t}))$ and a 7-dimensional $\boldsymbol{\phi}_t$ with one non-zero element for the "current day-of-week" seasonal factor and zeros for each other element. Further, the state evolution matrix in all cases is $\mathbf{G}_t \equiv \mathbf{G} = \mathbf{I}$. We use the same DGLM specification– but with component-specific state vectors– for each of the binary and shifted Poisson components of these DCMMs.

In principle, model $\mathcal{M}_0$ could be any external model used to map and predict patterns of traffic in the store impacting spaghetti purchases. Our example uses a simple model based on aggregate sales that highlights the relevance of the term multi-scale. This could obviously be modified to incorporate additional predictors of day-of-week effects, but serves well to illustrate the analysis here. Since pasta sells at high levels, aggregate-level sales of types of pasta such as spaghetti are typically high and more clearly reveal the weekly demand patterns as they change day-to-day. So a relevant aggregate series would be a natural candidate for $\mathcal{M}_0$. We take y_t to be the log of total spaghetti sales in this store on day t as our example, and specify $\mathcal{M}_0$ as a flexible dynamic linear model incorporating time-varying effects and stochastic volatility as in Section 2.2. This DLM $\mathcal{M}_0$ has a local linear trend, a regression component with the log average spaghetti price as a predictor, the full Fourier seasonal component for the 7−day seasonal pattern over the week, and a yearly seasonal effect with the first two harmonics of period 365. Discount factors in this DLM were chosen based on previous analyses of aggregate sales, resulting in $\delta = 0.995$ for each of the trend and regression components, $\delta = 0.999$ for each of the seasonals, and $\beta = 0.999$ for the residual stochastic variance process. Note that, at each time, the full posterior for the implied "current day effect" element in $\boldsymbol{\phi}_t$ can be trivially simulated from this model as it is simply one element of a linear transformation of the Fourier coefficients to the 7−dimensional seasonal factor vector (West and Harrison, 1997, section 8.6.5). In all analyses we used an initial three weeks of training data to specify initial priors for DCMM state vectors at the time index $t = 1$ representing the start of the 830 day modeling period. Analysis was then run over the first period of 100 days, prior to then forecasting 1:14 days ahead on each day of the following two years. Predictive performance is assessed across a range of random effect discount factors $\rho \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ in each $\mathcal{M}_i$. Initial priors for the benchmark and multi-scale DCMMs are matched, adjusting for the use of the latent seasonal factor $\boldsymbol{\phi}_t$ in the latter compared to the individual Fourier models in the former. We use fixed discount factors of $\delta = 0.999$ (Bernoulli), 0.99

(item B Poisson), and 0.995 (item A, C, D Poisson) for each of the local level, price regression state elements, and the state elements corresponding to the seasonal effects (the vector of time-varying Fourier coefficients in the benchmark DCMMs, and $\boldsymbol{\phi}_t$ in the multi-scale model, respectively). The lower discount factor in the Poisson component for item B reflects the more apparent non-stationarities in sales compared to lower-selling items.

4.5.2 Forecast Metrics

We have discussed general issues of forecast evaluation and comparison in Sections 2.6, 3.4.3 and 3.4.4, stressing the applied need for global probabilistic metrics in general, and particularly in contexts of low count time series. That said, for some basic comparisons of the benchmark with multi-scale DCMMs, it is also of interest to relate to traditional point forecast accuracy measures. We do that here with three empirical loss functions from the count forecasting literature (e.g. Fildes and Goodwin, 2007), namely mean absolute deviation (MAD), mean ranked probability score (MRPS), and scaled mean squared error (sMSE) metrics. Metrics are specific to a chosen lead-time $k > 0$. For any series y_t, denote by $f_{t,k}$ a point forecast of y_{t+k} made at time t. The MAD metric is standard, simply the time average of $|y_{t+k} - f_{t,t+k}|$, and the optimal point forecast is the $k-$step ahead predictive median. The RPS metric is a scoring rule related to earlier discussed and utilized PIT measures (Snyder et al., 2012; Kolassa, 2016). If the time t forecast distribution for y_{t+k} has c.d.f. $P_{t,k}(\cdot)$, then $RPS_t(k) = \sum_{j=0}^{\infty}(P_{t,k}(j) - \mathbb{1}(y_{t+k} \leqslant j))^2$, and we calculate the MRPS for forecast horizon k by averaging $RPS_t(k)$ across all days t. The scaled squared error (sSE) based on outcome y_{t+k} is $sSE_t(k) = (y_{t+k} - f_{t,t+k})^2/(\bar{y}_t)^2$ where $\bar{y}_t$ is the mean of $y_{1:t}$. Then the sMSE for forecast horizon k is calculated as the average of the $sSE_t(k)$ over all days t. This has become of interest in evaluating point forecast accuracy of count data as it is well defined (unless all observed values are zero), and it is not as sensitive to high forecast errors as the MSE (Kolassa, 2016). The optimal point forecast under sMSE is the $k-$step ahead predictive mean.

4.5.3 Forecasting Analysis and Comparisons

Items from the analyses are shown in Figure 4.2. These items represent the different levels of demand present in this store: low demand (item A), moderate demand (items C, D), and high demand (item B). For each of the three metrics, we evaluate forecasts across 1:14 days ahead at each day. The benchmark and multi-scale models are evaluated across a range of random effects discount factors $\rho \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. Since different values of ρ may provide better forecasts across the forecasting horizon, we display only the minimum results across each of the five models. That is, for illustration here we only present the results from the best benchmark and multi-scale model for each item, error metric, and forecast horizon.

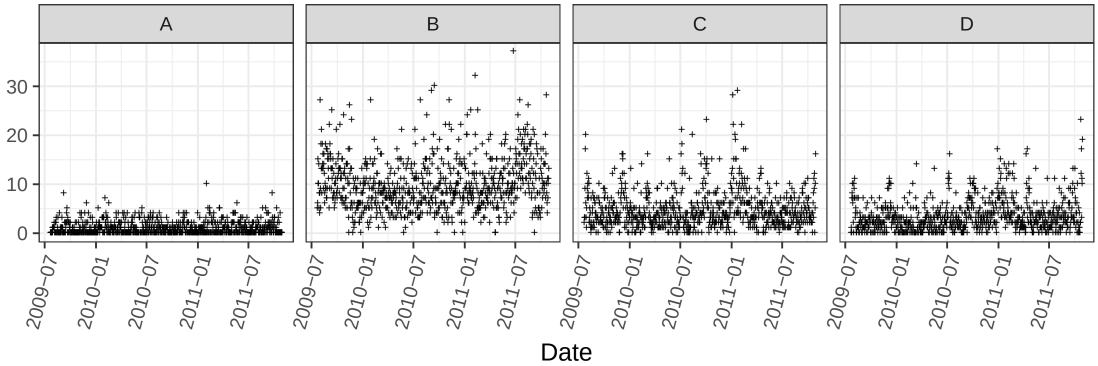


FIGURE 4.2: Daily unit sales (in counts per day) of four spaghetti items A – D in one store from July 22nd 2009 to October 29th 2011.

Comparisons under MAD: Figure 4.3 shows the mean absolute deviations (MAD) versus forecasting horizon for each item from the best performing multi-scale and benchmark DCMMs. For each item, the multi-scale DCMM has lower MAD than the benchmark DCMM for all 14 forecast horizons. The greatest decreases in MAD

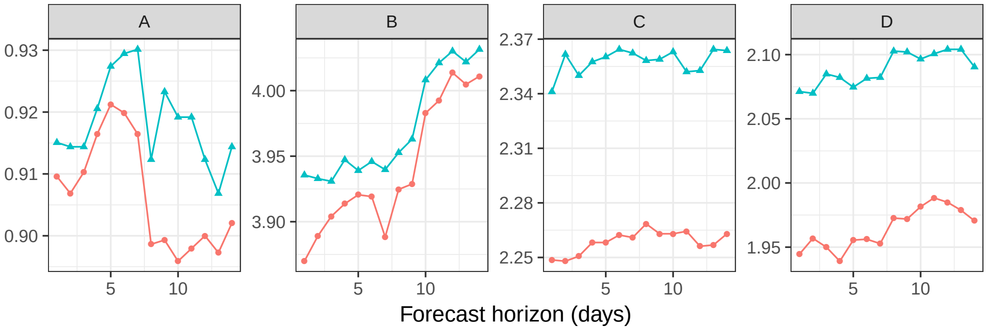


FIGURE 4.3: Mean absolute deviation (MAD) vs forecast horizon (days) for items A – D from the multi-scale (orange circles) and benchmark (blue triangles) models.

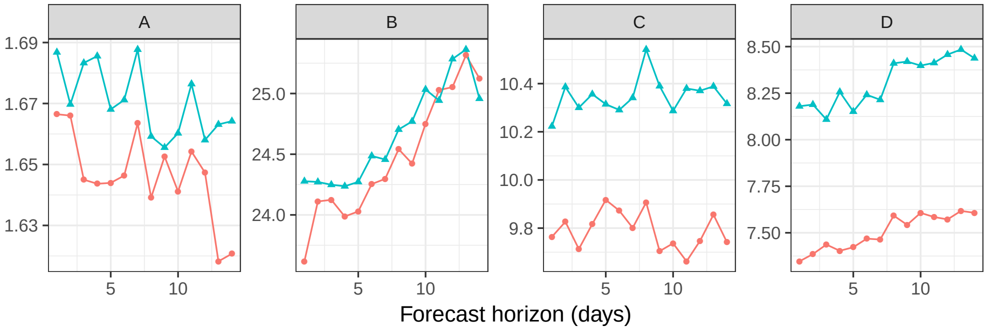


FIGURE 4.4: Mean rank probability score (MRPS) vs forecast horizon (days) for items A – D from the multi-scale (orange circles) and benchmark (blue triangles) models.

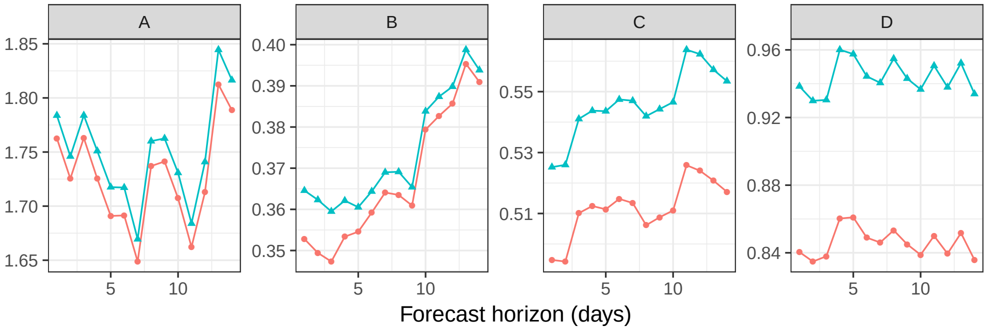


FIGURE 4.5: Scaled mean squared error (sMSE) vs forecast horizon (days) for items A – D from the multi-scale (orange circles) and benchmark (blue triangles) models.

occur for the moderate selling items. Decreases in MAD are small for low-selling items since the predictive median is typically zero or one. For high-selling items, the seasonal pattern may be more evident in the item-level data compared to lower-selling items, thus reducing the potential benefit of the proposed multi-scale approach. In the following comparisons, we present the raw differences in MAD as well as the difference in total absolute deviation. The total absolute deviation may be more interpretable in this context because it represents the error in terms of total number of units over the two-year period. Item specific differences are now noted.

A: The decreases in MAD are small (between 0.004 and 0.024) across the forecast horizons. The largest decreases occur when forecasting 6-14 days ahead. Across the two year period, this corresponds with between 3 to 18 units of accuracy gained from using the multi-scale over the benchmark DCMM.

B: The decreases in MAD range between 0.016 and 0.066 across the forecast horizons. Compared to the benchmark DCMM over this two year period, the multi-scale DCMM is 48, 38, and 32 units more accurate when forecasting 1, 7, and 2−days ahead, respectively.

C: The decreases in MAD range between 0.09 to 0.11 across the forecast horizons. Across this two year period, this corresponds with the multi-scale DCMM being between 64 and 83 units more accurate than the benchmark DCMM.

D: The decrease in MAD ranges between 0.11 to 0.14 across the forecast horizons. Across this two year period, the multi-scale DCMM is between 82 and 104 units more accurate than the benchmark DCMM.

Comparisons under RPS: Figure 4.4 shows the ranked probability score (RPS) versus the forecasting horizon for each item from the best performing multi-scale and benchmark DCMMs. Item specific differences are as follows.

A: The multi-scale DCMMs have lower RPS than the benchmark DCMMs for all 14 forecast horizons. Across the forecast horizons, the average percentage decrease in RPS is about 1.5% for the multi-scale versus benchmark DCMM.

B: The multi-scale DCMMs have lower RPS than the benchmark DCMMs for 12 of 14 forecast horizons. The largest percentage decrease in RPS of 2.7% occurs for 1−day ahead forecasts.

C: The multi-scale DCMMs have lower RPS than the benchmark DCMMs for all 14 forecast horizons. In general, the decreases in RPS are larger for mid-to-long range forecasting. Across the forecast horizon, the average percentage decrease in RPS is 5.4%.

D: The multi-scale DCMMs have lower RPS than the benchmark DCMMs for all 14 forecast horizons. The decreases in RPS are between 0.67 and 0.89 and consistent across the forecast horizons. The average percentage decrease in RPS across the forecast horizons is 9.7%.

Comparisons under sMSE: Figure 4.5 shows the sMSE versus forecast horizon for each item from the best performing multi-scale and benchmark DCMMs. For each item, the multi-scale DCMM has lower sMSE across all of the forecast horizons. Comments by specific items are as follows.

A: The decreases in sMSE are similar across the forecast horizons for the multi-scale versus benchmark DCMM. Averaging across the forecast horizons, the overall percentage decrease in sMSE of the multi-scale DCMM versus the benchmark DCMM is 1.4%.

B: The decreases in sMSE are largest when forecasting $1-7$ days ahead. When forecasting $1-3$ days ahead, the percentage decrease in sMSE of the multi-scale DCMM versus the benchmark DCMM is greater than 3%. Averaged across all of the forecast horizons, the percentage decrease in sMSE is 1.8%.

C: The decreases in sMSE are similar across the forecast horizons. Across the forecast horizons, the average percentage decrease in sMSE of the multi-scale versus benchmark DCMM is 6.3%.

D: The decreases in sMSE are consistent across the forecast horizons. Compared to the benchmark DCMM, the multi-scale DCMM has an average 10.3% decrease in sMSE across the forecast horizons.

Comparisons with Alternate Methods: We did an extensive search for alternative methods that are appropriate for our applied framework. Here we briefly describe the results of three models: the integer valued autoregressive (INAR) model (Al-Osh and Alzaid, 1987; McKenzie, 1988), autoregressive conditional Poisson (ACP) model (also known as the integer valued GARCH model) (Heinen, 2003; Ferland et al., 2006; Fokianos et al., 2009), and the generalized linear ARMA (GLARMA) model (Dunsmuir, 2015; Dunsmuir and Scott, 2015). In this application, we believe the state space modeling framework is more appealing than observation driven modeling due to the adaptability over time, interpretability of model components, and flexibility in modeling aspects of time series of counts. A detailed description of our comparisons with these models is discussed in Section 5. For the INAR model, we were unable to produce comparable forecasts for items A–D in the application. Compared to the ACP/INGARCH model, the multi-scale DCMM had much lower MAD across the forecast horizons for items A–D. The forecast intervals for high-selling item B showed significant undercoverage. Compared to the GLARMA models, the multi-scale DCMM had lower MAD for items A–D across the forecast horizons. For item A, the rPIT plots showed the GLARMA model produced too narrow of forecast distributions, and the binary calibration plots showed miscalibration of zero-versus-nonzero forecasts. Finally, a major concern was that the GLARMA model could not estimate the model for some items with intermittent demand characterized by high sales mixed with periods of zero sales. The GLARMA model performs well in point forecasts for some items, but had poor performance of probabilistic forecasts on high-demand and intermittent-demand items when compared to the DCMM.

4.6 An Example in Multi-Scale Model Inference

In this section, we discuss aspects of model inference in the multi-scale DCMM framework. When updating the multi-scale DCMM, in step 2 of Section 4.4, we take S Monte Carlo samples from $(\boldsymbol{\phi}_{t:t+k} \mid \mathcal{D}_{0,t-1}, \mathcal{I}_{0,t-1})$. Ideally, S is large enough to sufficiently explore the posterior of $\boldsymbol{\phi}_{t:t+k}$. However, as S increases, the computation burden rises as we must compute S DGLM model updates. Conditional on the Monte Carlo samples, these updates can be done in parallel, but the computational burden may still be too large if we have many individual series. We briefly discuss the practical effects that different values of S have on model estimation. We consider $S \in (2, 10, 50, 100, 500, 1000)$, and fit the multi-scale DCMM for each value of S. In step 5 of Section 4.4, S posterior means and covariance matrices are combined to approximate $(\boldsymbol{\theta}_{i,t} \mid \mathcal{D}_{i,t}) \sim (\mathbf{m}_t, \mathbf{C}_t)$. For each candidate value of S, we compare the resulting elements of $\mathbf{m}_t$ and $\mathbf{C}_t$. When S is 50 or less, the resulting elements of $\mathbf{m}_t$ differ noticeably from posteriors for larger values of S. When S was 100 or higher, the elements of resulting posterior means and covariance matrices were practically indistinguishable. Based on these results, we conclude that $S = 100$ is sufficient for model estimation, and ideal from a computational standpoint.

In the multi-scale DCMM, we are able to explore the effects of latent factors over time for different series $y_{i,t}$. If a latent factor has a weak effect on $y_{i,t}$, then $\boldsymbol{\beta}_{i,t}$ will be close to zero. If the latent factor has an effect similar to $\mathcal{M}_0$, then $\boldsymbol{\beta}_{i,t}$ will be close to one. Trajectories of $\boldsymbol{\beta}_{i,t}$ can capture time-variation in the effect of latent factors. In this section, we compare trajectories of $\boldsymbol{\beta}_{i,t}$ across two series to identify the strength latent factors have on each series. This example includes two items, E and F, which are sold in the same store. Figure 4.6 plots the daily sales of these two items from June 22nd 2015 to July 1st 2017. Item E has consistently high demand, and item F has low to medium demand with a strong seasonal pattern over the year. In this analysis, we take models $\mathcal{M}_i, i = 1, 2$ in which the individual DCMMs of eqn. (4.1) have $\mathbf{f}'_{i,t} = (1, \log(\text{price}_{i,t}), \text{promo}_{i,t})$ and $\boldsymbol{\phi}_t$ represents the weekly seasonal effect. The predictor $\text{promo}_{i,t}$ is a binary indicator for whether or not a promotion is occurring for item i on day t. Note that the state evolution matrix is $\mathbf{G}_t = \mathbf{G} = \mathbf{I}$. Discount factors in each $\mathcal{M}_i$ were set to $\delta = 0.995$ in the Poisson DGLM component, and $\delta = 0.999$ in the binary DGLM component.

For $\mathcal{M}_0$, we use a model based on aggregate sales, similar to Section 4.5. We take y_t to be the log of the total sales of all pasta products in this store on day t. We specify $\mathcal{M}_0$ to be a DLM incorporating time-varying effects and stochastic volatility as in Section 2.2. This DLM $\mathcal{M}_0$ has a local linear trend, a regression component with the log average pasta price as a predictor, the full Fourier seasonal component for the 7-day seasonal pattern over the week, and a yearly seasonal effect with the first two harmonics of period 365. Discount factors in this DLM were chosen to be $\delta = 0.995$ for the trend and regression components, $\delta = 0.999$ for the seasonal components, and $\beta = 0.999$ for the residual stochastic variance process.

Figure 4.7(i) plots the seasonal factors of the weekly seasonal effect over time from model $\mathcal{M}_0$. Each color in this plot represents a day of the week, and the effect can be interpreted as the additive increase in log sales, relative to the average, occurring on the specified day of the week. In Figure 4.7, seasonal factors for Tuesdays, Wednesdays, and Thursdays are negative indicating that demand on these days is below average. Similarly, demand on Saturdays and Sundays is above average. Since the seasonal factors are very close to zero, demand on Mondays and Fridays is average. This figure summarizes the weekly seasonality in total daily sales of pasta products within this store. Next, we study whether item E and F have similar weekly seasonal effects.

We denote item E with index $i = 1$ and item F with index $i = 2$. Each element of $\boldsymbol{\beta}_{i,t}$ corresponds to the coefficient on the latent factor representing a particular day of the week. For each item, $\beta_{i,t,1}$ represents the coefficient on the Monday latent factor, $\beta_{i,t,2}$ the coefficient for Tuesday, etc. Figure 4.8 plots the mean and ±2 standard deviations for each element of $\boldsymbol{\beta}_{1,t}$ over time for item E. Over the first few weeks, the coefficients vary quite a bit, and then appear to stabilize. After stabilizing, the coefficients for latent factors representing Wednesday through Sunday are close to one over time. The coefficient on the latent factor representing Monday is between 2 and 3, and has higher variance than the other coefficients. The coefficient for the Tuesday latent factor is about 0.5 over time. These coefficients indicate that the latent factor $\boldsymbol{\phi}_t$ affects $y_{i,t}$ similarly to how it affects the aggregate sales in $\mathcal{M}_0$. Figure 4.7(ii) plots the implied seasonal factors of the weekly seasonality of item E. Comparing this plot to Figure 4.7(i), we can understand the differences in the weekly seasonality of $\mathcal{M}_0$ and item E. As in $\mathcal{M}_0$, the demand on Tuesdays, Wednesdays, and Thursdays is below average. However, compared to $\mathcal{M}_0$, demand for item E on Tuesdays is higher than in the aggregated data. Just as in $\mathcal{M}_0$, the demand on Fridays is average, and the demand on the weekends is above average. In $\mathcal{M}_0$, the demand on Mondays was average, but for item E, we see that there is very slightly above average demand on Mondays.

Figure 4.9 plots the mean and ±2 standard deviations for each element of $\boldsymbol{\beta}_{2,t}$ over time for item F. For item E, after an initial learning period, the coefficients did not seem to vary over time. For item F, there is more variation in the coefficients for Monday, Wednesday, Thursday, and Friday over time. The coefficient on the latent factor representing Monday is centered around zero for the first year, and then begins to steadily decrease in mid-2016 to about -3. The coefficient on Tuesday is close to one for the entire time period. The coefficient on Wednesdays begins around zero, and then increases in mid-2016 to around one for the remainder of the time period. The coefficient on Thursdays begins around one, and then decreases to zero over time. The coefficient on Fridays increases from about three to five. The coefficient on Saturdays varies between two and three over time. The coefficient on Sundays is between zero and one during this time period. Figure 4.7(iii) plots the implied seasonal factor representation of the weekly seasonality of item F. Compared to item E and $\mathcal{M}_0$, there is more uncertainty about the seasonal factors for item F. Similar to $\mathcal{M}_0$, demand on Fridays, Saturdays, and Sundays appears to be above average over time. Demand on Mondays, Wednesdays, and Thursdays is average. Demand on Tuesdays is slightly below average.

The multi-scale DCMM allows us to compare dependence on shared latent effects $\boldsymbol{\phi}_t$ through examination of item-specific coefficients $\boldsymbol{\beta}_{i,t}$. As showcased by these two items, incorporation of item-specific coefficients allows each item to “modify” the dependence on $\boldsymbol{\phi}_t$. For example, the weekly seasonality of item E appears to be very similar to the seasonality of the store traffic. In general, the item-specific coefficients are close to one over time. The weekly seasonality of item F, however, has some slight differences from the seasonality of item E and the overall store traffic. Notably, item F has higher average sales on Fridays and Saturdays relative to the overall store traffic seasonality. Time-varying coefficients $\boldsymbol{\beta}_{i,t}$ allow each model to capture dynamic dependence on $\boldsymbol{\phi}_t$. Item E has a stable weekly seasonality over time while the weekly seasonality of item F appears to change over time. In product sales forecasting, identification of patterns and inference of this nature can tie into decisions regarding marketing and production. The straightforward interpretation of the multi-scale coefficients allows easy comparison of patterns across items and provides an added benefit of the multi-scale framework in addition to improving forecasting performance.

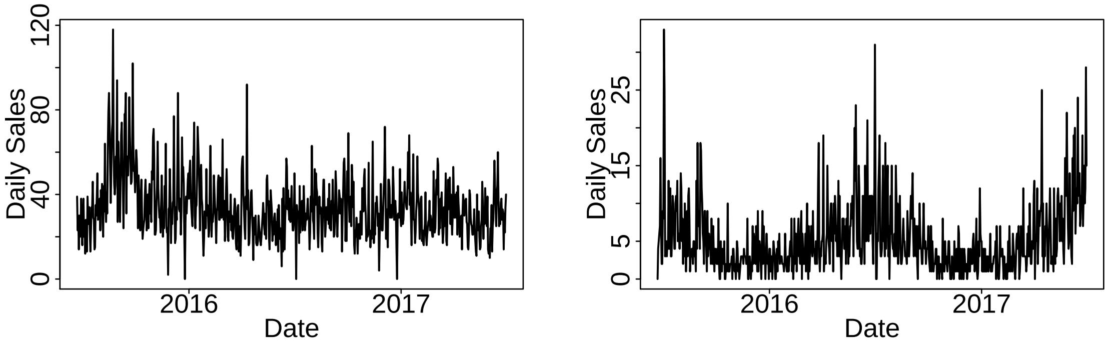
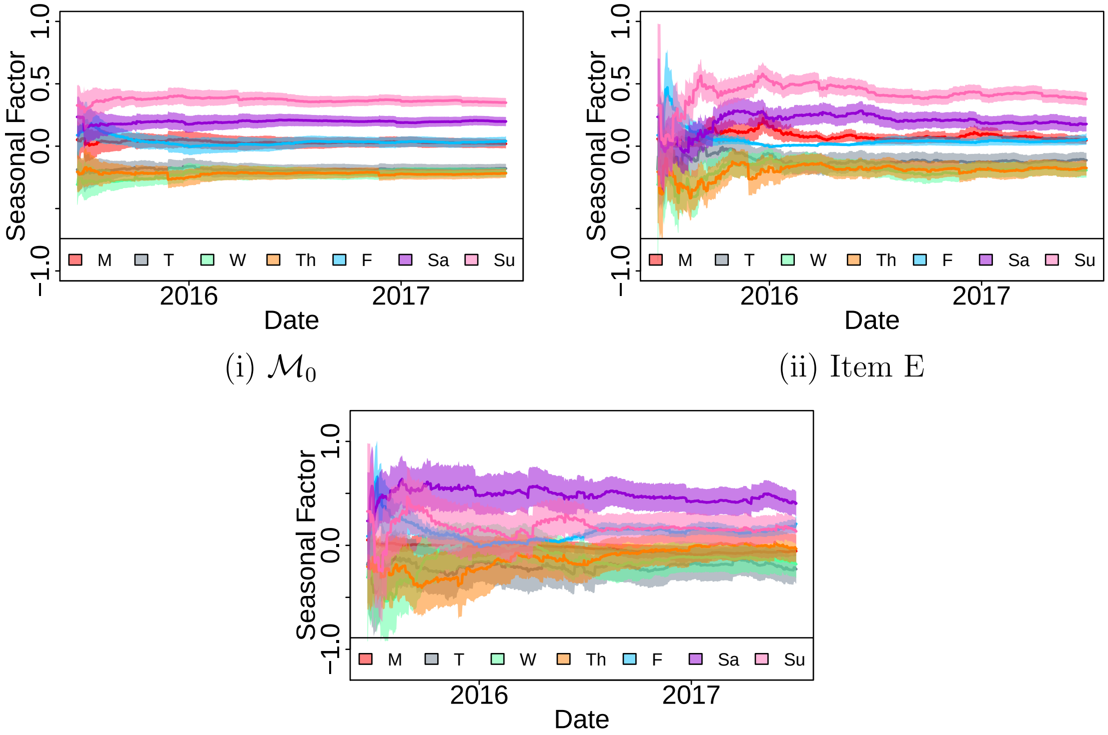


FIGURE 4.6: Daily sales of items E (left) and F (right) from June 22 2015 to July 1 2017.



(i) $\mathcal{M}_0$

(ii) Item E

(iii) Item F

FIGURE 4.7: Trajectories of weekly seasonal factors over time for the daily sales of (i) all pasta in the selected store, (ii) item E, and (iii) item F.

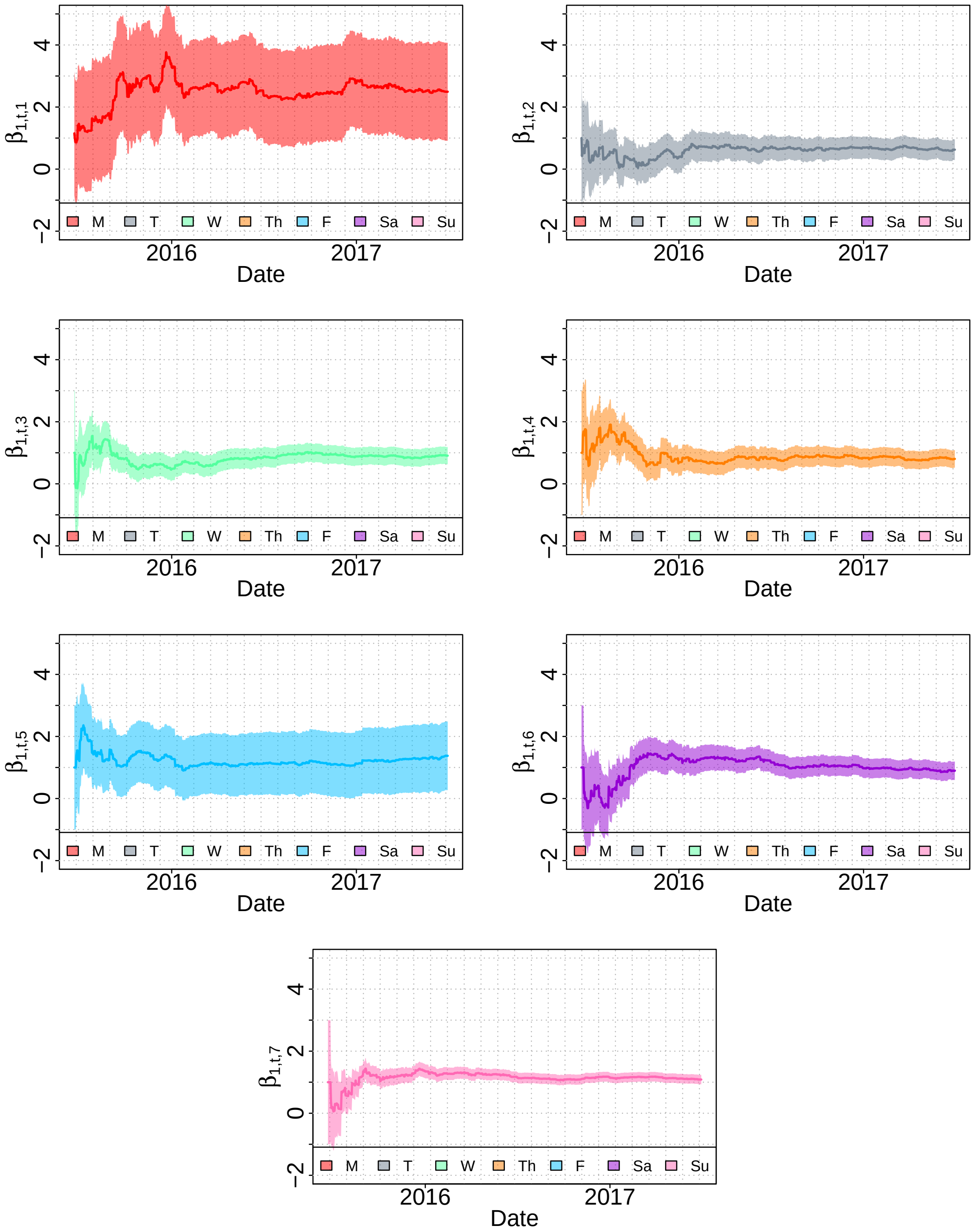


FIGURE 4.8: Trajectories of multi-scale coefficients $\beta_{i,t}, i \in 1{:}7$ over time for item E.

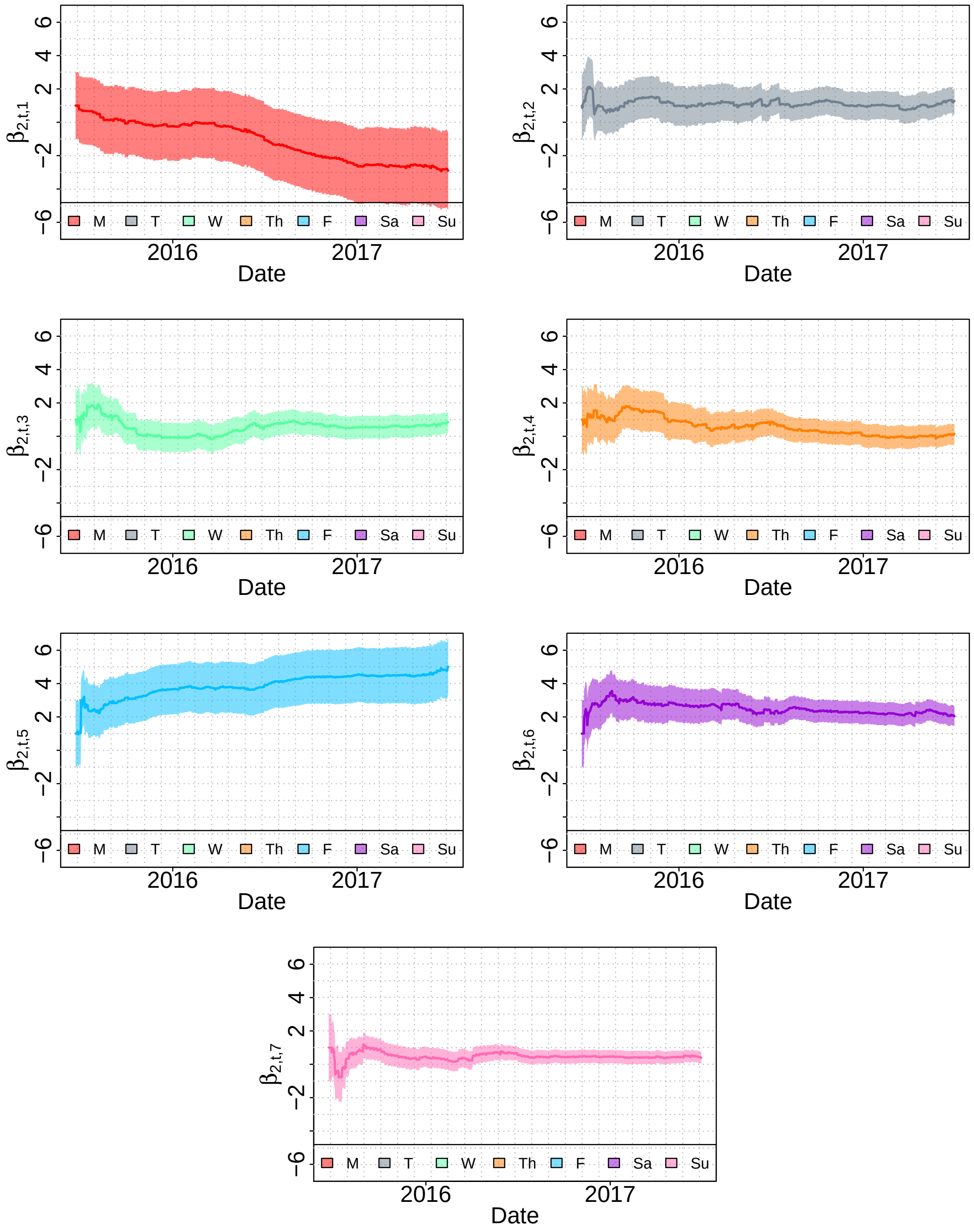


FIGURE 4.9: Trajectories of multi-scale coefficients $\beta_{i,t}, i \in 1{:}7$ over time for item F.

4.7 Multi-Scale Aggregate Model Exploration

Section 4.5 details an application of the multi-scale DCMM, and the results suggest the potential of this framework to improve $1 - 14$ day ahead forecasting at the item level compared to DCMMs with item specific seasonality. This application focused on items of one type (spaghetti) in a single store, amd $\mathcal{M}_0$ was specified as a lognormal DLM on the total daily spaghetti sales within the chosen store. This choice of $\mathcal{M}_0$ acted as a proxy for store traffic, and the common factor ϕ_t represented the store-spaghetti-level weekly seasonality. In this section, we compare item-level forecasting performance under different choices of $\mathcal{M}_0$. Specifically, we compare 1–14 day ahead point forecast performance of univariate DCMMs and multi-scale DCMMs with various choices of $\mathcal{M}_0$. This example includes 22 items of four pasta types (spaghetti, lasagna, egg, macaroni) within one store. For the multi-scale DCMMs, we specify $\mathcal{M}_0$ to be a lognormal DLM on the daily sales of various levels of aggregate data. For an item with UPC u, pasta type p, in store s, we consider multi-scale DCMMs with the following levels of aggregate data:

- **All-pasta:** Total pasta sales across all stores.
- **All-type:** Total sales of pasta type p across all stores.
- **All-UPC:** Total sales of UPC u across all stores.
- **Store-pasta:** Total pasta sales within store s.
- **Store-type:** Total sales of pasta type p within store s.

For each of the multi-scale DCMMs, the binary and conditionally Poisson DGLMs have a local level, regression component with log price as a predictor, and they inherit the weekly seasonal effect from $\mathcal{M}_0$. For both binary and Poisson DGLMs, the prior mean is set using 21 days of initial training data. The prior mean of the multi-scale seasonal coefficients is set to one, and the prior mean of the regression component is set to zero. The prior covariance matrix of the state vector is set to $0.1\mathbf{I}$. In the univariate DCMM, the binary and conditionally Poisson DGLMs have a local level, regression component with log price as a predictor, and a full Fourier form seasonal component with period 7. We use 21 days of initial training data to set the prior means of the state vector elements corresponding to the level and the seasonal component. We set the prior mean of the regression coefficient to zero. The prior covariance matrix of that state vector is set to $0.1\mathbf{I}$. For independent and multi-scale DCMMs, we fix discount factors in the binary DGLM to 0.999 on each model component, and to 0.99 in each component of the conditionally Poisson DGLM. In this analysis, we train the models on 100 days of data before we begin forecast 1–14 days ahead over the next 100 days. We use the MAD and RMSE to evaluate point forecast accuracy. For all models, we display error metrics averaged across the 100 day period for each forecast horizon. Results are shown in Figures 4.10, 4.11, 4.14, 4.12, 4.13, and 4.15. We now detail the results for each pasta type.

Results for spaghetti items: Figures 4.10 and 4.11 display the MAD and RMSE for the eight spaghetti items in this analysis. The black line corresponds to the univariate DCMM, and the colored lines correspond to the various multi-scale DCMMs. In general, most of the spaghetti items see some improvement in item-level forecasting under one of the multi-scale DCMMs compared to univariate DCMMs. For the highest selling spaghetti items 1–4, the Store-pasta and Store-spaghetti multi-scale model improves point forecasting performance compared to the independent DCMM. Results are varied for the lower selling spaghetti items 5–8, but in general it seems that the best choice of $\mathcal{M}_0$ for spaghetti items is a lognormal DLM fit to the total daily sales of all pasta items in the specific store. We now discuss additional item specific results.

1: The multi-scale DCMMs with the lowest MAD and RMSE are the Store-pasta and Store-spaghetti models. This suggests that this item's weekly seasonality is well represented by the store traffic weekly seasonality.

2: The multi-scale DCMMs with the lowest MAD and RMSE are the Store-pasta and All-pasta models. This suggests that this item's weekly seasonality is more similar to the overall Pasta category sales rather than the total spaghetti sales.

3: MAD for $1-14$ day ahead forecasts and RMSE for $6-14$ day ahead forecasts is lower for all multi-scale DCMMs compared to the univariate DCMM. There does not appear to be a significant difference in forecast error between the multi-scale models.

4: All of the multi-scale DCMMs have lower MAD and RMSE than the univariate DCMM. In general, the model with the lowest errors is the All-UPC model. This suggests that the weekly seasonality of this item's sales is very similar to the weekly seasonality of the total sales of this UPC across stores.

5,7,8: Results for MAD and RMSE are mixed for these items with one showing better performance of the multi-scale models and the other showing better performance of the univariate model.

6: MAD and RMSE is lower for the multi-scale models, although there is no discernible difference between the multi-scale forecast errors.

Results for macaroni items: Figures 4.12 and 4.13 show the MAD and RMSE versus forecast horizon for the five macaroni pasta items in this analysis. For four of these items, the multi-scale models have generally lower forecast errors than the univariate DCMM. However, for one item, the univariate DCMM has substantially lower RMSE and MAD than all of the multi-scale models. We now discuss item specific results.

1: The univariate DCMM and Store-macaroni multi-scale DCMM have the largest

MAD and RMSE. The multi-scale model with the lowest RMSE is the All-pasta model. For MAD, the All-pasta model has the lowest error for short to mid range forecasting, and for mid to long range forecasting, the All-UPC model has the lowest MAD. These results indicate that forecasting performance improves when the weekly seasonlity of this item is estimated using data aggregated to higher levels than the individual item or store-macaroni.

2: The MAD and RMSE of the multi-scale DCMMs is much lower than the error of the univariate DCMM. Relative to the independent DCMM, MAD decreases between 10–40% across the forecast horizon with the greatest improvements occurring for long term forecasting. Similarly, for RMSE, the multi-scale DCMMs decrease the error by more than 40% for long-term forecasting from 12–14 steps ahead. The forecasting performance of each of the multi-scale models is relatively similar, so it appears that any choice of aggregate data in $\mathcal{M}_0$ is beneficial.

3: In general, the multi-scale models have lower MAD and RMSE than the univariate DCMM.

4: Forecast performance of all of the models is similar, although the RMSE of the univariate DCMM is slightly higher than the multi-scale models for low-to-mid range forecasts.

5: The MAD and RMSE of the univariate DCMM is much lower than the multi-scale DCMMs. Each of the multi-scale DCMMs has the same MAD across the forecast horizons, suggesting that this is a low-selling item for which the multi-scale models predict all zeros.

Results for lasagna items: The lower three rows of figure 4.13 show the MAD and RMSE versus forecast horizon for the three lasagna items in this analysis. In general, the multi-scale models have lower forecast errors than the univariate DCMM.

1,2: The MAD and RMSE of the univariate DCMM is higher than that of the multi-scale DCMMs. The errors of the multi-scale DCMMS are all relatively similar.

3: The MAD of the univariate DCMM is higher than the MAD of the multi-scale models. The RMSE of the univariate DCMM is lower for short-range forecasts, and larger for longer-range forecast horizons.

Results for egg-based pasta items: MAD and RMSE versus forecast horizon for egg-based pasta items are shown in Figures 4.14 and 4.15. In general, the error of multi-scale DCMMs is lower than the error of the univariate DCMM. Item specific details are now discussed.

1: The multi-scale DCMMs have lower MAD and RMSE than the univariate DCMM. The model with the lowest forecast errors is the All-UPC model, suggesting that the weekly seasonality of the sales of this item are similar to the sales of the same UPC at other stores.

2: The MAD of the multi-scale and univariate DCMMs is similar across the forecast horizons. The multi-scale DCMMs have lower RMSE than the univariate DCMM across the forecast horizon.

3,4,6: The multi-scale DCMMs have lower MAD and RMSE than the univariate DCMM. The performance of all of the multi-scale DCMMs is relatively similar for both metrics.

5: Compared to the multi-scale DCMMs, the univariate DCMM has similar MAD for short-range forecasts and lower MAD for longer-range forecasts. The multi-scale DCMMs have lower RMSE than the univariate DCMM. The multi-scale models with the lowest RMSE are the All-pasta and All-egg models, although the forecasts across all models are similar.

In this section, we have explored different choices of $\mathcal{M}_0$ in the multi-scale DCMM framework. We have compared the point forecast accuracy of univariate DCMMs and five choices of multi-scale DCMM across a set of twenty-two items of differing demand levels and pasta types. Although there are a few exceptions, it appears that the multi-scale DCMM improves point forecast accuracy for most items. The performance of the five multi-scale DCMMs was similar for most items indicating that any aggregate series is appropriate for $\mathcal{M}_0$. For a few items, point forecast accuracy improved when the data modeled in $\mathcal{M}_0$ was aggregated to a higher level than the Store-type data. These results suggest that a good choice of $\mathcal{M}_0$ is a lognormal DLM fit to data aggregated to the level of total store-pasta sales or higher.

4.8 Summary Comments

In the context of a motivating case study and application in consumer sales and demand forecasting, we have introduced a novel class of dynamic state-space models for time series of non-negative counts, and a formal multivariate extension for many related series. The univariate DCMM framework builds on and extends prior approaches to univariate count time series, contributing a flexible, customizable class of models. The ability to explore and include covariates as potential predictors of both binary and positive count series is of interest in many areas, and the coupling of DGLMs for these two components addresses a very common need with count data. Motivated by problems in consumer demand and sales forecasting, the opportunity to apply these models and extend their use in commercial and socio-scientific forecasting is evident. In addition to the dynamic regression components, the time-varying state-space framework allows evaluation of changes over time in regressions, trends, seasonal patterns and other forms of predictor information, and adaptability to any such changes. The Bayesian framework defines probabilistic forecasts that, accessed trivially computationally through direct, forward simulation of predictive distributions, enables evaluation of various summary measures of uncertainty about forecast paths of time series into the future and arbitrary functions of sets of future outcomes. This is important in applications as a general matter of properly communicating forecast information, and also provides the basis for formal decision analysis in decision contexts reliant on forecasts. Our examples developed in consumer sales forecasting highlight the machinery of model fitting and forecasting, and demonstrate the utility of DCMMs with series exhibiting quite differing patterns and levels of outcome intensity. This is important in applications involving many series where it is desirable to have a single model class that is flexible and adaptable to individual series characteristics. A number of traditional point forecast metrics are also discussed, along with the point that they should always be considered so long as the background context supports the role of the implicit loss function underlying any specific point forecast. More broadly on forecast assessment and evaluation, we have stressed and exemplified the use of analyses addressing the full forecast distributions, to include frequency calibration of both binary and positive count models, empirical coverage of nominal forecast intervals.

The embedding of sets of DCMMs into a multivariate system defines a novel class of state-space models for many related time series of counts. Importantly, this maintains the flexibility of modeling at the univariate series level, using individual DCMMs that are linked across series via common latent factors. The linkages are series-specific, potentially time-varying random effects, so defining an overall, flexible hierarchical dynamic model framework. Also, critically, the new multivariate/multiscale approach maintains the ability to run fast, sequential Bayesian analysis of decoupled univariate analyses of many series, with recoupling across series based on information about common factors flowing from an external model. This strategy enables analytic computations and trivial forward simulation for sequential analysis and forecasting, and by design is scalable to many series (computations grow only linear with the number of series). Our example in the motivating consumer sales case study involves a common factor process related to shared seasonal patterns and in which the external model generating inferences on the factor process is a dynamic model applied to aggregate data in which the pattern is more precisely identified. Other implementations of the multi-scale approach, beyond purely seasonality, offer the possibility of learning additional relevant shared features across products, stores, or brands. For example, we could use this approach to learn the elasticity to promotions which are run in many stores. In future applications, the shared latent factor process will be multivariate, with dimensions reflecting different ways in which series are conceptually related. In product demand forecasting, for example, products can be grouped by product family, brand, store location and other factors, and both aggregate-level and external economic or business models may provide inputs to forecast several common factors representing the relevant cross-series linkages. Our examples illustrate the ability of the multivariate/multi-scale approach to improve forecasts at the individual series level, in both short and longer-term forecasting, and across series with intermittent, moderate, and high demand patterns. While forecast accuracy improvements cannot be expected for all series all of the time, even small increases in forecast accuracy on a number of items can have a profound impact on retail decision-making and costs. One important factor that plays a role here is the length of historical data available for each item. It is to be expected that series with shorter histories will most immediately benefit from the multivariate approach, as common features impacting demand on similar items will feed information relevant to forecasting the newer items, a context of clear interest in commercial settings when new or modified products are introduced.

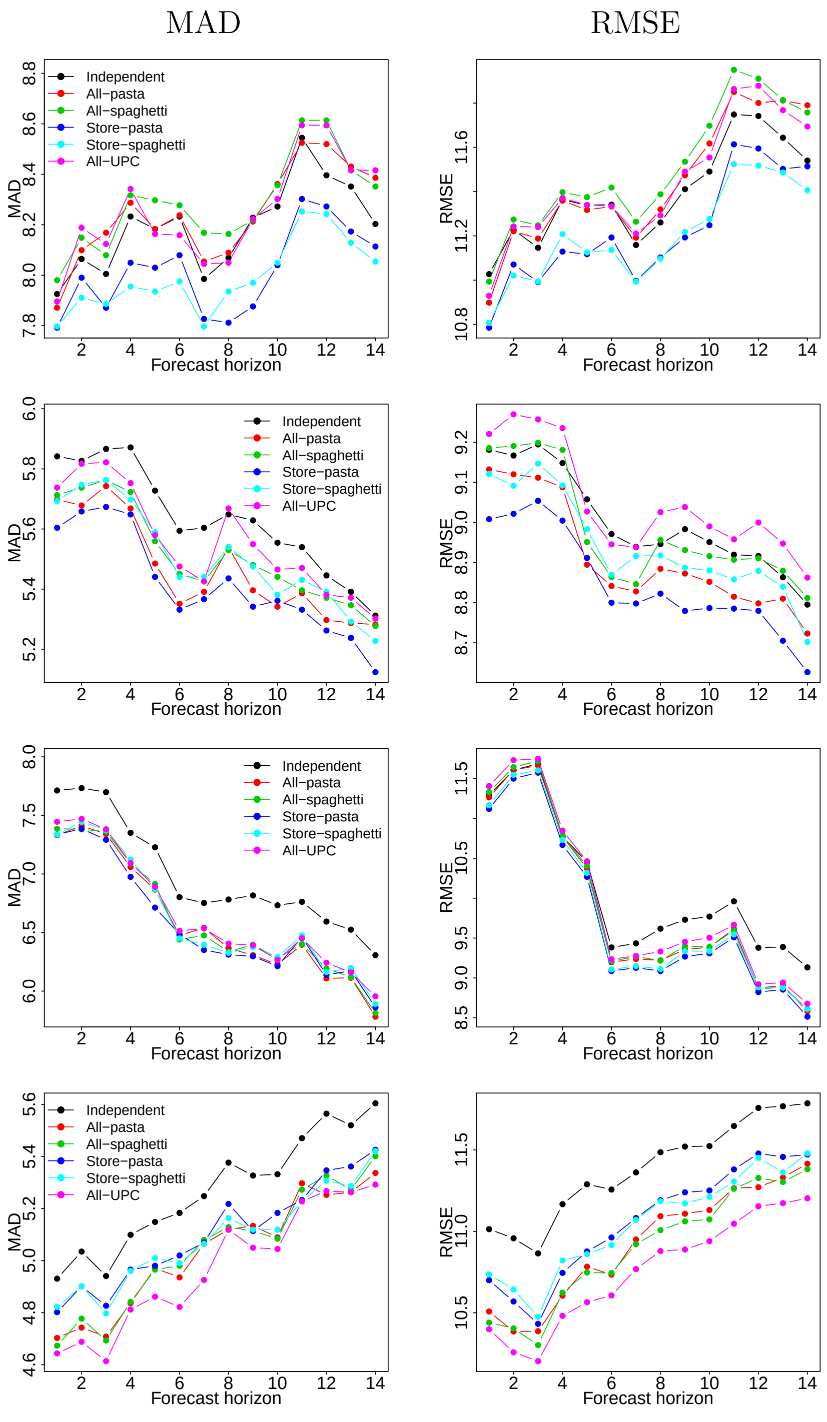


FIGURE 4.10: Forecast comparison in terms of MAD (left) and RMSE (right) for spaghetti items 1–4 using following DCMMs: independent (black), All-pasta (red), All-spaghetti (green), Store-pasta (dark blue), Store-spaghetti (light blue), and All-UPC (pink).

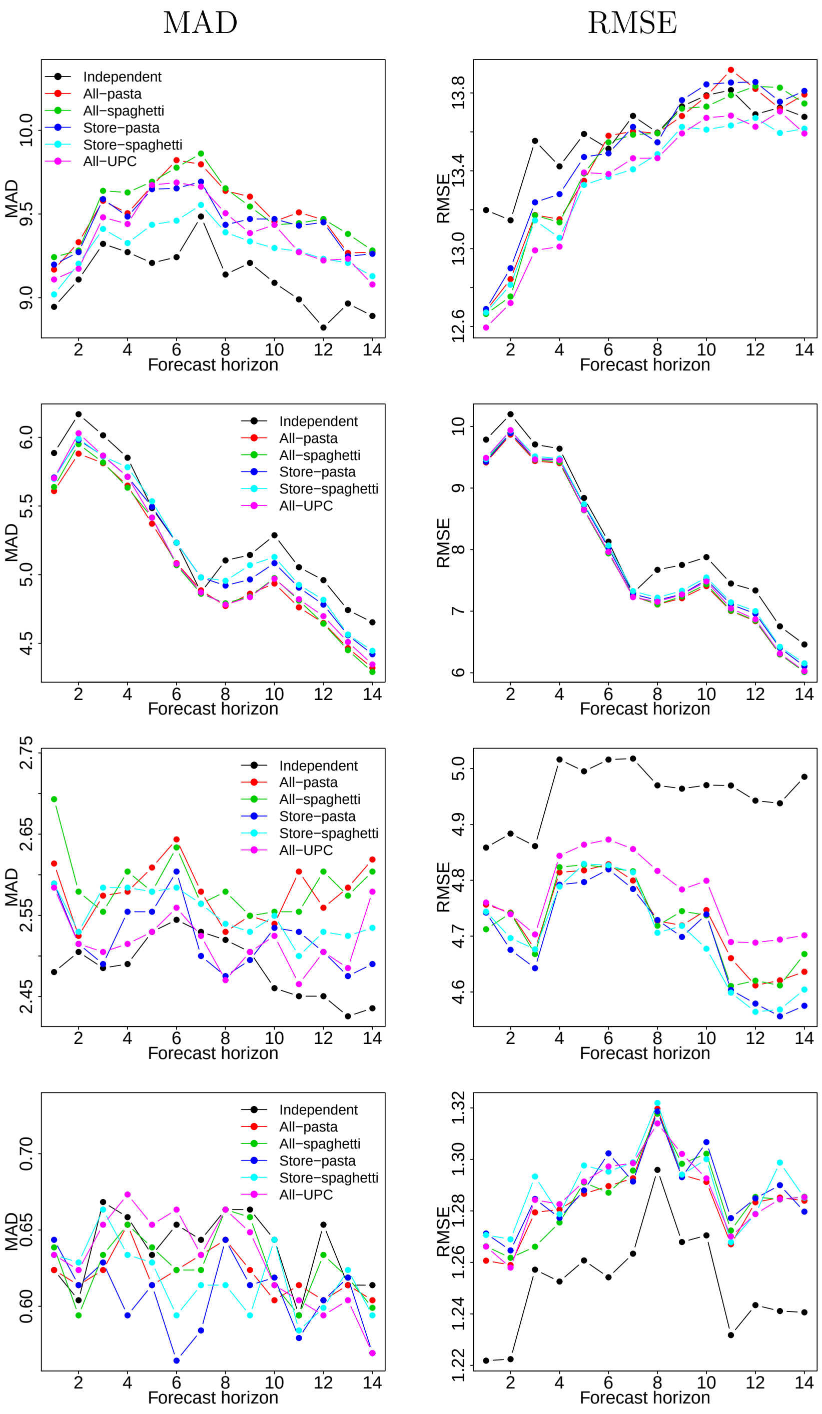


FIGURE 4.11: Forecast comparison in terms of MAD (left) and RMSE (right) for spaghetti items 5–8 using following DCMMs: independent (black), All-pasta (red), All-spaghetti (green), Store-pasta (dark blue), Store-spaghetti (light blue), and All-UPC (pink).

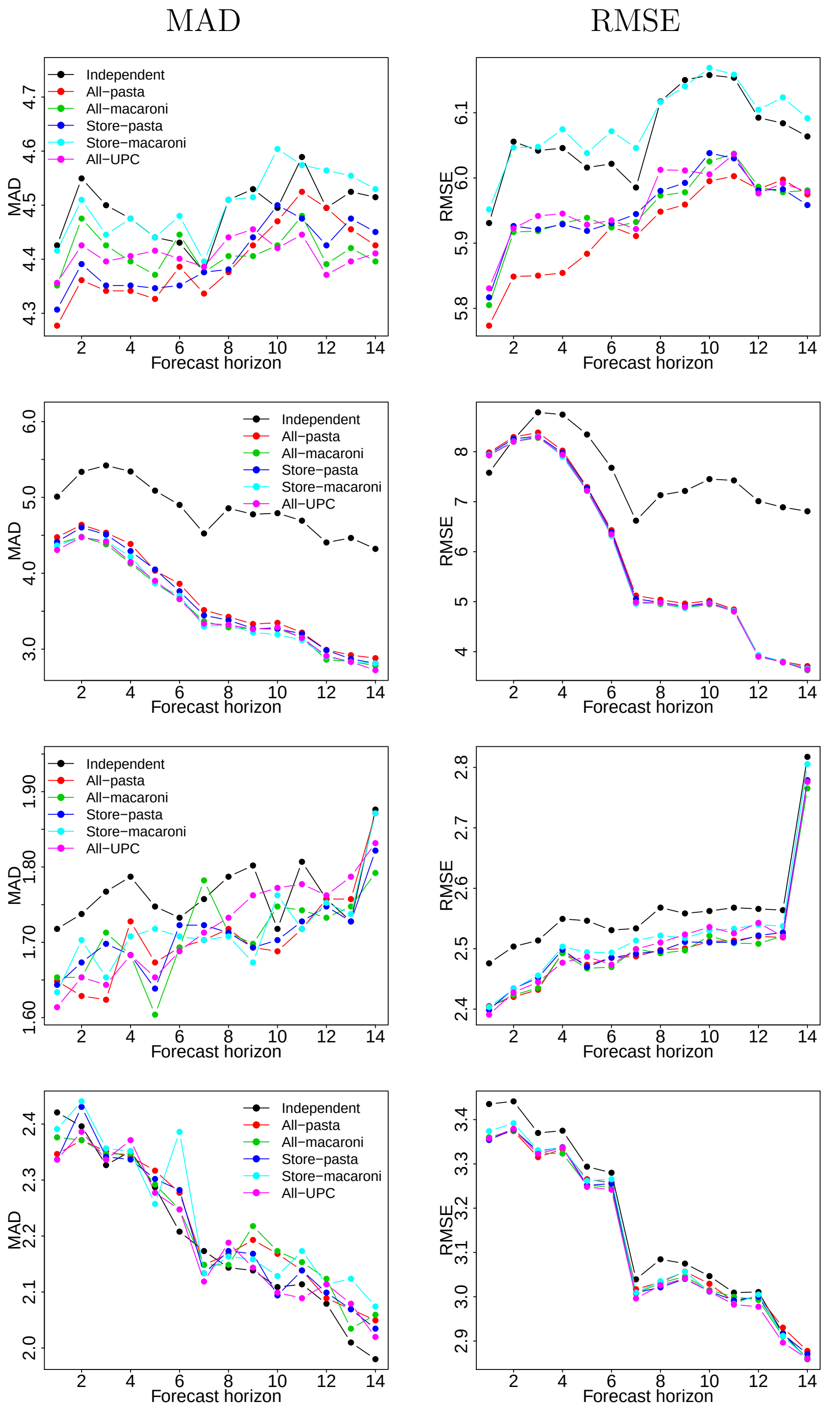


Figure 4.12: Forecast comparison in terms of MAD (left) and RMSE (right) for macaroni items 1–4 using following DCMMs: independent (black), All-pasta (red), All-spaghetti (green), Store-pasta (dark blue), Store-spaghetti (light blue), and All-UPC (pink).

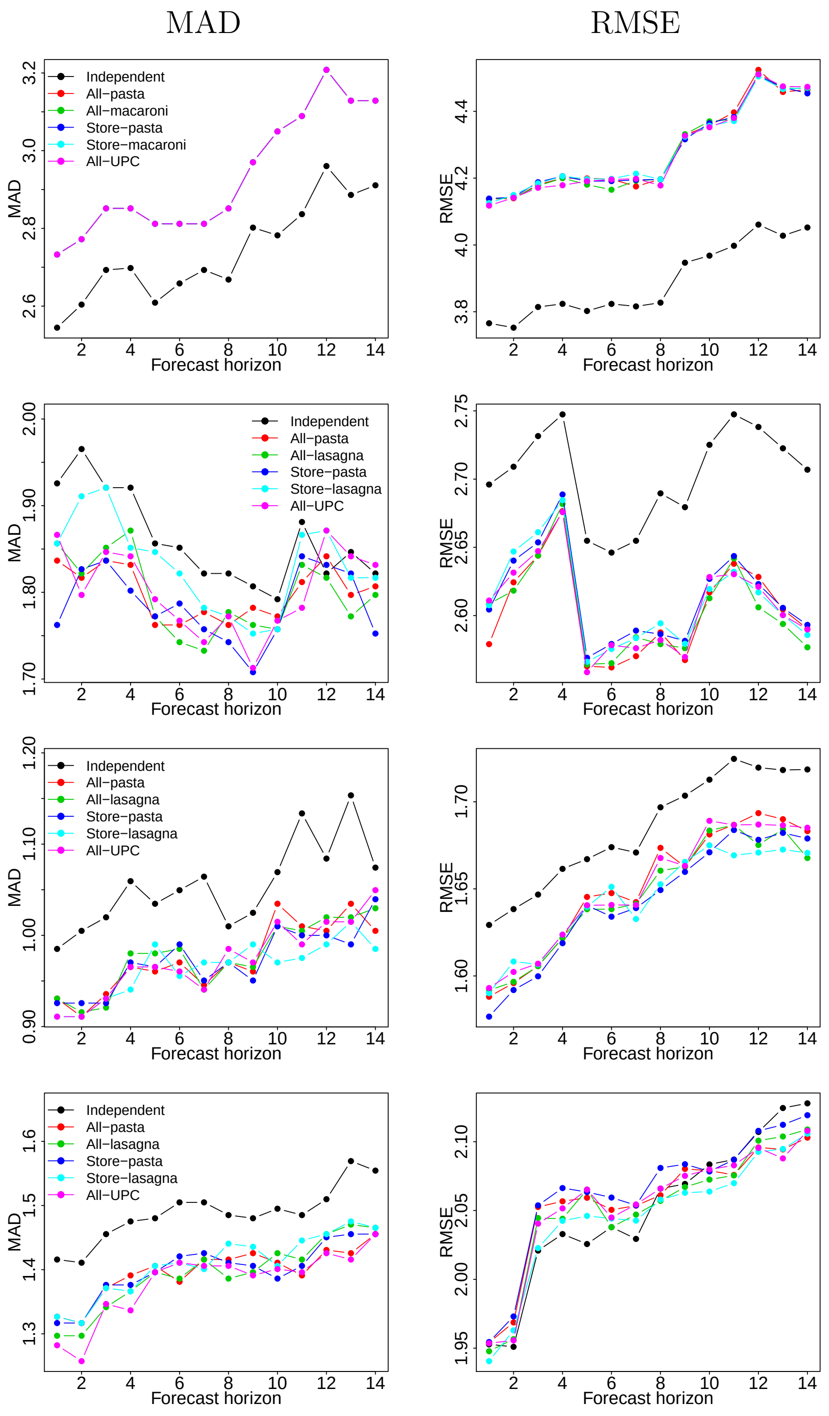


FIGURE 4.13: Forecast comparison in terms of MAD (left) and RMSE (right) for macaroni item 5 and lasagna items 1–3 using following DCMMs: independent (black), All-pasta (red), All-spaghetti (green), Store-pasta (dark blue), Store-spaghetti (light blue), and All-UPC (pink).

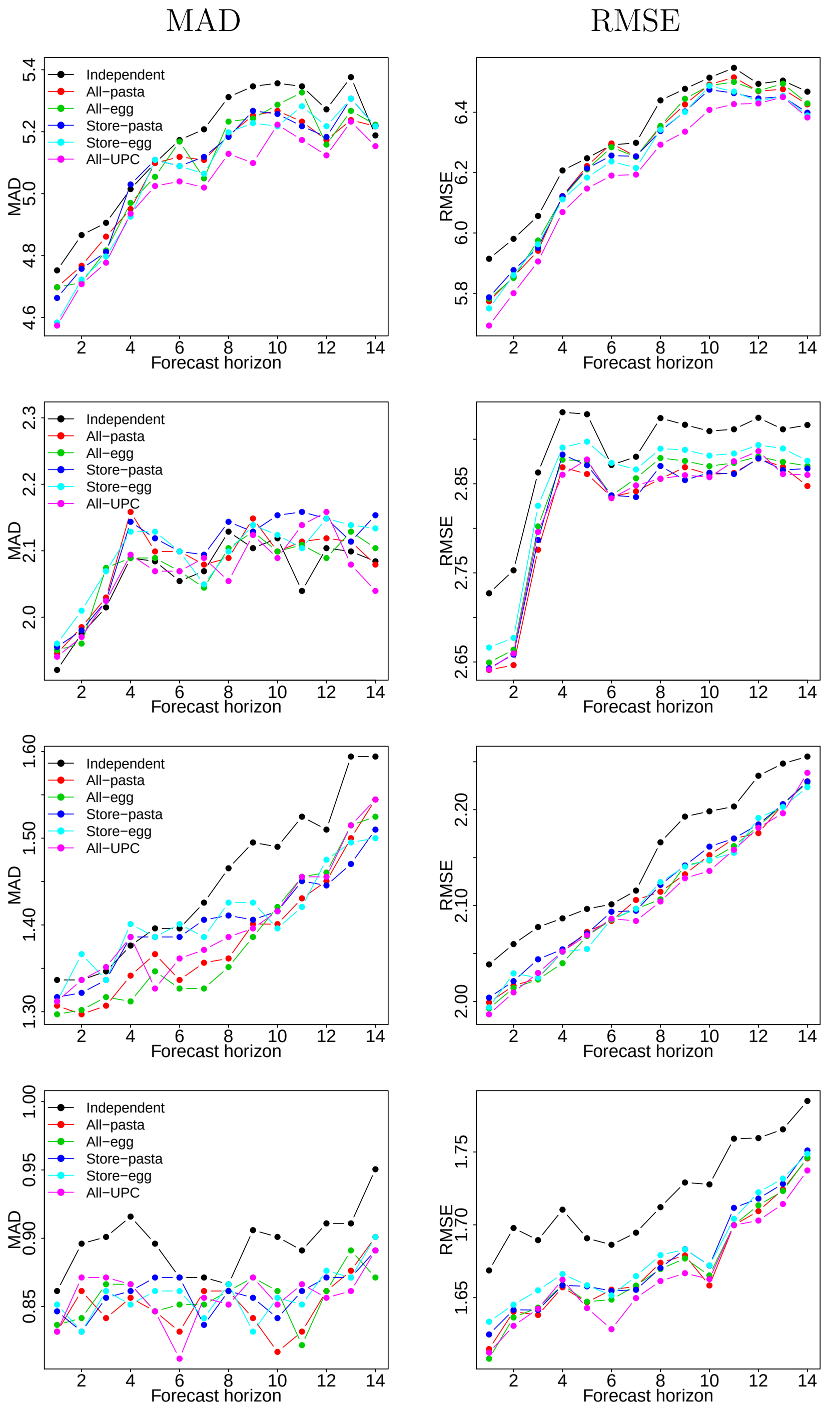


FIGURE 4.14: Forecast comparison in terms of MAD (left) and RMSE (right) for egg items 1–4 using following DCMMs: independent (black), All-pasta (red), All-spaghetti (green), Store-pasta (dark blue), Store-spaghetti (light blue), and All-UPC (pink).

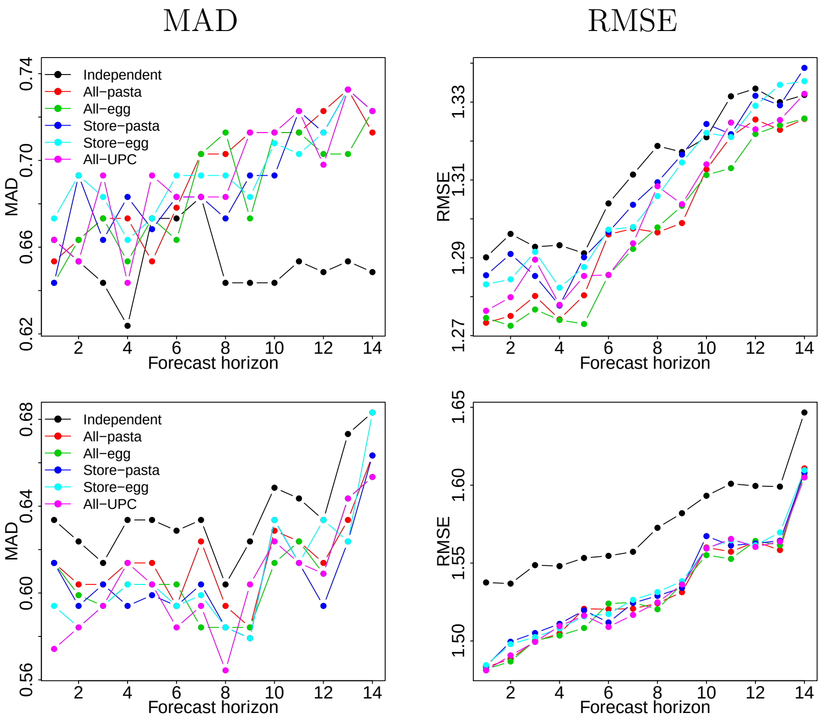


FIGURE 4.15: Forecast comparison in terms of MAD (left) and RMSE (right) for egg items 5–6 using following DCMMs: independent (black), All-pasta (red), All-spaghetti (green), Store-pasta (dark blue), Store-spaghetti (light blue), and All-UPC (pink).

5

Comparison of DCMM with Alternate Forecasting Methods

In this section, we discuss several existing models that could be used in the context of product demand forecasting. For each comparison method, we briefly describe the model, detail the availability of code/packages, and (if possible) summarize the performance of each model on items in our dataset. Given our focus on forecasting low-valued count data, we limit the scope of models considered here to those which produce coherent forecasts over the non-negative integers. To compare forecast results, we present various point forecast metrics as well as probabilistic forecast evaluation.

Each of the models considered here fall in the framework of observation-driven models with some flavor of autoregressive dependence structure between observations. Before describing model specific details, we detail a few key differences between DCMMs and the following methods:

- All of the following models are univariate, and we have proposed an *efficient* multi-scale framework for incorporating cross series dependence.

- In the following models, the ARMA, regression, and seasonal components are static while the DCMM can incorporate dynamic versions of these model components.

- Some of the models mentioned below account for overdispersion – typically through an extension from a Poisson to Negative Binomial distribution. However, in our applied framework with potentially millions of individual series, it is not realistic to individually specify the appropriate model for each individual item. Additionally, the appropriate model may change over time if the demand for an item changes (e.g. seasonal products). The DCMM random effects extension we have proposed can *automatically* account for overdispersion in time series of counts.

- In the following models, there is no way to account for data with more/fewer zeros than expected under Poisson/Negative Binomial models. In the DCMM, the binary component is automatic and flexible enough to model any time series of counts (with fewer/excess zeros) without the need for individual item model customization.

5.1 Integer Valued Autoregressive Model

Details of the integer valued autoregressive (INAR) model are given in Al-Osh and Alzaid (1987) and McKenzie (1988). An alternate name for the same model is the Poisson Autoregressive (PAR) model. Let y_t denote a non-negative count valued time series observed over time $t = 1, \dots, T$. The INAR model is defined by

$$y_t = \alpha \circ y_{t-1} + \epsilon_t$$

where y_0 has a Poisson distribution with mean λ_0, and $\epsilon_t \sim Po(\lambda)$, with $\epsilon_t \perp y_k$ for all t, k. The binomial thinning operator $\circ$ is defined as follows: given X_{t-1},

$\alpha X_{t-1} = \sum_{i=1}^{X_{t-1}} B_{it}$, where $B_{1t}, \ldots, B_{X_{t-1},t}$ are iid Bernoulli random variables with success probability α.

The intuition underlying the INAR model is that the count at time t is the sum of "survivors" from time $t-1$ and new arrivals between time $t-1$ and t. The survivors are represented by the binomial thinning, $\alpha \circ y_{t-1}$, and the new arrivals are represented by ϵ_t, generally a Poisson random variable. This simple intuition is appealing in many applications, and could be relevant when modeling the number of customers waiting in line at a store. However, when modeling the daily sales of a supermarket product, the birth-death concept does not seem to be a natural way of understanding the data.

Additionally, it is not trivial to extend this model to incorporate covariate effects. Incorporating item level price, promotions, and holiday effects is a very important part of product sales forecasting. Finally, it is unclear how this model would incorporate negative dependence between observations. Both of these tasks are possible in the DCMM framework.

We spent many hours searching, but we were unable to find any code/packages to implement the INAR model. It is possible for us to code up the estimation procedure to produce MLE estimates of the model parameters ourselves. Given these MLEs, we could write code to produce $k-$step mean forecasts for the INAR model. However, given that our time series are low-valued counts, the mean forecast is not especially interesting, and our main interest is in entire forecast distributions. It is unclear if $k-$step predictive distributions are available in this model, and would be a separate research project to implement this ourselves. Given these experiences, we have focused on other comparison models which have similarities to the INAR model in that they are observation-driven and have an autoregressive dependence structure.

5.2 Generalized Linear ARMA Model

The Generalized Linear ARMA model (GLARMA) is described in Dunsmuir (2015) and Dunsmuir and Scott (2015). Let y_t denote a non-negative count valued time series observed over time $t = 1, \dots, T$. Here we detail the Poisson GLARMA model, however there is also a Negative Binomial GLARMA for modeling overdispersed counts. The model is

$$y_t \mid \mathcal{D}_{t-1} \sim Po(\mu_t), \tag{5.1}$$

$$\log(\mu_t) = W_t = \mathbf{x}_t' \boldsymbol{\beta} + Z_t, \tag{5.2}$$

where

$$Z_t = \sum_{i=1}^{p} \phi_i (Z_{t-i} + e_{t-i}) + \sum_{i=1}^{q} \theta_i e_{t-i} \tag{5.3}$$

and

$$e_t = \frac{Y_t - \mu_t}{\nu_t}$$

for some scaling factor ν_t. For the models considered here, we set e_t to be Pearson residuals with $\nu_t = \sqrt{\mu_t}$.

The GLARMA model is implemented in the `glarma` R package. This package allows incorporation of static regression effects and static ARMA dependence structure. Multi-step forecasting is available through a simulation based approach – allowing access to the full predictive distribution k-steps ahead. Unlike the DCMM, the GLARMA model does not account for excess zeros. To account for overdispersion, this package allows both conditionally Poisson and Negative Binomial response distributions. Another important detail is that the GLARMA package does not handle time series with missing data which is straightforward in the DCMM framework.

To compare with the DCMM, we implement the GLARMA models with both Poisson and Negative Binomial response distributions. Each model includes the log price as a predictor and six dummy variables for the day of the week to incorporate the weekly seasonal effect. We also include a lag-1 autoregressive structure to incorporate time dependence in the model.

As in our DCMM analysis, we fit the model on the first 100 days of data, and then began forecasting $1 - 14$ days ahead on each day during the next two years. In the GLARMA framework, this requires re-estimating the model on all of the observed data at each time t, and then forecasting over the next 1-14 days using the built in GLARMA simulation method. We run this analysis across the same 4 items (A – D) in the multiscale forecasting application in Section 4.5. Figure 4.2 plots the daily sales of items A – D over this time period. Figure 5.1 displays the MAD of forecasts under each model for items A – D. Figures 5.2 and 5.3 display probabilistic forecast metrics for the GLARMA models. We now discuss the specifics of the forecasting results for each item.

Item A. Here we only present results for the Poisson GLARMA model since the Negative Binomial GLARMA model encountered an error during estimation for item A. Figure 5.1 (top, left) shows the MAD over this two-year period versus the forecast horizon. The multi-scale DCMM has the lowest MAD for 12 of the 14 forecast horizons; however, the MAD of all three models are fairly similar. This similarity is most likely due to the fact that this low-selling item will have a median of zero or one sales on most days, so the average absolute deviation of the median will be very small on average. The rPIT plot and coverage plots for the Poisson GLARMA model are shown in Figure 5.2. In the rPIT plot, there is a slight S-shape to the rPIT values, indicating that the forecast distribution is slightly too narrow.

Figure 5.4 shows binary calibration plots for $1-$day forecasts of non-zero sales of item A in the Poisson GLARMA model. Well calibrated binary forecasts means that the crosses or red bars should fall within the shaded gray region. For item A, we see that there are three bars that all fall below the gray shaded region. Furthermore, although the other bars fall within the gray regions, we do notice a trend of the crosses falling under the diagonal line. This plot indicates that the Poisson GLARMA model overpredicted the probability of zero sales occurring.

Item B. Item B is a relatively high-selling item. Figure 5.1 (top, right) displays the MAD for each of the four models. We see that the DCMM models both have lower MAD than the GLARMA models. The decrease in MAD of the multi-scale DCMM versus the Negative Binomial GLARMA model ranges from 0.11 to 0.20. Over this two year period, this corresponds to between 86 and 144 units more accurate when using the multi-scale DCMM.

The rPIT plot and coverage plots for the Poisson GLARMA model are shown in the 1st row of Figure 5.3. In the rPIT plot, we see that the Poisson rPIT values deviate from uniformity. This indicates that the forecast distribution is too narrow in the Poisson GLARMA model. The Negative Binomial rPIT values appear to conform to uniformity. In the Poisson coverage plot, we see that there is undercoverage of forecast intervals. Overall, it appears that the Poisson GLARMA model does not sufficiently account for the overdispersion apparent in the data. In the Negative Binomial coverage plot, we see that the coverage lies along the 45-degree line. The Negative Binomial GLARMA model appears to perform well in terms of probabilistic forecasting.

Item C. The multi-scale DCMM has the lowest MAD for 13 of the 14 forecast horizons. When forecasting 14-days ahead, the multi-scale DCMM and Negative Binomial GLARMA model actually have the exact same MAD. The GLARMA models have lower MAD than the univariate DCMM across the forecast horizon. The largest differences in MAD occur for short term forecasting. For 1, 2, and $5-$step forecasting, the differences in MAD correspond to between 25 and 30 units more accurate for the multi-scale DCMM versus the Negative Binomial GLARMA model.

The rPIT plot and coverage plots for the Poisson GLARMA model are shown in the 2nd row of Figure 5.3. In the rPIT plot, the Poisson rPIT values deviate from uniformity slightly. Based on the rPIT values, it appears that the predictive distribution is too light on the lower end, and slightly too narrow on the higher end. The Negative Binomial rPIT values are closer to uniformity, however there is a slight deviation on the upper end. This may indicate that the upper tail of the predictive distribution is too long. The Poisson coverage plot shows very slight undercoverage for intervals above 80%. The Negative Binomial coverage plot shows slight overcoverage for the forecast intervals.

Item D. The multi-scale DCMM has the lowest MAD for all of the forecast horizons. However, the GLARMA models have lower MAD than the univariate DCMM. The differences in MAD are consistent across forecast horizons. Across the two-year period, this corresponds with between 46 and 64 units of accuracy gained by using the multi-scale DCMM versus the Negative Binomial GLARMA model. Similarly, this corresponds with between 23 and 43 units of accuracy gained by using the multi-scale DCMM versus the Poisson GLARMA model.

The rPIT plot and coverage plots for the Poisson GLARMA model are shown in the 3rd row of Figure 5.3. In the rPIT plot, we see that the Poisson rPIT values slightly deviate from uniformity. These values indicate that the forecast distribution in the Poisson GLARMA model is too narrow on the upper and lower tail. The Negative Binomial rPIT values are closer to uniformity, but do not conform exactly to the $45-$degree line. The Poisson coverage plot shows very slight undercoverage. The Negative Binomial coverage plot is very close to the $45-$degree line.

Item E. The daily sales of item E are shown in Figure 5.5. We attempted to run the forecasting analysis on this item for the Poisson and Negative Binomial GLARMA models. However, both of these models encountered errors during the time period we were evaluating. This item has a high probability of zero sales, but also frequent bursts of large sales. It appears that the GLARMA models are not appropriate for this time series.

Discussion of results

Overall, the performance of the GLARMA models is good for the items discussed here. The performance of the GLARMA models under MAD is quite competitive with the DCMMs for items A, C, and D. However, both of the DCMMs outperform the GLARMA models for item B. We believe one of the underlying reasons for this is that the daily sales of item B are more nonstationary and more variable over time than the other three items. This theory is supported by the fact that the GLARMA models encountered errors for item E, which displays quite non-stationary demand.

For probabilistic forecasting, the Poisson GLARMA model consistently produces forecast distributions that are just too narrow. The performance of the Negative Binomial GLARMA model was often improved.

One important consideration for the GLARMA models is that there is no separate model for zero sales like in the DCMM. One result of this is that the binary forecasting of zero-versus-non-zero sales appears to be miscalibrated for item A. While the GLARMA models performed well in these comparisons, we believe that the multi-scale DCMM is the more appropriate choice for our application to product sales forecasting. The main reason for this is that the DCMM is more automatically flexible and robust to handling the common characteristics of time series of counts. The DCMM can automatically handle time series with excess zeros, overdispersion, and non-stationary model components. The GLARMA models require individual item customization by fitting either a Poisson or Negative Binomial model, do not have a separate binary component, and appear to struggle with time series with apparent non-stationarities. Another key issue with the GLARMA framework is that it cannot be used to model and forecast for every item. This would require further customization for individual items which is not realistic in our context.

5.3 Autoregressive Conditional Poisson Model

An alternate name for this model is the integer-valued GARCH model (INGARCH) of order p and q. Details about these models are available in Heinen (2003), Ferland et al. (2006), and Fokianos et al. (2009). Let y_t denote a non-negative count valued time series observed over time $t = 1, \ldots, T$. The ACP model has counts following a Poisson distribution with an autoregressive mean, namely

$$y_t \mid \mathcal{D}_{t-1} \sim Po(\mu_t), \tag{5.4}$$

$$E[y_t \mid \mathcal{D}_{t-1}] = \mu_t = \omega + \sum_{j=1}^{p} \alpha_j y_{t-j} + \sum_{j=1}^{q} \beta_j \mu_{t-j}, \tag{5.5}$$

for positive $\alpha_j, \beta_j, \omega$. Here, $\mathcal{D}_{t-1} = \{y_1, \ldots, y_{t-1}\}$. In this model, the order p describes the number of lagged observations that affect the mean at time t. The order q represents the number of lagged values of Poisson mean that appear in the model. Since the Poisson mean μ_t is positive, the values of $\alpha_j, \beta_j, \omega$ are constrained to positive values. The most commonly used form of this model is the ACP(1,1).

The ACP(p, q) model is implemented in the `acp` R package. Built-in functions in this package estimate the ACP(p, q) model with covariates, and provide static forecast means. In the ACP(p, q) model, time dependence is achieved through the latent ARMA structure, but the ARMA coefficients and the covariate effects are static in time. This model accounts for overdispersion through the autoregressive Poisson mean. There is no simple way to account for excess/fewer zeroes, and there is no multivariate extension implemented. To apply this model in an on-line setting, we refit the model to all of the observed data at each time $t = 1, \ldots, T$.

For this comparison with the DCMM, we attempt to implement the ACP(1,0) and ACP(1,1) models. Each time we tried to run the ACP(1,0) model, the code returned an error – it appears that this model form is not supported by the current software. We focused our comparison on the ACP(1,1) model instead.

The built-in function `predict.acp` produces a static forecast over the next k days. At time t, when forecasting time $t+k$, this function will use the parameter estimates at time t, and requires the observed data $y_{t+1}, \ldots, y_{t+k-1}$. In our application, we are interested in multi-step ahead predictive distributions. In order to compare results with the DCMM, we have written our own function to produce simulations from the k-step joint forecast distribution. After fitting the ACP(1,1) model on data from time 1 to t, we use the `predict.acp` function to forecast the $1-$step forecast mean for time $t+1$. Conditioning on the mean, we simulate a Poisson valued prediction y^*_{t+1} from the implied forecast distribution for y_{t+1}. Treating this simulated y^*_{t+1} as synthetic data, we use the `predict.acp` function to forecast the mean at time $t+2$, and then draw a value of y^*_{t+2}. We repeat this procedure up to time $t+14$ to produce a single joint draw form the 1:14 step forecast distribution.

After developing this forecasting procedure, we refit the ACP(1,1) model at each day in the two year time period of interest. On each day, we use the described procedure to forecast k-steps ahead. However, for each of the four items A – D in the analysis, the ACP package encountered an error at some point during this two year period.

Another R package `tscount` implements the INGARCH model. We were able to implement the INGARCH model for some items with this package. The built-in function `predict.tsglm` produces 1:k step forecast means, medians, and prediction intervals. It is only possible to extract prediction intervals rather than the entire forecast distributions. Here, we focus on the coverage of 95% prediction intervals.

Item A. The MAD for the INGARCH model for item A is shown in Figure 5.6. The multi-scale DCMM has lower MAD across all forecast horizons. The decrease in MAD varies between 0.06 and 0.10. Across the two-year period, these decreases correspond with between 48 and 70 units of accuracy gained under the multi-scale DCMM.

The 95% prediction intervals for each forecast horizon show slight overcoverage of the observed data. Across the forecast horizons, the 95% intervals contain over 98% of the observed data.

Item B. The MAD for item B is shown in Figure 5.6. The multi-scale DCMM has lower MAD across all 14 forecast horizons. The decrease in MAD varies from 0.11 to 0.20 for the multi-scale DCMM versus the INGARCH model. Across the two-year period, these decreases in MAD corresponded to between 84-119 units more accurate under the multi-scale DCMM.

The 95% prediction intervals at each forecast horizon showed undercoverage of the observed data. The empirical coverage across the forecast horizon varied from 79:81%.

Item C. The MAD results for item C are shown in Figure 5.6. The multi-scale DCMM has lower MAD across all 14 forecast horizons. The decrease in MAD ranges from 0.37 to 0.52 for the multi-scale DCMM versus the INGARCH model. Across the two-year period, these decreases correspond to between 271 to 382 units of accuracy gained by using the multi-scale DCMM versus the INGARCH model.

The 95% prediction intervals at each forecast horizon are very close to the nominal coverage. Empirical coverage varies from 95% to 98% across the forecast horizons.

Item D. The MAD results for item D are shown in Figure 5.6. The multi-scale DCMM has lower MAD across all 14 forecast horizons. The decrease in MAD ranges from 0.28 to 0.40 for the multi-scale DCMM versus the INGARCH model. Across the two-year period, these decreases correspond to between 206 and 295 units of accuracy gained by using the multi-scale DCMM versus the INGARCH model.

The 95% prediction intervals at each forecast horizon are very close to the nominal coverage.

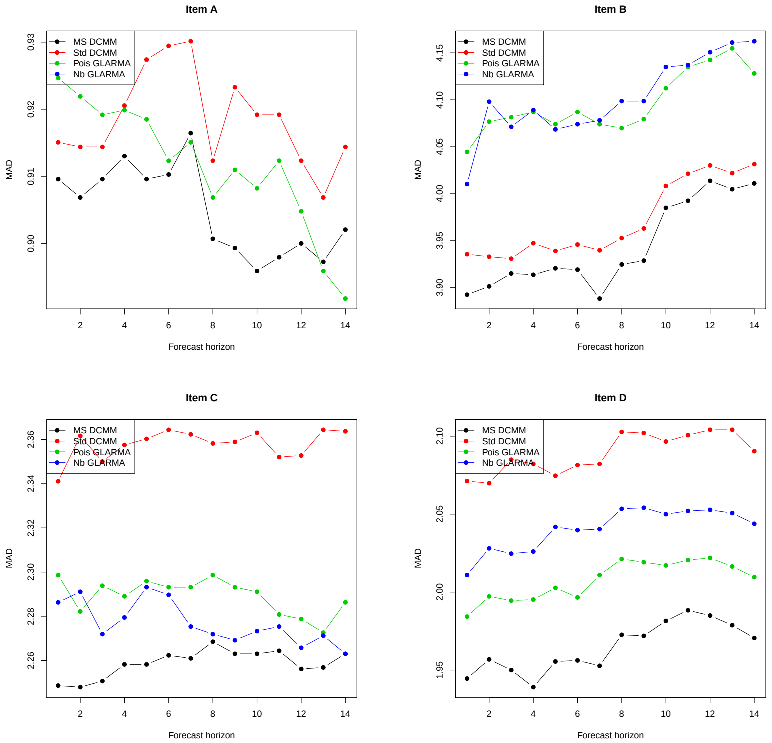
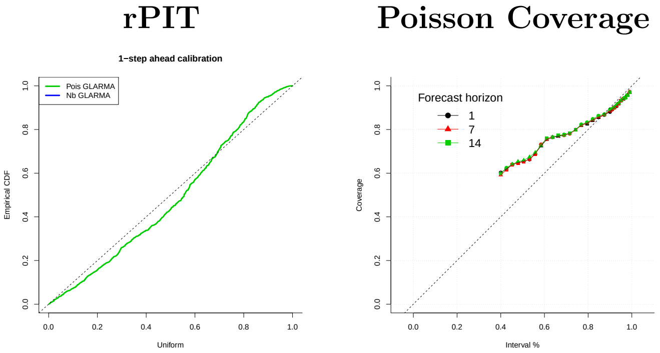
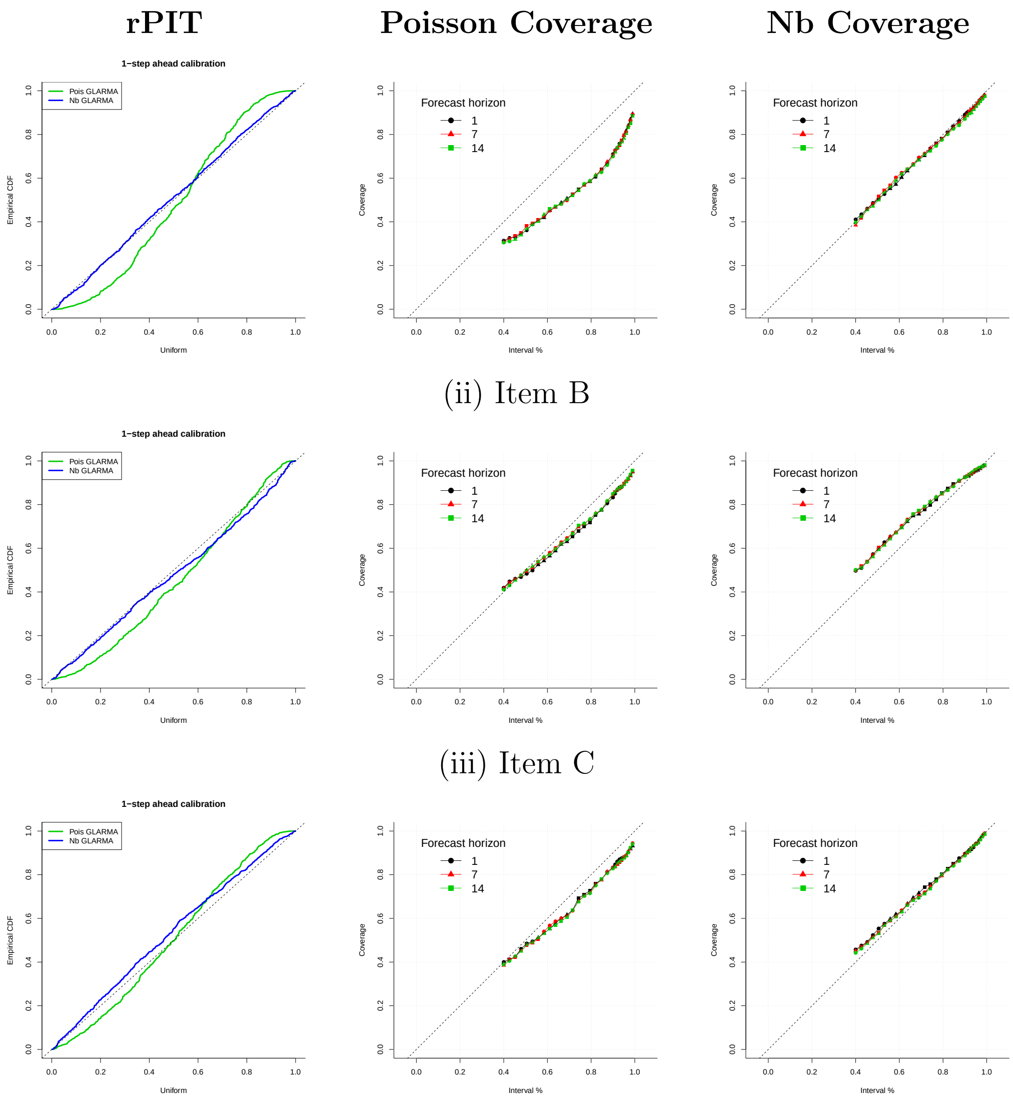


FIGURE 5.1: Forecast comparison in terms of MAD of items A – D of multi-scale (black) and univariate (red) DCMMs to the Poisson (green) and Negative Binomial (blue) GLARMAs.



(i) Item A

FIGURE 5.2: Probabilistic forecast evaluations for item A. The left column shows the rPIT plots for the Poisson (green) GLARMA model. The right column shows the coverage of HPD regions for the Poisson GLARMA model.



(ii) Item B

(iii) Item C

(iv) Item D

Figure 5.3: Probabilistic forecast evaluations for items B (1st row), C (2nd row), and D (3rd row). The left column shows the rPIT plots for the Poisson (green) and Negative Binomial (blue) GLARMA models. The middle column shows the coverage of HPD intervals for the Poisson GLARMA model. The 3rd column shows the coverage of HPD regions for the Negative Binomial GLARMA model.

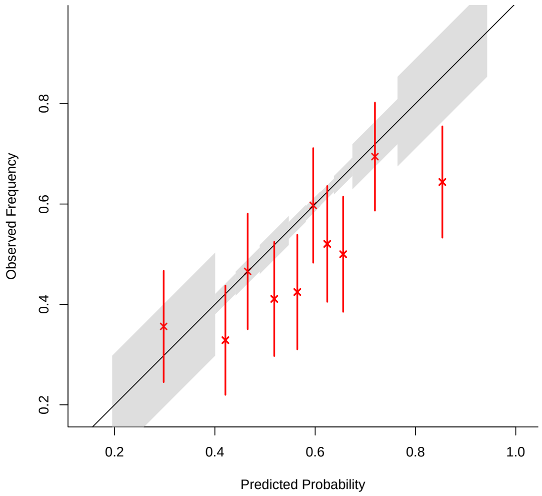


Figure 5.4: Binary calibration plots from 1-day ahead forecasting of non-zero sales of item A in the Poisson GLARMA model. Crosses mark observed frequencies in each bin, horizontal grey shading indicates variation of forecasts in each bin, and vertical bars indicate binomial variation based on the number of days in each bin.

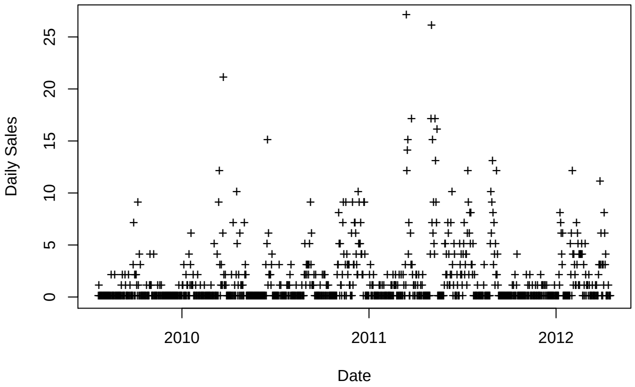


Figure 5.5: Daily sales of item E from mid 2009 to early 2012.

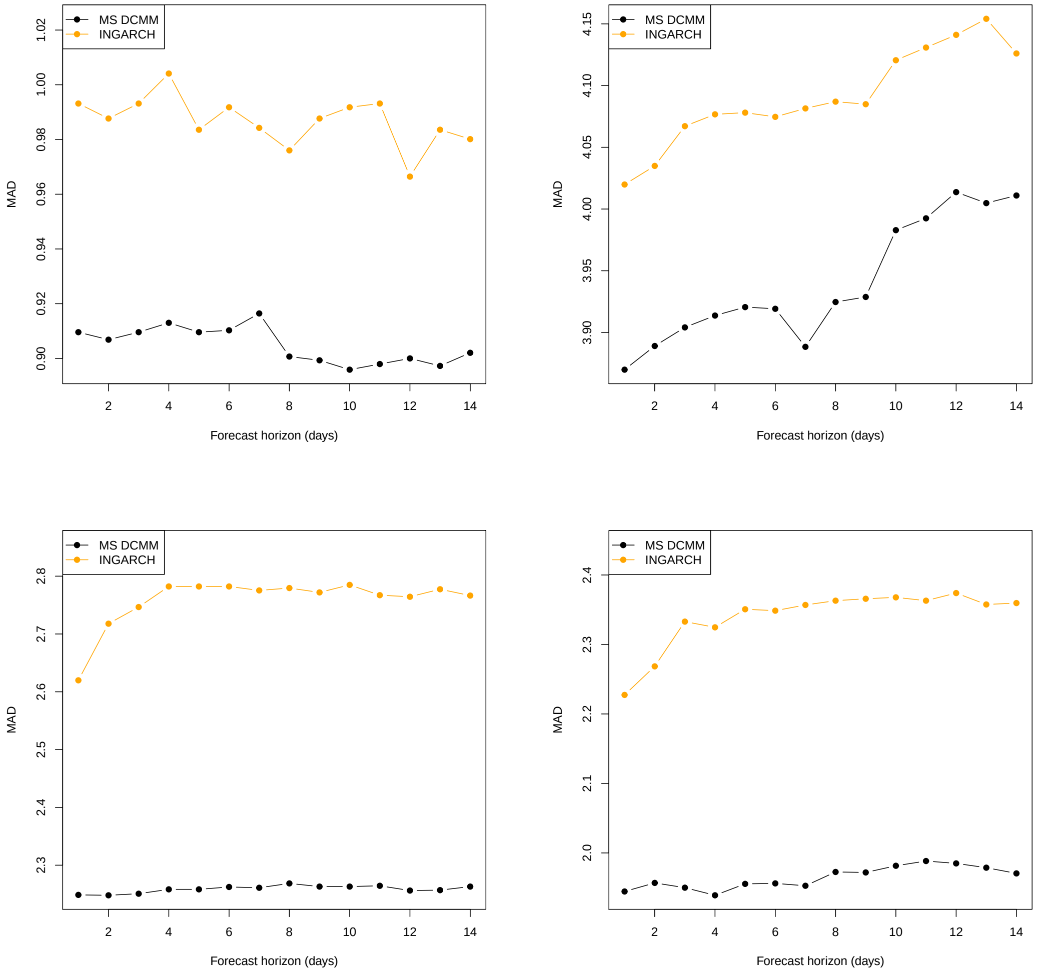


FIGURE 5.6: Forecast comparison in terms of MAD for item A,B,C,D for the multi-scale DCMM versus the INGARCH(1,0) model.

6

Dynamic Binary Cascade Model

6.1 Context and Models

6.1.1 Setting

The modeling advances in this work capitalize on availability of detailed point of sale data on transactions and sales-per-transaction information on supermarket items. Consider one specific item in a given store. Data are observed daily with day t records of (a) the number of transactions involving this item, i.e., of customers purchasing *some number* of the item, and (b) for each transaction, the number of units sold. Many items sell sporadically with no or few transactions per day, and with a high probability of only one unit sold per transaction. Many other items sell more frequently but again generally at 1 or perhaps 2 units per transaction. Then other items can sell at higher levels per transaction, though again generally small numbers. Infrequent bursts of item sales occur, often in the context of known promotions or pricing changes. Some items experience rare events in terms of larger numbers of sales in rare batch purchases.

Standing at the end of day t, the forecasting goal is to predict future sales over the coming period of k days; our applied context requires 2-week forecasts, so $k = 14$. We aim to do this in terms of a full probability forecast distribution for that coming period, and this process is repeated each day. The new model developed dissects and models item sales by transaction, with the following notation all indexed by day t:

- y_t is the total number of units sold.
- b_t is the number of transactions– or *baskets*– involving at least one unit sale.
- $z_t = \mathbb{1}(b_t > 0)$ where $\mathbb{1}(\cdot)$ is the indicator function; thus $z_t = 0$ implies zero transactions, while $z_t = 1$ indicates some transactions.
- $n_{r,t}$ is the number of transactions with *more than* r units, where $r = 0{:}d$ for some specified (small) positive integer d. By definition, $n_{0,t} \equiv b_t$. Evidently also, if $n_{r,t} = 0$ for some $r \leqslant d$ then $n_{r+1,t} = \cdots = n_{d,t} = 0$.
- $e_t \geqslant 0$ is the count of *excess sales* from any and all transactions that have more than d items. Evidently, $e_t = 0$ unless $n_{d,t} > 0$.
- With the above definitions, it follows that

$$y_t = \begin{cases} 0, & \text{if } z_t = 0, \\ \sum_{r=1:d} r(n_{r-1,t} - n_{r,t}) + e_t, & \text{if } z_t = 1. \end{cases} \tag{6.1}$$

The new dynamic models for forecasting the y_t series are built from coupled components separately modeling transactions $b_t = n_{0,t}$ and the sequence of values $n_{1:d,t}, e_t$, as now detailed.

6.1.2 Transaction Forecasting using Dynamic Count Mixture Models (DCMMs)

First, we utilize a dynamic count mixture model to represent and forecast the item-specific transaction process b_t over time. This class of DCMMs provide a flexible framework for modeling non-negative counts that is customized to dealing with zero counts together with potentially diverse patterns of variation of non-zero counts. Two state-space model components are involved. The first is a dynamic binary/logistic regression model for zero/non-zero transactions; the second is a dynamic, shifted Poisson log-linear model for transaction levels conditional on there being some transactions. Each model component may involve covariates– such as price and promotion predictors, seasonal effect variables, holiday effects, and so forth– that may partly explain and hence predict variation over time in transaction outcomes. An initiating application for the development of DCMMs was in forecasting item sales, and one important aspect of these models is that they naturally integrate time-specific random effects–e.g., daily random effects in the supermarket forecasting context. This anticipates and adapts to unpredictable levels of variation in outcomes over and above that explained by the conditional Bernoulli and Poisson dynamic models. In sales forecasting, this is particularly key in dealing with relatively common "extra-Poisson" variation and occasional bursts in sales levels.

The key point here is to adapt DCMMs to model transactions, not sales. The heterogeneity and over-dispersion seen in sales data is, in part, due to the compounding effect of varying sizes of transactions per customer throughout the day. When modeling transactions alone, this level of complexity and diversity in outcomes is diminished; the opportunity for improved forecasting accuracy at the level of transactions is then clear.

A DCMM for transaction outcomes b_t is defined by a coupled pair of observation distributions in which

$$z_t \sim Ber(\pi_t) \quad \text{and} \quad b_t|z_t = \begin{cases} 0, & \text{if } z_t = 0, \\ 1 + x_t, \quad x_t \sim Po(\mu_t), & \text{if } z_t = 1, \end{cases} \tag{6.2}$$

over all time t. Here $Ber(\pi)$ denotes the Bernoulli distribution with success probability π, while $Po(\mu)$ denotes the Poisson distribution with mean μ. The parameters π_t and μ_t are time-varying according to binary and Poisson dynamic generalized linear models (DGLMs: West and Harrison, 1997 chapter 15; Prado and West, 2010 section 4.4), respectively; that is,

$$\text{logit}(\pi_t) = \mathbf{F}_t^0 \boldsymbol{\xi}_t \quad \text{and} \quad \log(\mu_t) = \mathbf{F}_t^+ \boldsymbol{\theta}_t \tag{6.3}$$

with latent state vectors $\boldsymbol{\xi}_t$ and $\boldsymbol{\theta}_t$ and known dynamic regression vectors $\mathbf{F}_t^+$ and $\mathbf{F}_t^0$, in an obvious notation. The regression vectors can include different covariates and dummy variables, and the choices can be customized to item. Some aspects of variation over time– in both zero/non-zero transaction probabilities and in the conditional levels of non-zero transactions– comes through the specification of covariates in the regression vectors. Additional aspects of variation can be captured and adjusted for through time variation in the latent state vectors defining time-varying regression coefficients, in the usual state-space mode.

6.1.3 Dynamic Binary Cascade Models for Sales-per-Transaction

A central modeling and methodological innovation here is a new dynamic binary cascade model (DBCM) that directly addresses the interests in precision in dissecting heterogeneity in sales outcomes by focusing on an hierarchical decomposition of numbers of units per transaction. Many items sell just once per transaction, many others sell at perhaps 2 or 3 items, with higher numbers becoming increasingly rare. The multi-scale formulation of a DBCM is motivated by the reality that predicting rare events of any kind– here, larger numbers of units per transaction– is only and properly addressed using hierarchical sequences of conditional probabilities to define chances of outcomes.

The DCMM defines forecast distributions for transactions b_t into the future, and is used to compute predictive probabilities of transaction outcomes as well as– critically– to simulate representative future outcomes. Given a chosen or simulated/synthetic value of b_t, we then condition to model and forecast the daily sales conditional on that level of transactions using the DBCM defined below. In a Bayesian Monte Carlo analysis, repeatedly simulating many representative values of b_t and then sales coupled to each value defines formal computation from the required predictive distribution of sales. As we move across Monte Carlo samples, uncertainty about transaction levels is represented, and then the conditional uncertainty about sales per transaction factors in.

Consider then a given a value of $b_t \equiv n_{0,t}$. The DBCM defines a probability model for $y_t|b_t$. First, if $b_t = 0$ then sales $y_t = 0$, the trivial case. Consider now cases when $b_t > 0$ and refer to eqn. (6.2) to focus on uncertainty about the resulting sales count y_t. The model is structured as follows:

- For each $r = 1{:}d$, denote by $\pi_{r,t}$ the probability that the number of items sold per transaction exceeds r given that it exceeds $r-1$, and assume the numbers of units per transaction are conditionally independent across baskets.
- For any number $r = 1{:}d$, the (increasingly small) probability of more than r sales per basket is then implied as $\pi_{1,t}\pi_{2,t}\cdots\pi_{r,t}$.

  This is a key to the strategy and utility of the binary cascade concept: it models and hence forecasts rare events– unusually high levels sales for any one transaction– via a sequence of conditional probabilities, each of which is estimable from the data while their product can be very small.
- For each $r = 1{:}d$, the hierarchy of sales levels $n_{r,t}$ then follow a sequence of conditional binomial distributions, namely $n_{r,t}|n_{r-1,t} \sim Bin(n_{r-1,t}, \pi_{r,t})$ based on these probabilities. As we sequence through $r = 0, 1, \ldots,$ if we experience a level r with $n_{r,t} = 0$ this implies, of course, that $n_{j,t} = 0$ for all $j \geqslant r$.
- The excess sales e_t are computed by summing over possible transactions with more than d sales each. If $n_{d,t} = 0$, then $e_t = 0$. If, on the other hand, if $n_{d,t} > 0$ then $e_t \geqslant (d+1)n_{d,t}$.

> Given that the probability of more than $d+1$ per basket is generally expected to be quite small, the analysis will be quite robust to the conditional distribution of e_t. Hence we consider two strategies to quantifying the excess. One strategy is to leave the distribution of the excess completely unspecified and simply report the probability of $n_{d,t} > 0$ along with the forecast distribution of sales y_t *conditional on* $n_{d,t} = 0$. A second strategy is to simply use a bootstrap analysis in which a simulated forecast with $n_{d,t} > 0$ results in randomly sampling the corresponding forecast excess from the empirical distribution of past observed excess values. This is further discussed and developed in Sections 6.1.4 and 6.3.2, and exemplified in the application.

As with the Bernoulli model for zero/non-zero transactions z_t, we have access to the flexible class of dynamic logistic state-space models for each of the elements of the cascade across levels of sales per transaction. That is, the conditional model of $n_{r,t}$ has the dynamic binomial logistic form

$$n_{r,t}|n_{r-1,t} \sim Bin(n_{r-1,t}, \pi_{r,t}) \quad \text{where} \quad \text{logit}(\pi_{r,t}) = \mathbf{F}^0_{r,t}\boldsymbol{\xi}_{r,t} \tag{6.4}$$

with latent state vectors $\boldsymbol{\xi}_{r,t}$ and known dynamic regression vectors $\mathbf{F}^0_{r,t}$ in an obvious extension of the earlier notation. The regression vectors can include different covariates and dummy variables for each level r, and can be customized to level. The $\pi_{r,t}$ may be relatively stable over time, but impacted by price and promotion effects that increase relative probabilities of higher levels of sales per item, so that such information is candidate for inclusion in regression terms. As with the transaction events, aspects of variation over time comes through the covariates included, but is also potentially represented via time variation in the latent state vectors $\boldsymbol{\xi}_{r,t}$ of time-varying regression coefficients. Additional details of model specification and Bayesian filtering/forecasting analyses are summarized in Section 2.2.

6.1.4 Multi-Step Ahead Forecasting

Bayesian forecasting is based on full predictive distributions. In most applications, it is of interest to use direct/forward simulation of multi-step ahead predictive distributions. Among other things, this allows trivial computation of probabilistic forecast summaries for arbitrary functions of the future data over multiple steps ahead. In transactions and sale forecasting, generating Monte Carlo samples of synthetic futures over a series of days provides forecast summaries for sales each day, the patterns of variation and dependence day-to-day, and other aspects of applied relevance such as cumulative forecasts over a period of days. Thus, by "forecast" we now mean simulation – i.e., the generation of multiple random samples of transactions and sales outcomes over multiple days, defining "synthetic" futures that can be summarized to compute a range of point forecasts of interest under various utility functions, as well as full probabilistic summaries that formally capture and reflect predictive uncertainties.

Multi-step forecasting via simulation in dynamic transaction-sales models builds on basic simulations from the sets of DGLMs that define model components. On any day t looking ahead over the next k days based on current information $\{\mathcal{D}_t, \mathcal{I}_t\}$, the requirement is to generate a large Monte Carlo sample from the full Bayesian predictive distribution for transactions and sales of the item over days $t+1{:}t+k$. Denote by superscript $*$ a single Monte Carlo sample of relevant quantities, referred to as a "synthetic" outcome. We generate large Monte Carlo samples of outcomes by independently and repeatedly generating single synthetic outcomes as follows.

Forecast Transactions Indicators: Over coming days $j = 1{:}k$, generate the set of k synthetic transactions/no transactions indicators z^*_{t+j} from the binary DGLM component of the DCMM transaction model. This is a representative draw from the current $k-$dimensional predictive distribution of $(z_{t+1:t+k}|\mathcal{D}_t, \mathcal{I}_t)$.

Technically, this uses direct compositional sampling applying, at each day into the future, the forward filtering and updating analysis of the binary DGLMs. This exploits the representation

$$p(z_{t+1:t+k}|\mathcal{D}_t, \mathcal{I}_t) = p(z_{t+1}|\mathcal{D}_t, \mathcal{I}_t)\, p(z_{t+2}|z_{t+1}, \mathcal{D}_t, \mathcal{I}_t) \cdots p(z_{t+k}|z_{t+1:t+k-1}, \mathcal{D}_t, \mathcal{I}_t).$$

Outcomes are simulated by sequencing through the composition here. Sample z^*_{t+1} from the first component, simply the 1−step ahead distribution implied in the binary DGLM at time t. Condition on this value z^*_{t+1} to update the summary information in the DGLM, evolve one day and then predict z_{t+2} using $p(z_{t+2}|z^*_{t+1}, \mathcal{D}_t, \mathcal{I}_t)$; this is again just the 1−step ahead distribution in the binary DGLM moved along one day and conditional on the synthetic value z^*_{t+1}. This is recursively applied over the following days up to k−steps ahead to produce the full synthetic path $z^*_{t+1:t+k}$.

Forecast Non-Zero Transaction Levels: For each day ahead j such that $z^*_{t+j} = 1$, generate number of transactions $n^*_{0,t+j} = b^*_{t+j}$ from the shifted Poisson DGLM component of the DCMM transaction model. This gives a representative draw from the current conditional predictive distribution of $(b_{t+1:t+k}|z^*_{t+1:t+k}, \mathcal{D}_t, \mathcal{I}_t)$ with the implicit zero values implied on days such that $z^*_{t+j} = 0$.

Technically, this again uses direct compositional sampling, now based on the forward filtering and updating analysis of the Poisson DGLMs. The concept and format is just as in the above details for the binary DGLM, simply differing in the distributional forms involved.

Forecast Sales per Transaction: For each day ahead j for which $z^*_{t+j} = 1$, generate a set of basket sizes $n^*_{1:d,t}$ from the dynamic binary cascade model conditional on the number of transactions $n^*_{0,t+j} = b^*_{t+j}$. This gives a representative draw from the current conditional predictive distribution of the full sequence of baskets sizes $(n_{1:d,t+1:t+k}|b^*_{t+1:t+k}, z^*_{t+1:t+k}, \mathcal{D}_t, \mathcal{I}_t)$ with the implicit zero values implied on days such that $z^*_{t+j} = 0$. Technically, this is done by sequencing through the cascade on each day, generating the number of baskets with a single item, and conditional on that number simulating the number with two items, and so on up to d items. In cases when the total number of items simulated with fewer than $d+1$ items in any transaction reaches b^*_{t+j}, the implied synthetic number of items sold is established. Otherwise, the (generally few) remaining transactions involve more than d items each. If the excess distribution is unspecified, then the DBCM outputs the current synthetic probability of the excess sales event $e_{t+j} \geqslant (d+1)n^*_{d,t}$.

If the excess distribution in the DBCM has been specified, we can proceed by simulating from this excess distribution. One specific excess distribution that fits nicely in the compositional forecasting framework is simulating the excess sales from the empirical excess distribution up to time t. For example, prior to time t, assume we have observed excess sales-per-transactions of $(d+1, \ldots, D)$ with frequencies $(w_{d+1}, \ldots, w_D)$, where $\sum_{i=d+1}^{D} w_i$ is the total number of transactions with $n_{d,t} > 0$. Given $n_{d,t+k} > 0$, we can forecast the future excess sales e_{t+k} by sampling $n_{d,t+k}$ values with replacement from $(d+1, \ldots, D)$ with weight proportional to $(w_{d+1}, \ldots, w_D)$.

As with the transactions simulations above, moving ahead over days involves direct compositional sampling, now based on the forward filtering and updating analysis of the sets of conditional binomial DGLMs. The concept and format is just as in the above details for the binary DGLM, simply differing in the distributional forms involved.

Uncertainty about the underlying DGLM model components are fully accounted for in forward simulation of each of the state vectors. Critically also, each such synthetic outcome inherently reflects day-to-day dependencies as well as uncertainties about the underlying DGLM model state vectors; that is, we generate full predictive samples from the joint distribution of the binary, Poisson and binomial latent transactions and sales variables over the $k-$step ahead path. This means that summary inferences on aggregates and other functions of transactions indicators, transactions levels, basket sizes and sales can be directly deduced by simple numerical summaries of the set of Monte Carlo samples.

6.2 Cross-Series Linkages and Multi-Scale Extensions

In forecasting multiple items with potentially related patterns over time, the opportunity to improve forecast accuracy by integrating information across series arises. Introduced in Chapter 3 in DCMMs for sales forecasting, an approach using dynamic predictors related to cross-series relationships is relevant to potentially both DCMM and DBCM components of the new transaction-sales models here. The basic idea is to define one or more factors to be used as common predictors in the dynamic regression models for each item. This is summarized here in the context of a single DGLM component for each of a collection of (possibly many) time series. Let N be the number of time series and denote by $\mathcal{M}_i$ a DGLM component for series i. In the transactions-sales applications, this can be any one or each of the component binary, binomial and (shifted) Poisson DGLM components. One particularly relevant context is to share information about related patterns of daily variation withing the week, i.e., weekly seasonal patterns, in which case the DGLM component $\mathcal{M}_i$ is the shifted Poisson for non-zero transactions for item i at the daily level.

A multivariate dynamic factor model incorporating cross-series linkages has state and regression vectors defined by

$$\mathcal{M}_i: \qquad \boldsymbol{\theta}_{i,t} = \begin{pmatrix} \boldsymbol{\gamma}_{i,t} \\ \boldsymbol{\beta}_{i,t} \end{pmatrix}, \quad \mathbf{F}_{i,t} = \begin{pmatrix} \mathbf{f}_{i,t} \\ \boldsymbol{\phi}_t \end{pmatrix}, \qquad i = 1{:}N, \tag{6.5}$$

with subvectors of conformable dimensions; the linear predictor is then $\lambda_{i,t} = \boldsymbol{\gamma}'_{i,t}\mathbf{f}_{i,t} + \boldsymbol{\beta}'_{i,t}\boldsymbol{\phi}_t$. Here $\mathbf{f}_{i,t}$ contains constants and series-specific predictors– such as item-specific prices and promotions in the sales forecasting context. The latent factor vector $\boldsymbol{\phi}_t$ is common to all series– such as seasonal or brand effects in the sales forecasting context. Each series has its own state component $\boldsymbol{\beta}_{i,t}$ so that the impacts of common factors are series-specific as well as time-varying.

A separate model depends on $\boldsymbol{\phi}_t$ and possibly other factors. Denote this model by $\mathcal{M}_0$. Forward sequential analysis of data relevant to $\mathcal{M}_0$ defines posterior distributions for $\boldsymbol{\phi}_t$ at any time t that can be used to infer and forecast the $\boldsymbol{\phi}_t$ process as desired. These inferences on the common factors are then forwarded to each model $\mathcal{M}_i$ to use in forecasting the individual series. Technically, this is done via direct simulation, so that current and future values $\boldsymbol{\phi}_*$ are simulated from the current posterior and predictive distributions under $\mathcal{M}_0$, and then forwarded to each $\mathcal{M}_i$. At each simulated value, each single posterior and forecast simulation in $\mathcal{M}_i$ conditions on one sampled $\boldsymbol{\phi}_*$, so that inferences under $\mathcal{M}_i$ are then available using the standard computations for individual models. Critically, the updates and forecasting computations in each $\mathcal{M}_i$ are performed separately and in parallel, conditional on values of the common factors $\boldsymbol{\phi}_*$; this decoupling of series for core computations enables scaling in the number N of items, while maintaining the information sharing across items.

Model $\mathcal{M}_0$ can be any external model generating information on common factors. Key special cases relevant to DCMMs for transactions are referred to as multi-scale models. This is highlighted in cases of collections of items within a store that naturally share common patterns of weekly seasonality based on customer traffic through the store. In such cases, $\boldsymbol{\phi}_t$ may be a scalar factor representing the current day-of-week based on an external model of traffic. The multi-scale special case arises when using aggregate transaction data– such as the total number of transactions on all products, or on some specific subgroup of products– to define $\mathcal{M}_0$. Each item-level model is then built on the predictions about daily variation from the aggregate model, while the elements $\boldsymbol{\beta}_{i,t}$ provide for item-specific, idiosyncratic deviations from the imputed aggregate values.

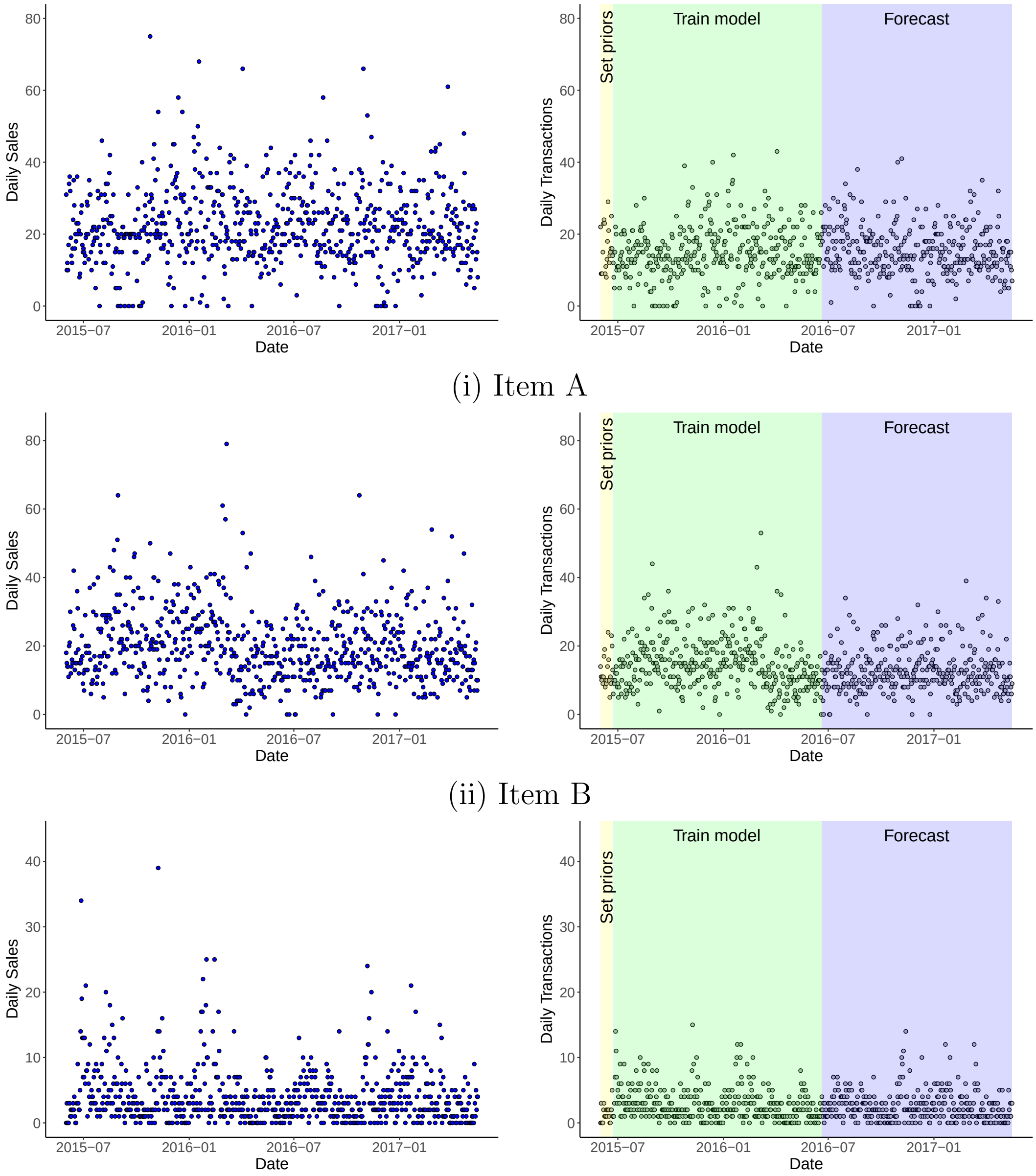
6.3 An Example in Forecasting with the DBCM

6.3.1 Data

The goal of our case study is to predict future sales of individual supermarket items 1:14−days ahead. We compare the forecasting performance of the binary cascade framework to a benchmark model; the latter is a DCMM for daily sales as in Chapters 3 and 4. This benchmark meets key desiderata of defining full predictive forecasts, flexibility in modeling diverse patterns in series of counts, incorporation of potentially time-varying dynamic seasonal and regression effects, and adaptability to heterogeneous patterns of otherwise unpredictable variability.

The data set records transaction-level purchases of supermarket items in one store of a major retail chain during the 762 day period from June 1st 2015 to July 1st 2017. Each row in the transaction-level data set represents one consumer's purchase of one or more units of a single item. Items are identified by a unique base universal product code (UPC) in the "Dry Noodles and Pasta" category. For each transaction event, the data includes item UPC, the purchase date, the effective price per unit, whether or not the item was purchased on promotion, and the unit sales in the given transaction. The daily transactions count for an item is the number of rows on a given day with the item's UPC; the total daily sales is then the sum of unit sales across transactions.

We explore forecasting of three spaghetti items to illustrate the potential improvements offered by decomposing heterogeneity into transactions and sales-per-transaction. These items represent a range of transactions-sales patterns and typify the features of data across many items. Table 6.1 reports summaries of the daily transactions and sales-per-transaction for item A,B and C. Figure 6.1 displays the daily transactions and sales for each item to illustrate the diminished diversity of item-level daily transactions in comparison to daily sales. Within this chosen cat-



(i) Item A

(ii) Item B

(iii) Item C

Figure 6.1: Daily sales and transactions of three spaghetti items (A-C) sold in one store from June 1st 2015 to July 1st 2017.

egory and store, items A and B are moderate to high selling items, and item C is a relatively low-selling item. Each item's daily sales and transactions share similar features such as the overall level and trends over time, and the evident day-of-week effect. Both series also share the feature of somewhat rare extreme values, although the diminished variability of the transaction data is evident.

Table 6.1: Some summaries of daily transactions and sales-per-transaction data for 3 spaghetti items.

	Daily transactions			Sales-per-transaction		
Item	Mean	Median	Variance	Mean	Median	% < 5
A	22.84	21	100.52	1.46	1	98.9
B	19.75	18	101.15	1.44	1	99.0
C	4.66	3	18.70	1.53	1	98.4

6.3.2 Model Specification

Transactions DCMM Specification

As described in Section 6.1.2, the DBCM framework utilizes a DCMM to forecast daily transactions. In this analysis, we consider two DCMMs for forecasting transactions: independent DCMMs with item-specific weekly seasonal effects, and a multi-scale DCMM that shares information on the weekly seasonal effect across all spaghetti items. The same form of DCMM is specified in the independent DBCM framework and the benchmark DCMMs on daily sales. In these independent DCMMs, each Bernoulli and conditionally Poisson component includes a local level, a full Fourier form seasonal component with period 7, and a regression component with log price and a binary indicator of promotions as predictors. Each binary and conditionally Poisson DGLM can be defined through regression vectors and state evolution matrices of the form

$$\mathbf{F}_t' = \big(1,\, \log(\text{price}_t),\, \text{promo}_t,\, 1, 0,\, 1, 0,\, 1, 0\big)\,,$$

$$\mathbf{G}_t = \text{blockdiag}[1,\, 1,\, 1,\, \mathbf{H}_1,\, \mathbf{H}_2,\, \mathbf{H}_3],$$

with

$$\mathbf{H}_j = \begin{pmatrix} \cos(2\pi j/7) & \sin(2\pi j/7) \\ -\sin(2\pi j/7) & \cos(2\pi j/7) \end{pmatrix}, \quad j = 1{:}3.$$

where price_t is the item-specific price on day t, and promo_t is equal to 1 if the item is on promotion on day t, and 0 if not. Through the standard use of discount factors, each component is dynamic, allowing for time variation in the level, weekly seasonality, and price and promotion effects. Based on previous analyses of item-level sales and transactions, we set fixed discount factors of 0.99 (Poisson) and 0.999 (Bernoulli) on each component.

For the multi-scale DBCM framework, we specify item-level models $\mathcal{M}_i$ with $\mathbf{f}'_{i,t} = \big(1, \log(\text{price}_{i,t}), \text{promo}_{i,t}\big)$, and a $7-$dimensional $\boldsymbol{\phi}_t$ with one non-zero element representing the current day-of-week effect. In this multi-scale analysis, $\mathcal{M}_0$ is a dynamic linear model (DLM) on the aggregate log daily transactions of all spaghetti items in the chosen store. This aggregate DLM includes a local linear trend, the scaled log average spaghetti price as a predictor, and full Fourier form seasonal components of periods 7 and 365 representing the weekly and yearly seasonal effects. We allow for dynamic level, trend, regression effects, and seasonality with discount factors of $\delta = 0.995$ for the trend and regression components, $\delta = 0.999$ for each of the seasonal components, and $\beta = 0.999$ for the residual stochastic variance process. Predictive performance in all sales/transactions DCMMs is evaluated across a range of random effects discount factors, $\rho \in (.2, .4, .6, .8, 1)$.

The shading in Figure 6.1 indicates analysis set-up. For each DCMM and the aggregate DLM, initial priors using three weeks of training data (yellow shading). For the aggregate lognormal DLM and the conditionally Poisson DGLMs, we define approximate prior moments for the state vectors based on the posterior moments in a standard reference analysis of a Bayesian linear model of the log daily sales/transactions. For the binary DGLMs, we estimate the prior mean of the level to be $\log(p/(1-p))$, where p is the observed proportion of the first 21 days with at least one transaction. All other prior means in the binary DGLM are set to zero, with the prior covariance matrix as the identity. The green shaded region in Figure 6.1 denotes the one year period beginning on day 22 (denoted $t = 1$) in which our models are trained. After this one year period, in the blue shaded region, forecasting 1:14-days ahead is performed on each of the 332 days.

Binary Cascade Model Specification

Based on an exploratory analysis of typical sales-per-transaction, we set $d = 4$ for all items in this analysis. As seen in Table 6.1, around 99% of all transactions of the chosen items include four or fewer unit sales. The form of the binomial logistic DGLMs is the same across items and for all $r = 1{:}d$. Each conditional model of $n_{r,t}$ includes a dynamic local level, and a static regression component with a binary indicator of promotion as a predictor. Each binomial DGLM allows for slow time variation in the level through a discount factor of $\delta = 0.999$. In previous analyses, we found a static promotional effect, with $\delta = 1$, to be sufficient. For each binomial logistic DGLM, we specify

$$\mathbf{F}^0_{r,t} = \big(1,\, \text{promo}_t\big)' \qquad \text{and} \qquad \mathbf{G}_{r,t} = \mathbf{I}$$

where the promo_t is an item-specific indicator of a promotion at time t. Again, we use three weeks of training data to specify the prior mean of the level. In a logistic model of $\pi_{r,t}$, we set the prior mean of the level to be $\log(p/(1-p))$ where p is the proportion of transactions with exactly r unit sales out of all transactions with at least r unit sales. We set the prior mean of the promotion coefficient to be zero, and the prior covariance matrix for the state vector to $0.1\mathbf{I}$.

Excess Distribution

We consider two perspectives: leaving the excess distribution completely unspecified, or bootstrapping from the empirical excess distribution. In this context of daily sales forecasting, unpredictable and relatively rare situations may arise where, for example, a consumer purchases dozens or hundreds of units in a single bulk order. Due to lack of relevant data and predictors that would make modeling these rare outcomes possible, it is often preferable to leave the tail of the sale-per-transaction distribution unspecified. However, without constraints or assumptions on the excess distribution, we are limited in the conclusions we can make about the predictive distribution. At time $t-1$, the 1-step forecast density of y_t is

$$p(y_t \mid \mathcal{D}_{t-1}, \mathcal{I}_{t-1}) = q_t f(y_t) + (1-q_t)p_d(y_t)$$

where: (i) $q_t = Pr(n_{d,t} > 0)$ is the probability that $n_{d,t} > 0$, i.e., that *some* of the transactions have more than d units; (ii) $f(y_t)$ is the p.d.f. of the sales distribution given that $n_{d,t} > 0$; and $p_d(y_t)$ is the p.d.f. of the (specified) distribution given that $n_{d,t} = 0$. The forecast p.d.f.s for multi-steps ahead have similar forms. If $f(\cdot)$ is unspecified, we cannot exactly identify the mean or quantiles of the distribution. It is possible to identify lower/upper bounds for any quantile of the forecast distribution, including the median, but without additional assumptions about f, bounds on the mean of the forecast distribution are not available.

The second perspective is to utilize the empirical distribution of excess sales over a past period of time. Simulating excess sales-per-transaction from the empirical excess distribution results in access to the entire predictive distribution through Monte Carlo samples. With this approach, we can report any quantity of interest from the forecast distribution. Since forecasters are often interested in the accuracy of many different error metrics (and the corresponding optimal point forecasts), we present the results of the DBCM models using the empirical excess distribution. A potential downside of this approach is that the only possible values of sales-per-transaction are those that have previously been observed; that excesses are very rare ameliorates this concern. Other specifications that may be of utility are noted in the concluding section.

6.3.3 Examples and Evaluations

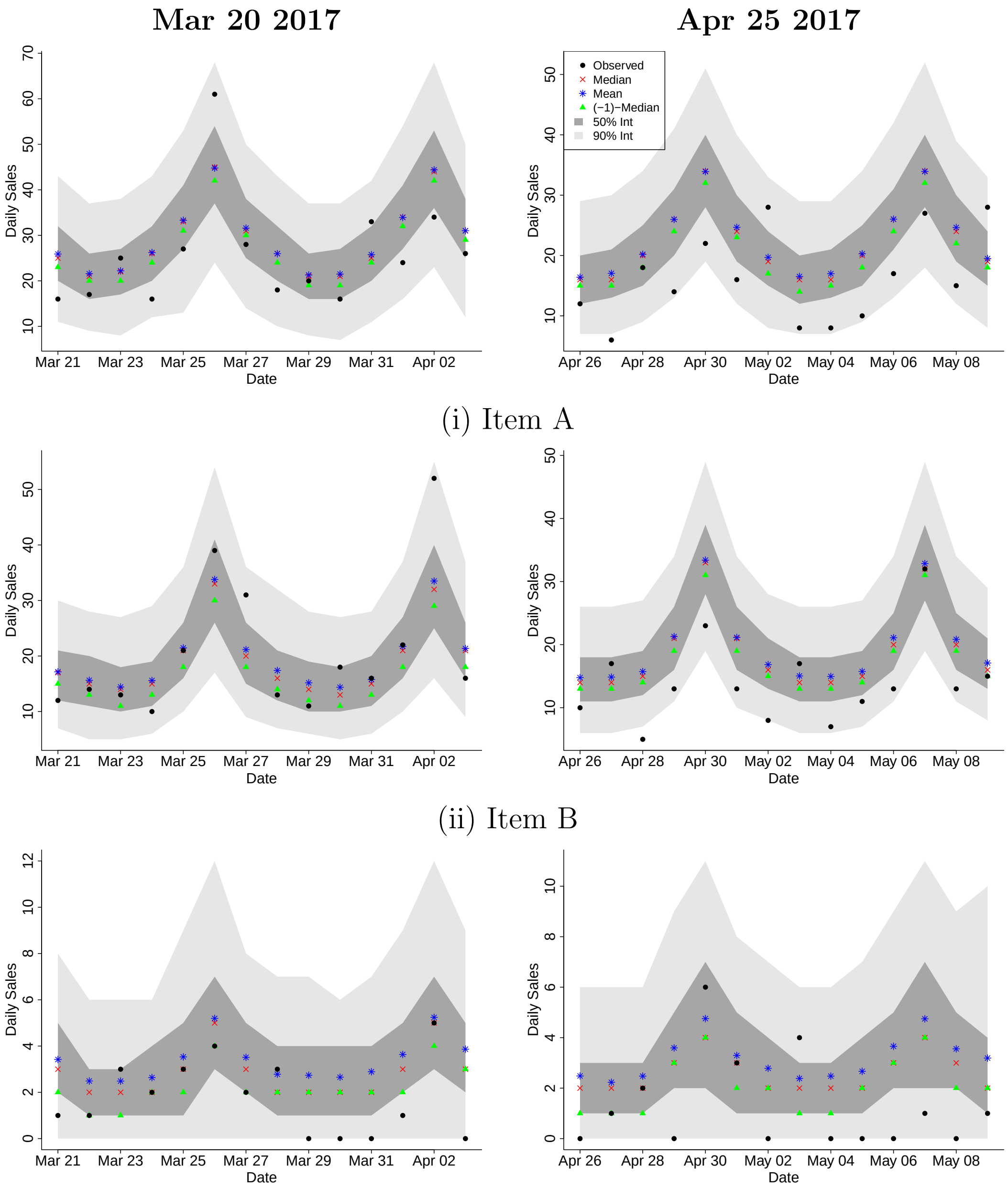
Joint Forecast Trajectories and Probabilistic Evaluation

Example forecast trajectories from this analysis are shown in Figure 6.2. These plots illustrate 1:14-day ahead joint forecasts on two days, Mar 20th 2017 (left column) and Apr 25th 2017 (right column). For each item, these forecasts were generated from the multi-scale binary cascade model, and the excess sales was drawn from empirical excess distribution. The displayed forecasts from the DBCM model are based on transaction forecasts from a DCMM with a random effects discount factor of $\rho = 1$. These plots provide insight into the spread of the forecast distribution $(50, 90\%$ credible intervals in gray shading), as well as the location of common point forecasts (mean, median, and (-1)-median). Observed daily sales are shown as black circles.

In general, forecasts made on Mar 20th were accurate in terms of location and spread. For item A, 7/14 days are contained in the 50% credible intervals, and 14/14 in the 90% intervals. For item B, the 50% intervals contain 11/14 days, and the 90% intervals contain 14/14 days. For item C, the 50% intervals contain 8/14 days, and the 90% intervals contain 14/14 days. On Apr 25th, the point forecasts are somewhat over-estimates, while 50% intervals show some under-coverage. For items A, B, and C, the 50% intervals contain only 2/14, 4/14, and 5/14 days, respectively. However, 90% intervals for each item are more accurate, containing 13/14, 13/14, and 14/14 observations, respectively. These trajectories simply provide snapshots of forecasts on two single days, to highlight the underlying forecasting process; coupled with this, we now evaluated aspects of longer-term forecasting performance.

Figure 6.3 (left column) displays coverage of the forecast distributions for 1, 7, and 14-day ahead forecasts for each item. These plots show the empirical coverage obtained over the 322-day forecast period for predictive credible intervals (HPD - highest posterior density) of different percentages. Ideally, the empirical coverage of our credible intervals is close to the nominal level, resulting in coverage close to the $45-$degree line. For item A, the empirical coverage of credible intervals is close to the nominal coverage, although there is some evidence of slight under-coverage. For example, empirical coverage of 1-step ahead 65% credible intervals is about 60%. For item B, the empirical coverage of credible intervals is close to the nominal coverage. For 5% and 20% credible intervals, there is of slight over-coverage and for 65% and 80% intervals, there is slight under-coverage. For item C, forecast intervals have slight over-coverage. For example, the empirical coverage of 1-day ahead 65% intervals is about 71%.

Figure 6.3 (right column) displays randomized probabilistic integral transform (PIT; Kolassa, 2016) values. If count valued data y is forecast with predictive c.d.f., $P(\cdot)$, define $P(-1) = 0$ and draw a random quantity $p_y \sim U(P(y-1), P(y))$ given the observed value of y. Over repeat forecasts, an ideal model would generate values of p_y that are approximately uniformly distributed. Figure 6.3 plots ordered randomized PIT values for 1:14-day ahead forecasts versus uniform quantiles. For item A, the values appear relatively uniform. Slight dips below the 45-degree line could be random variation, or may indicate that the lower tail of the forecast distribution is too light. For item B, randomized PIT values appear to closely reflect uniform quantiles. For item C, randomized PIT values are close to uniformity; there are small dips below the $45-$degree line that could reflect random variability, or slightly underweight lower tails of forecast distributions.

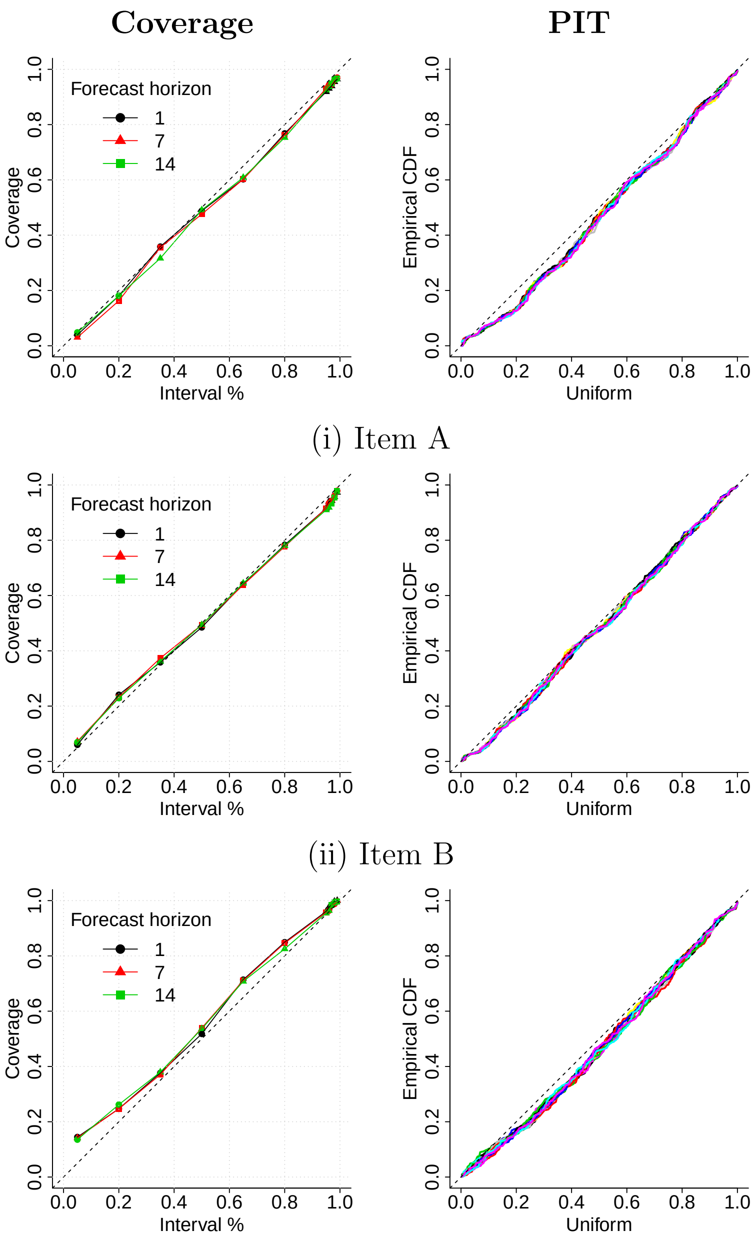


(i) Item A

(ii) Item B

(iii) Item C

FIGURE 6.2: 1-14 day joint forecast trajectories on Mar 20th 2017 (left) and Apr 25 2017 (right). Observed daily sales shown as a circle, forecast median as an x, forecast mean as a diamond, and forecast (-1)-median as a triangle. Light and dark shading indicate the forecast 50 and 90% credible intervals, respectively.

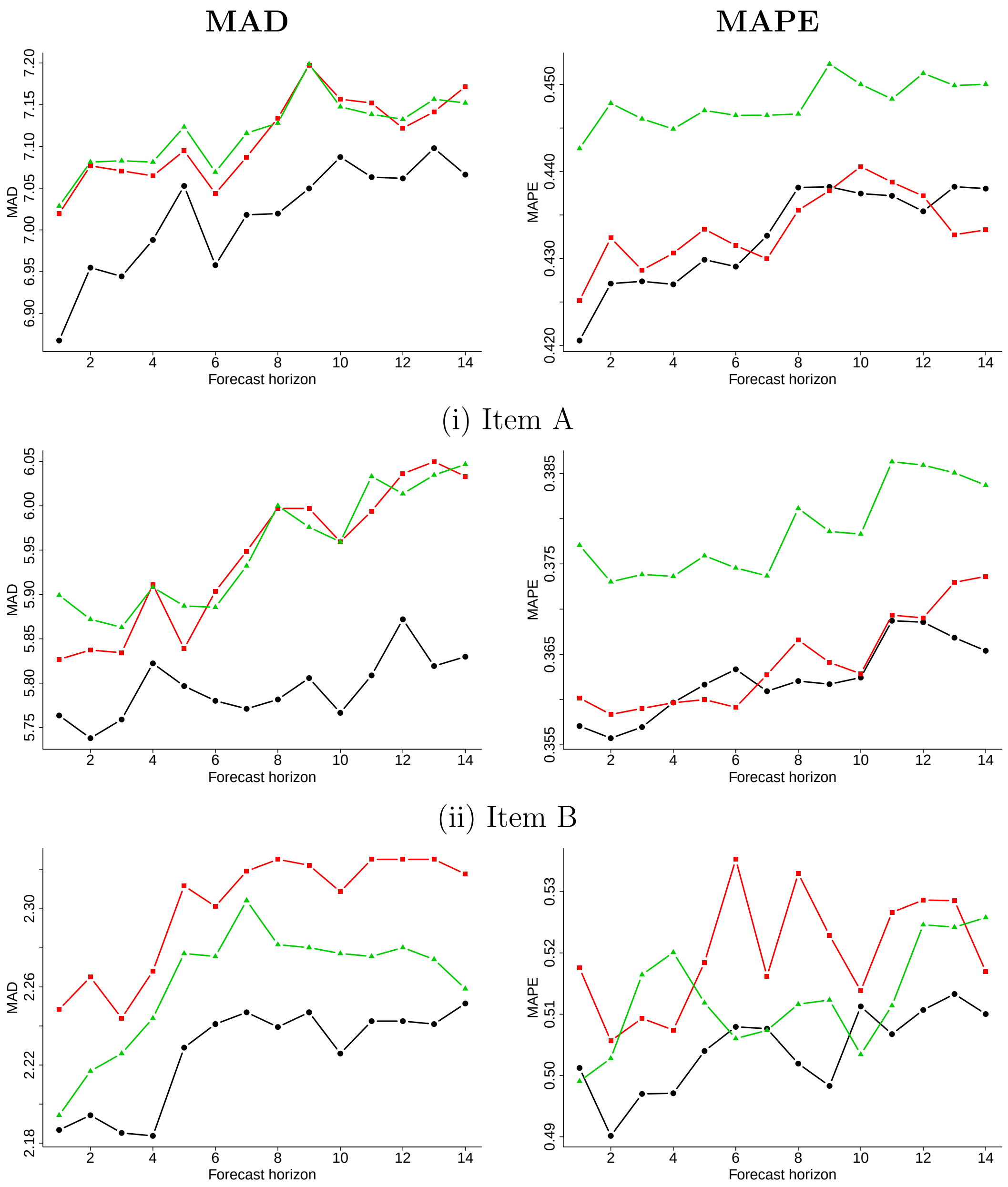


(i) Item A

(ii) Item B

(iii) Item C

FIGURE 6.3: Empirical coverage plots (left) for $1, 7, 14$-step forecasts and randomized PIT plot (right) for 1-14 step forecasts of items A (top), B (middle), and C (bottom) using the multi-scale DBCM with empirical excess distribution and random effect discount factor of $\rho = 1$.



(i) Item A

(ii) Item B

(iii) Item C

FIGURE 6.4: Mean absolute deviation (MAD: left) and mean absolute percentage error (MAPE: right) vs forecast horizon (days) for items A (top), B (middle), and C (bottom) from the multi-scale DBCM (black circles), independent DBCM (red squares), and independent DCMM (green triangles).

Point Forecasts

Error metrics for selected point forecasts are shown in Figure 6.4. We focus on two standard point forecast metrics, the mean absolute deviation (MAD) and the mean absolute percentage error (MAPE). Metrics are specific to a chosen lead-time $k > 0$. For a series y_t, denote by f_{t+k} a forecast of y_{t+k} made at time t. MAD is the time average of the absolute deviation, $|y_{t+k} - f_{t+k}|$, and the optimal point forecast is the k-step ahead predictive median. MAPE, a common error metric in demand forecasting, is simply the time average of $|y_{t+k} - f_{t+k}|/y_{t+k}$, and the optimal point forecast is the k-step predictive (-1)-median. The (-1)-median of a distribution $f(y)$ is the median of $g(y)$ where $g \propto f(y)/y$. When evaluating the chosen error metrics, we use the corresponding optimal point forecast from each model. For each metric, we evaluate the error across 1:14 days ahead on each day. The benchmark DCMM and both DBCM models (multi-scale and independent) are evaluated across a range of DCMM random effect discount factors, $\rho \in \{.2, .4, .6, .8, 1\}$. The accuracy of forecasting under each random effect may depend on the forecasting horizon, so we report only the lowest error across each of the five discount factors. Figure 6.4 displays the error from the best baseline DCMM, independent DBCM, and multi-scale DBCM across item, forecasting horizon, and metric.

Comparisons under MAD:

A: The multi-scale DBCM has the lowest MAD across the entire forecast horizon. Across the forecast horizon, the multi-scale DBCM has an average 1.4% decrease in MAD compared to the DCMM. The multi-scale DBCM results in the largest percentage decreases in MAD for short- and mid-range forecasts of $1-3$ and $6-9$ days ahead. The independent DBCM and DCMM have similar MAD performance.

B: The multi-scale DBCM has the lowest MAD across the entire forecast horizon. Across the forecast horizon, the multi-scale DBCM has a average of a 2.6% decrease in MAD compared to the DCMM. The largest percentage decreases in MAD occur for mid- to long-range forecasts of $7 - 14$ days ahead. The independent DBCM and DCMM have similar MAD performance.

C: The multi-scale DBCM has the lowest MAD across the entire forecast horizon. Across the forecast horizon, the multi-scale DBCM has a average of a 1.6% decrease in MAD compared to the DCMM. The multi-scale DBCM has the largest percentage decrease in MAD in mid-range forecasts of $3, 4, 5, 7, 8,$ and 10-days ahead. The DCMM has lower MAD than the independent DBCM across the entire forecasting horizon.

Comparisons under MAPE:

A: The multi-scale and independent DBCMs have lower MAPE across the entire forecast horizon. Across the forecast horizon, the multi-scale DBCM had an average decrease in MAPE of 3.4% compared to the DCMM. The largest percentage drops in MAPE occurred for shorter-term forests from $1 - 6$ days ahead.

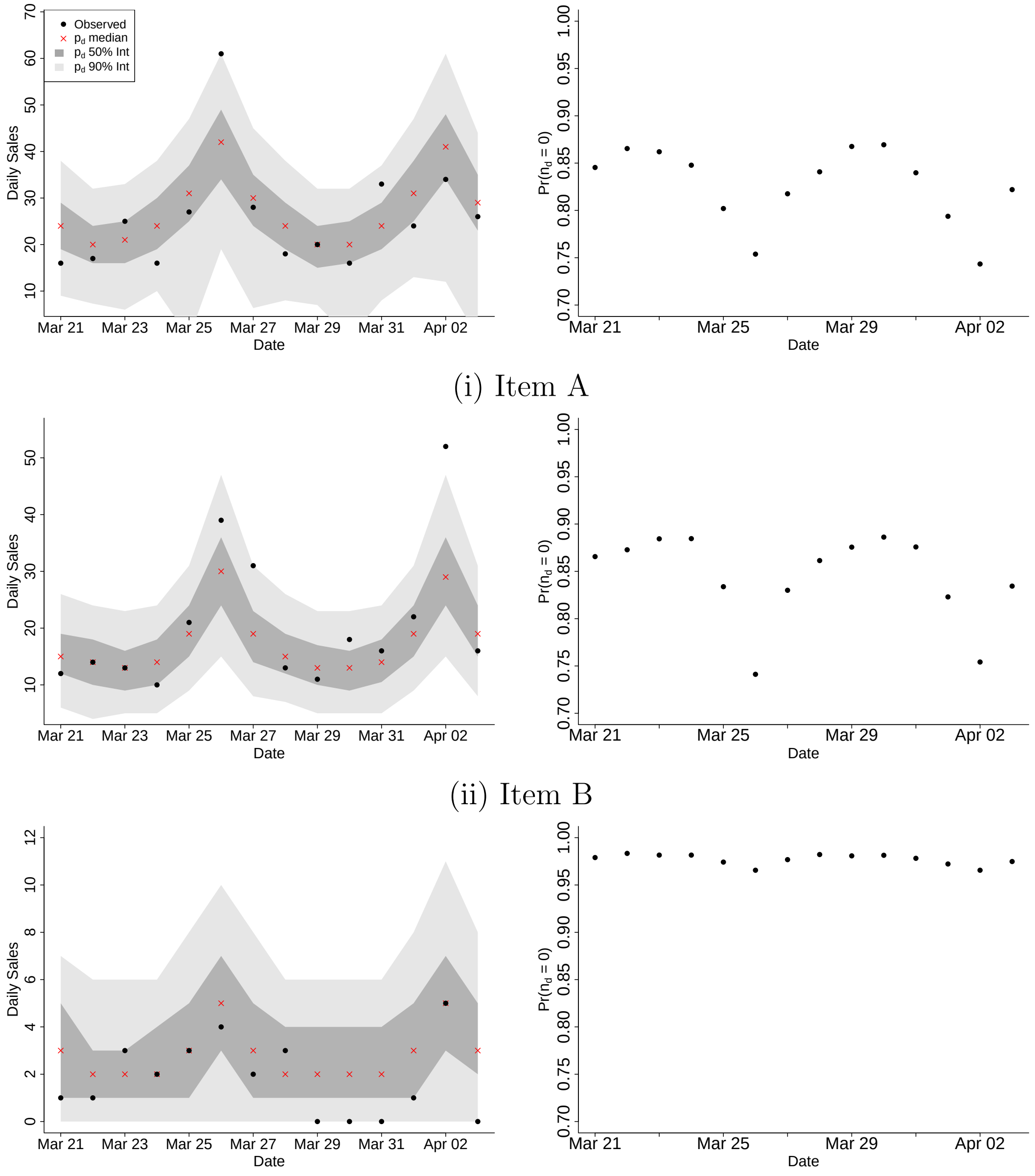
B: The multi-scale and independent DBCMs have lower MAPE across the entire forecast horizon. Across the forecast horizon, the multi-scale DBCM had an average decrease in MAPE of 4.3% compared to the DCMM. The largest percentage drops in MAPE occurred sporadically when forecasting $1, 2, 8, 11, 13,$ and 14-days ahead.

C: The multi-scale DBCM has the lowest MAPE for 10 of 14 forecast horizons. Across the entire forecast horizon, the multi-scale DBCM had an average decrease in of 1.6% compared to the DCMM. The largest improvements in MAPE occurred sporadically when forecasting $3, 4, 9$, and 14-days ahead. The DCMM has lower MAPE than the independent DBCM for 11 out of 14 forecast horizons.

Forecasting and Impact of Excess

It is also of interest to exemplify the dissection of forecasts based on the binary cascade excess distribution, and explore the impact on forecast uncertainties in particular. From the simulation-based DBCM joint forecast distributions we can trivially extract predicted probabilities of no excess on a future day– the probability than none of the transactions on that day sell more than the specified d items. At the store level, this is potentially useful additional summary information in its own right. Further, looking at the sales forecast distributions conditional on no excess baskets on a particular day provides insights into the impact– on both forecast level and uncertainties– of the excess component of the model.

One selected example is summarized in Figure 6.5 using 1-14 day forecasts for each item made at the earlier selected date of Mar 20th 2017. The figure shows the trajectories of joint forecast distributions over the next 14 days now conditional on no excess (i.e., conditional on predicted $n_{d,t+k} = 0$ for $k = 1 : 14$ where t indexes Mar 20th 2017). These figures have the same format as those for the full unconditional forecasts shown in Figure 6.2. Small differences can be seen, with the conditional forecast distributions naturally favoring slightly lower values while being less diffuse; this is also naturally more pronounced for higher levels of sales such as for item A. Figure 6.5 also displays trajectories of the predictive probabilities of no excess over the next 14 days, naturally indicating higher probabilities for the lower levels of sales exhibited by item C.



(i) Item A

(ii) Item B

(iii) Item C

FIGURE 6.5: 1-14 day forecasts made on on Mar 20th 2017. Joint forecast trajectories conditional on no excess baskets (left), with details as in unconditional trajectories in Figure 6.2, and of corresponding probabilities of no excess (right).

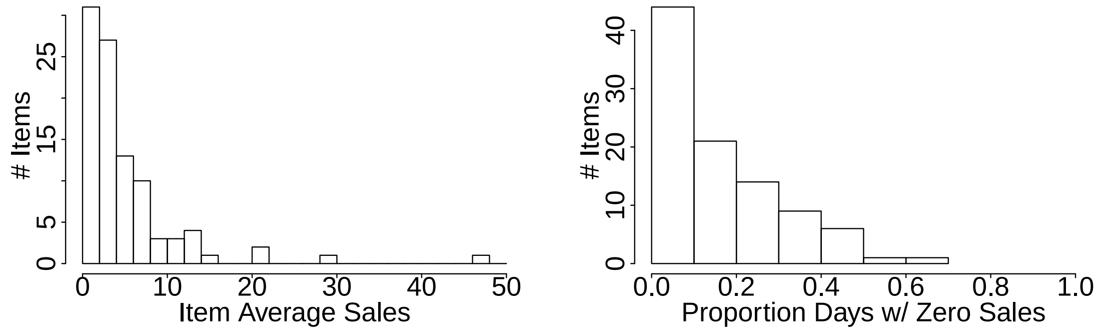


FIGURE 6.6: Histograms of (left) average daily pasta sales and (right) the proportion of days with zero sales across the 96 items in DBCM comparison.

6.4 DBCM Multi-Item Comparison

Section 6.3.3 provides a detailed understanding of the point and probabilistic forecast performance of the DBCM for three supermarket items of varying demand levels. However, we are also interested in the performance of the DCBM approach across a larger number of items and retail outlets. In this section, we apply the multi-scale DBCM approach to 96 items sold across different store locations. We consider products sold at three store locations—denoted A, B, and C—which represent outlets of high, medium, and low sales levels, respectively. Then, we consider 32 UPCs sold at each of the three chosen stores for a total of 96 unique store-UPC combinations. Histograms in Figure 6.6 summarize aspects of the demand of the chosen items. We categorize items into three groups based on the average daily sales in a training data set: low-sellers (0–2), medium-sellers (2–15), and high-sellers (>15). In our selected group of items, there are 31 low-sellers, 61 medium-sellers, and 4 high-sellers.

In this analysis, we implement a multi-scale DCMM to forecast daily transactions of each item. The multi-scale DCMM shares information on the weekly seasonal effect across all pasta products within each store. The item-level models $\mathcal{M}_i$ have $\mathbf{f}_{i,t}' = (1, \log(\text{price}_t), \text{promo}_{i,t})$ where $\text{price}_{i,t}$ is the item-level modal price and $\text{promo}_{i,t}$ indicates whether item i was on promotion on day t. Each $\mathcal{M}_i$ also includes a

7−dimensional $\boldsymbol{\phi}_{s_i,t}$ with one non-zero element representing the current day-of-week effect where s_i indicates the store of item i. That is, each item in store A, with $s_i = A$, inherits the weekly seasonality estimated from an external $\mathcal{M}_0$ fit to the total daily transaction of pasta products in store A. We set fixed discount factors of $\delta = 0.99$ on the Poisson trend and $\delta = 0.995$ for all other binary and Poisson model components. In this multi-scale analysis, $\mathcal{M}_0$ is a DLM on the aggregate log daily transactions of all pasta items in the chosen store $s_i \in \{A, B, C\}$. Each of these aggregate DLMs includes a local linear trend, the scaled log average pasta price as a predictor, and full Fourier form seasonal components of periods 7 and 365 representing the weekly and yearly seasonal effects. We allow for dynamic level, trend, regression effects, and seasonal with discount factors of $\delta = 0.995$ for the trend/regression components, $\delta = 0.999$ for each of the seasonal components, and $\beta = 0.999$ for the residual stochastic variance process. As in previous examples, priors for the model components of $\mathcal{M}_0$ are specified using the reference posteriors from a static Bayesian linear model fit to 21 days of training data. Predictive performance in each DCMM is evaluated and compared using models based on $\rho = \{0.2, 1\}$.

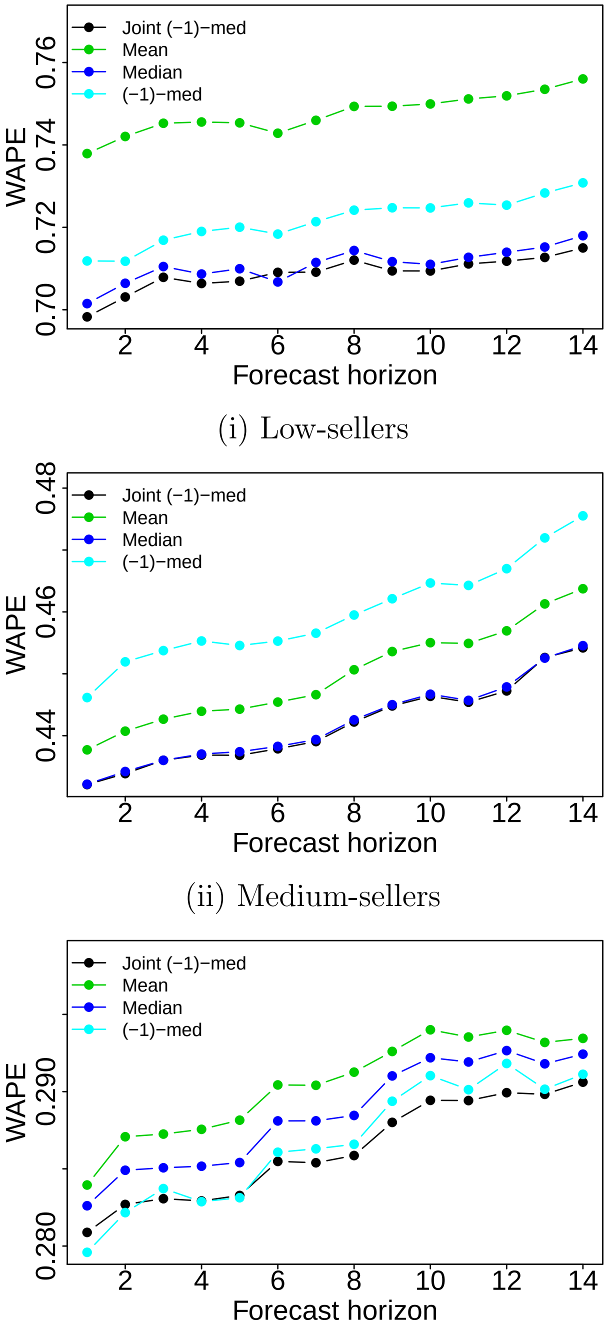
As in our previous example, we set $d = 4$ for all items in this analysis. The form of the binomial logistic DGLM is the same across all items and $r = 1{:}d$. Each conditional model of $n_{r,t}$ includes a dynamic local level and a static regression component with $\text{promo}_{i,t}$ as a predictor. To allow for very slow variation in the level over time, we specify a fixed discount factor of $\delta = 0.999$ on the trend component. We specify priors using the same approach detailed in Section 6.3.3. For this approach, we present results using the empirical approach to the DBCM excess distribution.

For each item, we train our models for one year before forecasting 1:14−days ahead on each of the next 332 days. An initial motivation for this comparison was comparing the large-scale forecasting performance of the multi-scale DBCM to a current industry standard forecasting model. Due to confidentiality issues, we provide only summaries of the results for the industry standard. On each day, the industry standard model produces a predictive mean over the next 1:14−days.

Figures 6.7(i)-(iii) show the weighted absolute percentage error (WAPE) versus forecast horizon for each category of items using several DBCM point forecasts. In Section 2.6.3, we showed that the optimal point forecast under WAPE is an extension of the (−1)-median to a joint distribution. For each category of items, we compute the joint (−1)-median at each time by importance sampling the simulated joint forecasts across each item. The resulting WAPE under the optimal joint (−1)-median is shown in black, along with the WAPE when using the non-optimal mean, median, and (−1)-median point forecasts. For each product category, we see that, with a few exceptions, the WAPE is minimized by the joint (−1)-median. Theoretically, the joint (−1)-median is the optimal point forecast under WAPE, but realized values of WAPE losses are dependent on the outcomes and may, in some cases not be lowest across the set of outcomes being compared. In these examples, the theoretical optimal forecasts generally lead to the lowest realized losses.

For low-sellers, the WAPE under the median is very similar to the optimal WAPE, and there is a slight increase in WAPE for the (−1)-median. The largest values of WAPE occur when forecasting with the predictive mean, and, averaged across the forecast horizons, WAPE decreases by 5.2% under the joint (−1)-median versus the mean. For medium-sellers, the WAPE is again minimized by the joint (−1)-median although WAPE under the median is nearly indistinguishable. In this category, we see that the WAPE under the mean is lower than the WAPE under the (−1)-median. This occurs because the (−1)-median is the furthest point forecast from the joint (−1)-median in terms of absolute distance. Compared to the (−1)-median, the WAPE under the joint (−1)-median decreases by 3.9% on average across the forecast horizon. Compared to the mean, the WAPE under the joint (−1)-median decreases by 1.8% averaged across the forecast horizon. Finally, for high-sellers, we again see that the WAPE is minimized by the joint (-1) median followed closely by the (-1)-median, the median, and then the mean. The WAPE is decreased by 1.5% averaged across the forecast horizon. These results show, in terms of minimizing WAPE, the use of optimal point forecasts is clearly advantageous, especially for low-selling items. Additionally, we see that the WAPE under the median point forecast is very similar to the optimal WAPE. This suggests that the median, which does not require importance sampling to compute, is a reasonable alternative to the optimal point forecast of WAPE. A key takeaway from these results is that the mean is not an appropriate point forecast if our goal is to minimize WAPE. Many models, including many industry standard models, that simply output a predictive mean will be at a disadvantage if minimizing WAPE is a business objective.

Other error metrics we considered for this comparison are RMSE, MAD, and ZAPE. In this analysis, we define $\text{ZAPE}(y, f) = f\mathbb{1}(y = 0) + |y - f|/y\mathbb{1}(y > 0)$. For each of these error metrics, we compare each item's average error over time under the industry standard and the DBCM optimal point forecast. Since the optimal point forecast for RMSE is the mean, both the DBCM and the industry standard produce optimal point forecasts. In our comparison, we found that the DBCM reduces the RMSE for about 50% of the items across the forecast horizon. This result suggests that, for RMSE, neither model is evidently dominant across all items, and that both the industry standard and the DBCM perform similarly on the large-scale with regards to the mean point forecast. For the MAD, minimized by the median, we found that the DBCM reduced the MAD for about 75% of items across the forecast horizon. This reduction in MAD is likely due, in part, to the fact that the DBCM produces the optimal median forecast. Finally, we found that the DBCM reduces ZAPE for close to 100% of the selected items. The ability to compute the ZAPE-optimal point forecast through the DBCM forecast distribution is clearly advantageous in terms of minimizing ZAPE compared to using a predictive mean.



(i) Low-sellers

(ii) Medium-sellers

(iii) High-sellers

FIGURE 6.7: Weighted absolute percentage error (WAPE) vs forecast horizon (days) for (i) low-sellers, (ii) medium-sellers, and (iii) high-sellers from the multi-scale DBCM joint (-1)-median (black), (-1)-median (light blue), mean (green), and median (dark blue).

6.5 Summary Comments

Motivated by an application to product demand forecasting, and enabled by the availability of rich point-of-sale data, we have introduced a novel framework for Bayesian state space modeling of heterogeneous transactions-sales time series. This work stems from the recognition that variability seen in high frequency sales arises from the compounding effect of variability in the number of transactions as well as the number of sales-per-transaction. The dynamic binary cascade model builds upon earlier chapters by adapting the DCMM to model transactions rather than sales. Given the reduced variability of transactions relative to sales, this is a promising application in which the DCMM may improve forecasting accuracy.

Application of the DCMM to transactions of related items offers an opportunity to integrate information across series through a multi-scale, multivariate dynamic factor model. Coupled with the DCMM on transactions, the binary cascade concept involves a sequence of Bayesian models to predict the number of units sold per transaction. The motivation behind this binary cascade is that the appropriate way to forecast rare events is through a sequence of conditional probabilities which define chances of outcomes of increasingly higher– and rarer– sales per transaction. The final stage of the DBCM framework is the choice of excess distribution – leaving it unspecified or choosing a specific form. Leaving the excess distribution unspecified avoids the difficult task of fitting the long tail of the sales-per-transaction distribution, however, this approach limits the conclusions we can present about the forecast distribution. We also present a logical nonparametric choice for the excess distribution which involves bootstrapping from the empirical excess distribution.

In addition to the incorporation of covariates into the binary and Poisson DGLM components of the DCMM, the DBCM framework extends the hierarchical decomposition further by incorporating covariates into the cascade of binomial logistic

DGLMs. This allows incorporation of complex price/promotion effects which may impact the overall traffic in the store, the probability that a customer makes a purchase, and the number of units purchased given that a transaction occurs. The Bayesian framework used for the DBCM allows direct/forward simulation of multi-step ahead predictions, enabling trivial computation of forecast summaries of interest. Selected examples of sales forecasting show the promise for forecast improvement of the DBCM across demand sizes, error metrics, and forecast horizon, emphasizing assessment of probabilistic forecasting accuracy in multiple metrics as well as via standard point forecast summaries.

7

Conclusion

The research presented in this dissertation has focused on modeling and forecasting time series of counts, with a particular focus on high-dimensional settings. Chapter 2 gave details of sequential learning and forecasting in DGLMs and a discussion of the benefits of Bayesian state-space modeling in this applied setting. One section of this chapter describes three variational Bayes methods for conjugate prior specification in DGLMs. The chapter concluded with a discussion of appropriate evaluation of both point and probabilistic forecasts for count data. Chapter 3 presented the DCMM which involves dynamic generalized linear models for binary and conditionally Poisson time series, with dynamic random effects for over-dispersion, allowing use of dynamic covariates in both binary and non-zero count components. In Chapter 4, we extend this framework to an efficient multivariate model that allows borrowing of information across related series. A novel decouple/recouple strategy incorporates cross series linkages while assuring parallelization is possible, resulting in a scalable multi-scale framework as the number of series increases. We present a case study in multi-step forecasting of sales of a number of related items that showcases forecasting of multiple series, with discussion of forecast accuracy metrics and broader questions of probabilistic forecast accuracy assessment. The chapter concluded with an exploration of inference in the multi-scale framework and of the effect of different choices of $\mathcal{M}_0$. Chapter 5 extended the previous examples to compare the proposed DCMM with alternative count forecasting models. Examples demonstrate improved forecasts under the DCMM for a range of point forecast metrics, and illustrate the general improvement in probabilistic forecasting across various items. Chapter 6 introduced a framework which extends the DCMM to apply to forecasting individual customer transactions, coupled with a novel probabilistic model for predicting counts of items per transaction. A central modeling innovation is the new DBCM that addresses interests dissecting heterogeneity in sales outcomes by decomposing sales into transaction counts and units per transaction. A key idea underlying this strategy is modeling and forecasting rare events via a sequence of conditional probabilities, each of which are estimable but their product can be very small. A second case study of multi-step forecasting across several supermarket items shows the improvement of the DBCM compared to the DCMM. I now present some areas of possible future work.

Data sparsity may cause issues in the binary DGLM component of the DCMM for items with very high/low daily sales. Items with very high sales will have sparse binary series z_t which are almost always equal to 1 (equivalently, for low selling items, z_t is almost always zero). When there is sparsity in the data, we do not want to discount the little information we may learn about the success probability over time. Instead of our current approach of fixing a constant discount factor, we may want to consider either a static binary DGLM or a time-varying discount factor. In the context of modeling sparse network flows, Chen et al. (2019) presented an approach for a dynamic discount factor which converges to one when there is limited information in the data. A promising strategy may be to modify their proposed approach to our context of a binary DGLM with conjugate Beta priors.

One specific applied component of the models open to further development is the integration of additional, feed-forward information about promotions at the item level. This of particular interest in connection with forecasting infrequent higher basket sizes based on, for example, "buy 1, get 1 free" types of promotion. Such information can be incorporated in modified models of the excess distribution in a number of ways that should yield practical forecast improvements in such cases.

Another applied topic of interest is the potential forecasting improvement from incorporating additional levels of hierarchy or multiple latent factors. We have focused on latent factors which represent aggregate weekly seasonality, but the proposed framework is generalizable to any other shared factor including, but not limited to, promotions, pricing, weather, brand name, or aggregate trends. Additionally, the multi-scale framework offers additional promise of a method of incorporating effects that otherwise would be too difficult to estimate with limited historical data. For example, yearly seasonality is evident for many supermarket items when modeling daily sales, however, it is difficult to estimate yearly seasonality without many years of historical data. With this multi-scale approach, item-level models could inherit the yearly seasonality learned from aggregate store-level models of similar products. Similarly, it can be difficult to learn the effects of promotions on sales given relative rarity of promotions occurring for individual items. However, if we grouped together similar items and estimated an overall promotion effect, the multi-scale framework offers a way to share information and more precisely estimate shared promotional effects. Future studies will explore the effects of incorporating multiple latent factors on forecasting performance.

In addition to contributing advances in dynamic model-based forecasting for consumer sales, the proposed classes of models (DCMM, DBCMs) should be of interest in other areas involving multiple heterogeneous time series of non-negative integers. Some possible areas of application of the DCMM to count time series include crime forecasting, infectious disease epidemiology, and forecasting demand for emergency services. In each of these applications, forecasts may be required across many individual series, and we may expect the series to share some underlying features based on location or seasonality. The DBCM can be applied to other areas where counts arise from underlying compound processes. One example includes forecasting visitors to different tourist sites by first forecasting number of vehicles, and then number of passengers per vehicle. Similarly, the DBCM could be used to forecast counts of animal species by forecasting the number of herds of animals, and then number of animals per herd.

The proposed DBCM framework offers a general approach to modeling and forecasting compound count processes. Beyond the proposed cascade of binomial models, we could use a shifted Poisson/Negative Binomial or a multinomial distribution for units-per-transaction if we are not concerned with rare events. Another methodological advance would be considering a continuous distribution for the value-per-transaction. For example, in supermarket sales forecasting, goods such as produce and meat are often sold by weight rather than by unit. An extension of our proposed work could be to first forecast the number of transactions which include any positive weight, and then forecast the total weight per day conditional on the number of transactions. For example, consider forecasting the number of transactions b_t through a DCMM, and recall that if $b_t = 0$, then the total weight sold on day t is also zero. There are many possible models for the weight sold per transaction including a log Normal DLM. Additionally, for each individual transaction, we could consider independently modeling the weight-per-transaction with an exponential DGLM with rate λ_t. Conditional on b_t, this implies the total weight can be modeled as $y_t \mid b_t \sim Ga(b_t, \lambda_t)$. This extension would allow modeling of non-negative real-valued outcomes with potentially many days of exactly zero sales.

Bibliography

Aguilar, O. and West, M. (2000), "Bayesian dynamic factor models and portfolio allocation," *Journal of Business and Economic Statistics*, 18, 338–357.

Aguilar, O., Prado, R., Huerta, G., and West, M. (1999), "Bayesian inference on latent structure in time series (with discussion)," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 3–26, Oxford University Press, Oxford.

Aktekin, T., Polson, N. G., and Soyer, R. (2018), "Sequential Bayesian analysis of multivariate count data," *Bayesian Analysis*, 13, 385–409.

Al-Osh, M. A. and Alzaid, A. A. (1987), "First-order integer valued autoregressive (INAR(1)) process," *Journal of Time Series Analysis*, 8, 261–275.

Ali, Ö. G., Sayın, S., van Woensel, T., and Fransoo, J. (2009), "SKU demand forecasting in the presence of promotions," *Expert Systems with Applications*, 36, 12340–12348.

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002), "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, 50, 174–188.

Berry, L. R. and West, M. (2019), "Bayesian forecasting of many count-valued time series," *Journal of Business and Economic Statistics (to appear)*, arXiv:1805.05232.

Berry, L. R., Helman, P., and West, M. (2019), "Probabilistic forecasting of heterogeneous consumer transaction-sales time series," *Submitted for publication*, arXiv:1808.04698.

Boone, T., Ganeshan, R., Jain, A., and Sanders, N. R. (2019), "Forecasting sales in the supply chain: Consumer analytics in the big data era," *International Journal of Forecasting*, 35, 170–180.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (2008), *Time Series Analysis: Forecasting and Control*, John Wiley & Sons, 4th edn.

Cargnoni, C., Müller, P., and West, M. (1997), "Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models," *Journal of the American Statistical Association*, 92, 640–647.

Carvalho, C. M. and West, M. (2007), "Dynamic matrix-variate graphical models," *Bayesian Analysis*, 2, 69–98.

Chen, C. W. S. and Lee, S. (2017), "Bayesian causality test for integer-valued time series models with applications to climate and crime data," *Journal of the Royal of Statistical Society (Series C: Applied Statistics)*, 66, 797–814.

Chen, C. W. S., So, M. K. P., Li, J., and Sriboonchitta, S. (2016), "Autoregressive conditional negative binomial model applied to over-dispersed time series of counts," *Statistical Methodology*, 31, 73–90.

Chen, H. and Boylan, J. E. (2007), "Use of individual and group seasonal indices in subaggregate demand forecasting," *Journal of the Operational Research Society*, 58, 1660–1671.

Chen, X., Irie, K., Banks, D., Haslinger, R., Thomas, J., and West, M. (2018), "Scalable Bayesian modeling, monitoring and analysis of dynamic network flow data," *Journal of the American Statistical Association*, 113, 519–533.

Chen, X., Banks, D., and West, M. (2019), "Bayesian dynamic modeling and monitoring of network flows," *Network Science*, to appear, arXiv:1805.04667.

Croston, J. D. (1972), "Forecasting and stock control for intermittent demands," *Operational Research Quarterly (1970-1977)*, 23, 289–303.

Czado, C., Gneiting, T., and Held, L. (2009), "Predictive model assessment for count data," *Biometrics*, 65, 1254–1261.

Da-Silva, C. Q. and Migon, H. S. (2016), "Hierarchical dynamic beta model," *Revstat*, 14, 49–73.

Dunsmuir, W. T. M. (2015), "Generalized linear autoregressive moving average models," in *Handbook of Discrete-Valued Time Series*, eds. R. Davis, S. Holan, R. Lund, and N. Ravishanker, chap. 3, pp. 51–57, CRC Monographs.

Dunsmuir, W. T. M. and Scott, D. J. (2015), "The glarma package for observation-driven time series regression of counts," *Journal of Statistical Software, Articles*, 67, 1–36.

Fahrmeir, L. (1992), "Posterior mode estimation by extended Kalman filtering for multivariate dynamic linear models," *Journal of the American Statistical Association*, 87, 501–509.

Ferland, R., Latour, A., and Oraichi, D. (2006), "Integer-valued GARCH process," *Journal of Time Series Analysis*, 27, 923–942.

Ferreira, M. A. R., Bi, Z., West, M., Lee, H. K. H., and Higdon, D. M. (2003), "Multiscale modelling of 1-D permeability fields," in *Bayesian Statistics 7*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. David, D. Heckerman, A. F. M. Smith, and M. West, Oxford University Press.

Ferreira, M. A. R., West, M., Lee, H., and Higdon, D. M. (2006), "Multiscale and hidden resolution time series models," *Bayesian Analysis*, 2, 294–314.

Fildes, R. and Goodwin, P. (2007), "Against your better judgment? How organizations can improve their use of management judgment in forecasting," *Interfaces*, 37, 570–576.

Fokianos, K., Rahbek, A., and Tjøstheim, D. (2009), "Poisson autoregression," *Journal of the American Statistical Association*, 104, 1430–1439.

Gneiting, T. (2011), "Making and evaluating point forecasts," *Journal of the American Statistical Association*, 106, 746–762.

Gruber, L. F. and West, M. (2016), "GPU-accelerated Bayesian learning in simultaneous graphical dynamic linear models," *Bayesian Analysis*, 11, 125–149.

Gruber, L. F. and West, M. (2017), "Bayesian forecasting and scalable multivariate volatility analysis using simultaneous graphical dynamic linear models," *Econometrics and Statistics*, 3, 3–22.

Hartmann, H. C., Pagano, T. C., Sorooshian, S., and Bales, R. (2002), "Confidence builders: Evaluating seasonal climate forecasts from user perspectives," *Bulletin of the American Meteorological Society*, 83, 683–698.

Heinen, A. (2003), "Modelling time series count data: An autoregressive conditional Poisson model," *Munich Personal RePEc Archive*, 8113, Electronic publication.

Hyndman, R., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008), *Forecasting with Exponential Smoothing: the State Space Approach*, Springer, Berlin and Heidelberg.

Hyndman, R. J. and Koehler, A. B. (2006), "Another look at measures of forecast accuracy," *International Journal of Forecasting*, 22, 679–688.

Hyndman, R. J., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2005), "Prediction intervals for exponential smoothing using two new classes of state space models," *Journal of Forecasting*, 24, 17–37.

Kolassa, S. (2016), "Evaluating predictive count data distributions in retail sales forecasting," *International Journal of Forecasting*, 32, 788–803.

Kourentzes, N., Petropoulos, F., and Trapero, J. R. (2014), "Improving forecasting by estimating time series structural components across multiple frequencies," *International Journal of Forecasting*, 30, 291–302.

Lopes, H. F., Johannes, M. S., Carvalho, C. M., and Polson, N. G. (2011), "Particle learning for sequential Bayesian computation (with discussion)," in *Bayesian Statistics 9*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, pp. 317–360, Oxford University Press.

Ma, S., Fildes, R., and Huang, T. (2016), "Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information," *European Journal of Operational Research*, 249, 245–257.

McCabe, B. P. M. and Martin, G. M. (2005), "Bayesian predictions of low count time series," *International Journal of Forecasting*, 21, 315–330.

McKenzie, E. (1988), "Some ARMA models for dependent sequences of Poisson coutns," *Advances in Applied Probability*, 20, 822–835.

Morlidge, S. (2015), "Measuring the quality of intermittent-demand forecasts: It's worse than we've thought!" *Foresight: The International Journal of Applied Forecasting*, 37, 37–42.

Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F., and Assimakopoulos, V. (2011), "An aggregate–disaggregate intermittent demand approach (ADIDA) to forecasting: An empirical proposition and analysis," *Journal of the Operational Research Society*, 62, 544–554.

Prado, R. and West, M. (2010), *Time Series: Modeling, Computation & Inference*, Chapman & Hall/CRC Press.

Quintana, J. M. and West, M. (1988), "Time series analysis of compositional data," in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, pp. 747–756, Oxford University Press.

Rosenblatt, M. (1952), "Remarks on a multivariate transformation," *The Annals of Mathematical Statistics*, 23, 470–472.

Seaman, B. (2018), "Considerations of a retail forecasting practitioner," *International Journal of Forecasting*, 34, 822–829.

Snyder, R. D., Ord, J. K., and Beaumont, A. (2012), "Forecasting the intermittent demand for slow-moving inventories: A modelling approach," *International Journal of Forecasting*, 28, 485–496.

Taylor, J. W. (2007), "Forecasting daily supermarket sales using exponentially weighted quantile regression," *European Journal of Operational Research*, 178, 154–167.

Trapero, J. R., Cardós, M., and Kourentzes, N. (2019), "Quantile forecast optimal combination to enhance safety stock estimation," *International Journal of Forecasting*, 35, 239–250.

Triantafyllopoulos, K. (2009), "Inference of dynamic generalized linear models: Online computation and appraisal," *International Statistical Review*, 77, 430–450.

Weisheimer, A. and Palmer, T. N. (2014), "On the reliability of seasonal climate forecasts," *Journal of the Royal Society Interface*, 11, 1–10.

West, M. (1981), "Robust sequential approximate Bayesian estimation," *Journal of the Royal Statistical Society (Series B: Methodological)*, 43, 157–166.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer-Verlag, New York, Inc, 2nd edn.

West, M., Harrison, P. J., and Migon, H. S. (1985), "Dynamic generalised linear models and Bayesian forecasting (with discussion)," *Journal of the American Statistical Association*, 80, 73–97.

Withycombe, R. (1989), "Forecasting with combined seasonal indices," *International Journal of Forecasting*, 5, 547–552.

Yelland, P. M. (2009), "Bayesian forecasting for low-count time series using state-space models: An empirical evaluation for inventory management," *International Journal of Production Economics*, 118, 95–103.

Biography

Lindsay Berry began her undergraduate studies at the University of Texas at Austin in 2011, and completed her Bachelor of Science in Mathematics Honors in May 2015. As a member of the Dean's Scholars, Lindsay completed an honors thesis on "Simulation Control of Seamless Phase II/III Clinical Trials" under the supervision of Peter Mueller. In August 2015, Lindsay began her graduate studies in the Department of Statistical Science at Duke University. Her graduate research with Mike West on modeling and forecasting multivariate time series of non-negative counts appears in Berry and West (2019) and Berry et al. (2019). She graduated with her Ph.D in May 2019, and became a Statistical Scientist at Berry Consultants in July 2019.