Aalto University
School of Science
Master's Programme in Mathematics and Operations Research

Tommi Vainio

On Hierarchical Campaign Forecasting

Master's Thesis
Helsinki, February 17, 2021

Supervisor: Assistant Professor Pauliina Ilmonen
Advisor: Erkka Saarinen M.Soc.Sci. (Econ.)

Aalto University
School of Science
Master's Programme in Mathematics and Operations Research

ABSTRACT OF MASTER'S THESIS

Author:	Tommi Vainio		
Title:	On Hierarchical Campaign Forecasting		
Date:	February 17, 2021	**Pages:**	62
Major:	Systems and Operations Research	**Code:**	SCI3055
Supervisor:	Assistant Professor Pauliina Ilmonen		
Advisor:	Erkka Saarinen M.Soc.Sci. (Econ.)		

As retail world is growing more and more competitive, retailers are forced to optimize their operations as well as possible. Accurate demand forecasting is pivotal in optimizing and automizing the replenishment of a retail chain. However, many different factors, such as promotional activities, induce drastic variation in the demand, making forecasting difficult if no appropriate methods or sufficient amount of data is available for the retailer.

This thesis specifies different designs of hierarchical, multilevel regression models for demand and campaign effect forecasting utilizing a natural product hierarchy within a retail chain. The performance of the models is evaluated with a dataset obtained from a large European retailer. The model evaluation is threefold, consisting of assessing the model estimation times, forecast accuracy, and the number of excess errors produced by each model. A particular focus is on cases where there is little data of promotional activities on the lowest hierarchy level, which is a case where conventional models struggle the most. Both frequentist and Bayesian approach is experimented. For the models specified in Bayesian fashion, both Maximum a Posteriori estimation of the parameters and full posterior distribution estimation with Variational Inference methods are implemented, and the results are compared.

The thesis aims to find out whether the specified hierarchical models outperform regular, single-level Generalized Linear Models specified similarly and to understand the advantages and disadvantages of the differently specified hierarchical models. Also, differences in the results of different estimation algorithms and regression types applicable to the problem are of interest.

It is concluded that the parameter estimation of most of the hierarchical models evaluated here is computationally too heavy to be used in operational demand forecasting of large retail chains, while producing little advantages compared to the single-level models. However, Empirical Bayes models seem very promising, as they outperform both the benchmark and the other hierarchical models in almost every evaluated attribute. In addition, a significant finding is that no clear advantages are obtained by using non-normal regression models, in comparison to linear regression where the objective variable is assumed to be normally distributed.

Keywords:	demand forecasting, promotion forecasting, generalized linear mixed models, hierarchical modelling, Bayesian data analysis, Empirical Bayes, Maximum a Posteriori, Variational Inference
Language:	English

Aalto-yliopisto
Perustieteiden korkeakoulu
Matematiikan ja operaatiotutkimuksen maisteriohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Tommi Vainio		
Työn nimi:	Hierarkkisesta kampanjaennustamisesta		
Päiväys:	17. helmikuuta 2021	**Sivumäärä:**	62
Pääaine:	Systeemi- ja operaatiotutkimus	**Koodi:**	SCI3055
Valvoja:	Apulaisprofessori Pauliina Ilmonen		
Ohjaaja:	VTM Erkka Saarinen		

Jatkuvasti kasvava kilpailu vähittäiskaupan alalla pakottaa kauppiaat optimoimaan toimintaansa niin hyvin kuin mahdollista. Tarkka kysyntäennuste on eräs tärkeimmistä tekijöistä toimitusketjun optimoinnissa ja täydennystilaamisen automatisoinnissa. Erilaiset kampanjat, sekä muut tekijät, jotka aiheuttavat suurta vaihtelua kysyntään vaikeuttavat kuitenkin kysynnän ennustamista merkittävästi. Erityisiä haasteita luovat tilanteet, joissa näistä kampanjoista ei ole riittävästi historiallista tietoa, tai mikäli vähittäiskauppiaan käytössä ei ole sopivia työkaluja ennusteen muodostamiseksi.

Tässä diplomityössä selvitetään erilaisten hierarkkisten mallien kykyä hyödyntää vähittäiskauppaketjujen tuotehierarkiaa kysynnän ennustamisessa, erityisesti keskittyen kampanjoiden vaikutuksen ennustamiseen tilanteissa, joissa dataa historiallisista kampanjoista on matalimmalla hierarkiatasolla vain vähän. Mallien suorituskykyä mitataan niiden estimointiin kuluvan ajan, ennustetarkkuuden, sekä poikkeuksellisen suurten ennustevirheiden lukumäärän suhteen eurooppalaiselta vähittäiskauppaketjulta saadun aineiston avulla. Arvioitujen mallien joukkoon kuuluu niin Bayesläisen, kuin myös tilastollisen todennäköisyystulkinnan näkökulmasta määriteltyjä malleja. Bayesläisittäin määriteltyjen mallien parametriestimointi suoritetaan mahdollisuuksien mukaan kahdella eri tavalla, MAP-estimoinnilla sekä Variaatiopäättely-menetelmällä, ja näiden menetelmien tuloksia vertaillaan.

Työn tavoitteena on selvittää, että soveltuvatko hierarkkiset mallit perinteisiä yleistettyjä lineaarisia malleja paremmin kampanjaefektien ennustamiseen. Myös eri regressiotyyppien välisiä eroja tutkitaan ja niiden soveltuvuudesta tämänkaltaiseen tilanteeseen käydään keskustelua.

Tutkimuksen tuloksena huomataan, että lähes kaikkien kokeiltujen hierarkkisten mallien parametriestimointi on liian hidasta, jotta niitä voitaisiin hyödyntää suurten vähittäiskauppaketjujen päivittäisessä kysyntäennustamisessa. Empiirinen Bayes -menetelmällä määritellyt ja sovitetut mallit nousevat kuitenkin esiin erittäin lupaavana vaihtoehtona hierarkisten mallien joukosta, suoriutuen muita malleja paremmin lähes kaikilla mittareilla mitattuna. Mielenkiintoisena voidaan myös pitää havaintoa, että normaalijakauma osoittautui kilpailukykyiseksi vaihtoehdoksi yleistettyjen lineaaristen mallien kohdemuuttujan jakaumalle, muiden ongelman luonteeseen sopivien eksponenttiperheen jakaumien rinnalla.

Asiasanat:	kysynnän ennustaminen, kampanjaennustaminen, yleistetty lineaarinen sekamalli, hierarkkinen mallinnus, Bayesiläinen data-analyysi, Empiirinen Bayes, MAP, Variaatiopäättely
Kieli:	Englanti

Acknowledgements

This thesis has been one of the most complex and rewarding projects I have ever done. I wish to thank my employer for providing this interesting topic for me and offering me the possibility to grow professionally during my studies together with the best possible colleagues.

I also want to express my gratitude to my supervisor Pauliina Ilmonen for the continuous support and invaluable guidance during this project. Your positive attitude and spot-on advice always helped me to maintain my motivation and stay on schedule even when things seemed confusing and even a little hopeless. Likewise, I want to thank my advisor Erkka for an outstanding level of commitment to this project and for extremely valuable comments and remarks throughout the whole process. The frequent meetings with you always sparked new ideas and helped me to understand this interesting topic better.

The past few years I have spent at Aalto University have been amongst the very best in my life. I have been fortunate to be surrounded by so many amazing people from whom I have learned so much more than I ever thought possible. I have made incredible friends who I wish to keep in my life indefinitely. Thank you for the most exciting adventures, selfless help and support, and for making the many long days at the university something that I will most definitely miss in the future.

Finally, and most importantly, I want to thank my family for always being there for me and believing in me in everything I have decided to do. I would not be who I am without you.

Helsinki, February 17, 2021

Tommi Vainio

Contents

Chapter 1

Introduction

In the highly competitive world of retail, accurate demand forecasting and automated just-in-time replenishment have grown more and more critical as many companies are optimizing their supply chains as well as possible. Simultaneously, promotional activities, such as campaigns, are often one of the main drivers of demand for many products, which poses challenges for accurate demand forecasting. This is especially true for one of the most commonly used demand forecasting methods: time-series forecasting. Those methods generally only consider the previous observations without acknowledging any external variables, such as promotional activities. The traditional way of dealing with external variables in demand forecasting has been to clean the sales data from the effect of such variables and use a separate process and forecasting method to forecast the effects of these variables. This separately forecasted effect can then be added to, or subtracted from, the baseline forecast. Such a separate forecasting method could be as simple as calculating an average of the external variables' past effects. In promotion forecasting, for example, the time-series methods can be used to obtain a baseline forecast, on top of which separately calculated campaign uplifts could then be added for the days in the future where similar promotional activities are planned to take place. Another option is to use regression modeling, where the variance in the sales quantities is explained by different independent variables (i.e., regressors). The promotional activities can then be included in the model as regressors, similarly to any other factors affecting the sales.

Both of the above-mentioned methods provide good and reasonable forecasts in situations when there exists a suitable grouping for different promotion categories and types, such that the promotions within the same type or category have a similar effect on sales, and when there exists enough data in the history for each of the promotion types to estimate this effect reliably. However, often the case is that such categorization is not readily available

or that the typing is so specific that there are not enough data points in history to estimate the effect of each promotion type reliably. Another common problem arises in situations where a new type of promotion is introduced for a product, as the effect of that type can not be directly estimated due to an absence of data of that promotion type for that specific product. For solving these problems, it is common to utilize some of the hierarchical structures that exist within retail store chains. For example, the products are usually divided into a few different categories, which are further divided into different types, et cetera. A final level in such a hierarchy would be product-store combinations, where one instance represents one product sold in a specific store. This level is the level where the sales are most often recorded, and thus also the level on which the forecast is usually calculated. An example of such a hierarchy is illustrated in Figure 1.1. Another type of hierarchy often present in retail chains is related to stores.

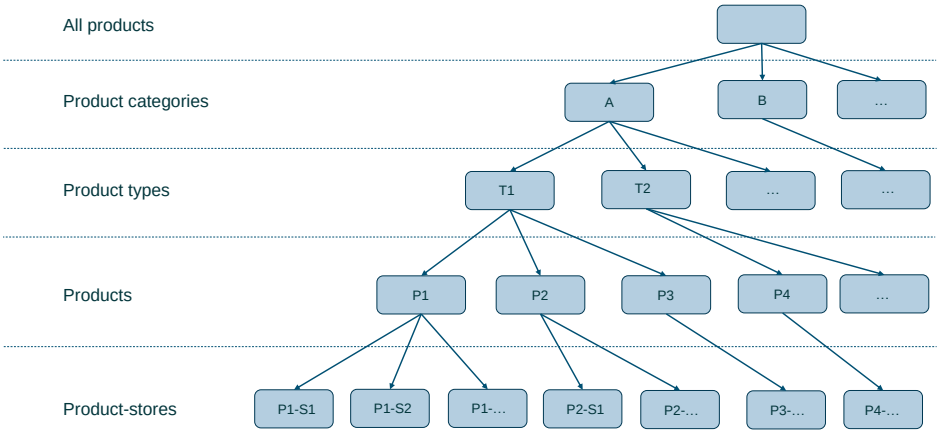


Figure 1.1: An illustration of a product group hierarchy within retail chains.

A common approach in estimating a reliable campaign forecast for product-stores that lack campaign data is to utilize the product hierarchy and use data from the higher hierarchy levels based on some heuristic to estimate the uplift. The problem with this is that setting up these heuristics requires much manual work, and even then, it might not produce reasonable forecasts. Furthermore, the amount of campaign data on the product-store level is often rather limited. Thus, obtaining accurate campaign forecasts is highly dependent on the quality of the heuristics based on which the data on higher hierarchy levels are utilized. In this thesis, hierarchical models that complement the data on the lowest level with the information from the higher levels

by design are specified, tested, and evaluated in the context of a retail chain, with data obtained from a large European food retailer. As there are multiple ways to specify and estimate the parameters of such models, different methods are experimented, and the advantages and disadvantages of them are analyzed and discussed.

This thesis is structured as follows. Chapter 2 reviews earlier literature on hierarchical modelling and presents the mathematical foundations for such methods. In Chapter 3, the models and algorithms used in this thesis are presented. Chapter 4 introduces the data used for fitting and evaluating the models, and Chapter 5 presents and discusses the results obtained with them. Finally, the conclusions drawn from the results, together with ideas for future research, are discussed in Chapter 6.

Chapter 2

Background

2.1 Generalized Linear Models

The term Generalized Linear Model (GLM) refers to a class of parametric regression models, where the response variable is assumed to belong to an exponential family, which includes normal, Poisson, binomial and gamma distributions. GLMs can thus be considered as a generalization of the classical linear model, where the response variable is assumed to be normally distributed. This generalization is also useful in demand forecasting, especially when the demand for a product is relatively low, as the sales numbers are non-negative and can often obtain only integer values. The concept of GLMs was first proposed by Nelder and Wedderburn [1972], where they also presented a unified estimation procedure for all of the models within the GLM family. This popularized the use of non-normal regression models, and they have been widely utilized in different fields and applications since.

The basic form of a GLM can be written for example as

$$\begin{aligned} y &\sim F(h(\lambda), \omega) \\ \lambda &= g(E\left[y|X\right]) = X\beta \end{aligned} \tag{2.1}$$

where F is a distribution from the exponential family, h the inverse function of a smooth and monotonic link function g, λ is the linear predictor for the response variable y, representing the expectation of conditional to the predictors X, β are estimated coefficients for the predictors, and ω includes the additional parameters required to parametrize the distribution F. The inverse $h = g^{-1}$ of the link function is used to map the values of the linear predictor λ to correspond to the appropriate distribution.

In demand forecasting, GLMs can be used to incorporate the effects of external variables, such as different promotional activities. As discussed in

Chapter 1, GLMs can either be used to model the full time-series or together with some time-series based forecasting method to model only the effect of the external variables not incorporated by the time-series based method. Problems with GLMs and most other single-level models in demand forecasting are that the number of regressors can quickly grow very large. This is especially problematic because, as mentioned in Chapter 1, there are often a very limited number of data points in history to estimate the effect of promotions of certain types or certain product-store combinations.

2.2 Hierarchical modelling

Hierarchical modelling, often also referred to as Multilevel modelling, is referring to modelling systems where observations can be grouped hierarchically, such that the information on higher levels can be utilized to complement the estimation of the lower-level parameters, thus decreasing the number of data points required for reliable inference. A commonly used example of a hierarchical data structure in educational sciences is students, which can be grouped into classes, which can be grouped into schools. Schools can, of course, also be grouped into different geographical clusters and so on.

Hierarchical modelling techniques have been widely utilized in social sciences, where the data is often hierarchical. For example, Bryk and Raudenbush [1987] utilized a two-stage hierarchical model to assess the structure of the psychological growth and the factors affecting it on individual and group levels. They noted already then that the performance and flexibility provided by the hierarchical offered excellent predictive performance on an individual level. Later, also Hofmann et al. [2000] took a similar hierarchical approach for growth modelling, but also incorporated longitudinal data of the persons and obtained good results. Gelman et al. [2006], on the other hand, used hierarchical modelling to analyze whether some racial groups were stopped more often than others and found out that persons of African and Hispanic descent were overrepresented in the data.

The problem with using traditional GLMs with data that has a hierarchical structure is that the values of the response variable are often correlated within the groups, leading to violating the i.i.d. requirements of the errors of the model if the group information is not taken into account. The approach for solving this problem depends on how the value of the response variable is assumed to be varying between the groups. If only the intercept value is expected to vary, one can simply add an indicator variable for each group into the GLM to incorporate the different baseline. Then again, if also the coefficients of the regressors are expected to vary between the different groups,

separate models need to be estimated for each group to explain the variation and obtain i.i.d. residuals. However, adding more regressors or estimating the models purely on lower levels leads to a need for an increased number of data points to obtain reliable estimates for the parameters. [Meijer et al., 2008]

Hierarchical modelling techniques offer a possibility to estimate these group-level coefficients while also utilizing the information on the higher levels, assuming that the coefficients of the different groups follow the same distribution. In practice, this can be implemented by specifying a separate probability model for the coefficients of the top-level regression model. The probability model of the coefficients has its own parameters, which are usually referred to as hyperparameters. [Gelman and Hill, 2006] These probability models can also include regressors on the group level if some part of the group-level variation in the parameters can be assumed to be explained with some group level information.

Such hierarchical models are referred to in literature with multiple different terms. Commonly used terms include Hierarchical Linear Models or Multilevel Linear Models, but perhaps the most common one is Generalized Linear Mixed Models, or GLMMs, which are extensively discussed by Raudenbush [1988]. The term "mixed model" refers to having both "fixed" and "random" components in the model. However, as pointed out by Gelman and Hill [2006], the term "fixed effect" is often used ambiguously when referring to both parameters that are constant for all groups and to the parameters estimated separately for each group, but without the hierarchical correlation structure. Thus, as proposed by Gelman and Hill [2006], we refer to the parameters either as "constant" or "varying" intercepts/slopes depending on whether the values of them vary by group or not, with the varying parameters further distinguished to hierarchical and non-hierarchical parameters. That can be handled in the model specification phase by specifying the variance components of the varying parameters appropriately, which will be explained more in detail later in this Chapter. One possible way to specify a Generalized Linear Mixed Model where there are no group-level predictors, as can be assumed to be the case when forecasting the effect of promotions based on only indicators of different promotion types, following the notation used in Gelman and Hill [2006], is presented below:

$$
\begin{aligned}
y_i =& \sum_{n=1}^{N} \alpha^n x_i^n + \sum_{m=1}^{M} \beta_{j(i)}^m z_i^m + \sum_{k=1}^{K} \gamma_{j(i)}^k u_i^k + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_y^2) \\
\beta_j^m =& \mu_{\beta^m} + \eta_j^m, \qquad \eta_j^m \sim N(0, \infty) \\
\gamma_j^k =& \mu_{\gamma^k} + \xi_j^k, \qquad \xi_j^k \sim N(0, \Sigma_\gamma).
\end{aligned} \tag{2.2}
$$

In this formulation, y_i denotes the response variable value of observation i, $j(i)$ specifies the group j to which the observation i belongs to. x_i^n denotes each of the N variables for which the coefficients are modelled to be constant across the different groups, z_i^m are the M variables that are modelled so that their coefficients vary across the groups but without the between-group correlation, whereas u_i^k are the K parameters in the model whose coefficients are assumed to vary between the groups and to be correlated across the groups. $\alpha^n, \beta_{j(i)}^m$ and $\gamma_{j(i)}^k$ are the effective coefficients of the respective variables. ϵ_i is used to denote the individual level variance, and η_j^m and ξ_j^k are used to denote the group-level variance in the uncorrelated and correlated parameters, respectively. μ_{β^m} and μ_{γ^k} represent the means of their corresponding parameters, across which the group-specific coefficients vary. As can be seen from the second line of Equation (2.2), the η_j^m:s are specified to follow a zero-mean normal distribution with infinite variance, essentially making the specification of μ_{β^m} obsolete and leading to group-specific OLS-estimates for the coefficients, whereas the coefficients ξ_j^k follow a multivariate normal distribution with zero means and correlation σ_γ. The notation could be made even more general by specifying a probability model also for the parameters α^n and restricting the varying component to zero by specifying them to follow a normal distribution with zero mean and zero variance.

2.3 Bayesian models

The distinctive feature of Bayesian data analysis methods is that the coefficients are treated as random variables, and thus the goal is to estimate them as probability distributions. This approach is useful as it provides more intuitive interpretation for each parameter estimate's uncertainty and allows incorporating prior knowledge of the values about the coefficients into the model. This is done using Bayes' theorem, which is defined as

$$
P(\Phi|y) = \frac{P(y|\Phi)P(\Phi)}{P(y)}, \tag{2.3}
$$

which can be interpreted as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}. \tag{2.4}$$

Within the Bayesian framework, the subject of interest is the posterior distribution of the parameters, indicating the most probable distribution for the estimated parameters, given the determined likelihood, prior belief, and the observed evidence. This contrasts with the traditional frequentist approach, where the goal is usually to maximize the likelihood of the data with respect to the parameters. Incorporating prior knowledge into the model can have many advantages, especially when the amount of data points is low compared to the number of parameters estimated for the model, as the priors also work as a sort of regularization, keeping the parameters in the proximity of this prior knowledge if no evidence would indicate otherwise. However, one must be careful not to specify prior distributions that would lead to false conclusions about the posterior densities, and in case there is no meaningful knowledge about some parameters, a non-informative uniform prior should be used. [Gelman et al., 2013]

The Bayesian framework offers a natural way to specify hierarchical models, as the probability models of the parameters and hyperparameters can be specified as prior distributions. For example, if the Equation (2.2) were to be modelled in a Bayesian way, the probability models of ϵ, η and ξ would act as their prior distribution, which would have their own hyperparameters. In addition, prior distributions would be specified for these hyperparameters and all other parameters denoted with greek letters in Equation (2.2). A model specified this way is often referred to as Bayesian GLMM, which are discussed more in-depth, for example, in Fong et al. [2010] or in Zhao et al. [2006], where Zhao et al. presents a general design for Bayesian GLMMs. Another possible way to model hierarchical structures in the Bayesian context is to use so-called Empirical Bayes methods, where the prior distributions of the model are specified based on values estimated from higher levels, discussed for example by Morris [1983]. There is no established definition for Empirical Bayes models, but in this context, it could mean e.g. estimating the regression model first solely on the highest level (i.e., estimating a pooled model) and using the marginal posterior densities of the parameters of the higher-level model as the prior distributions on the lower level.

2.4 Model estimation methods

The parameter values in hierarchical mixed models can be estimated in multiple different ways, which can roughly be divided into two categories, point-

estimate methods, and distribution estimation methods. In methods yielding point estimates, a single value for each parameter is searched, even though the estimator itself can have a distribution. A typical procedure for this is to formulate some criterion that is maximized/minimized by the parameter values that possess some desired qualities and then find the maximizing/minimizing values of the parameters with some numerical optimization technique.

For demonstrative purposes, we will consider a general formulation of a hierarchical linear model defined as

$$\begin{aligned} y &= X\beta + ZB + \epsilon, \\ B &\sim N(0, \Sigma_\theta), \\ \epsilon &\sim N(0, \sigma^2 I), \end{aligned} \tag{2.5}$$

where y is the vector of observations of the response variable, vector β denotes the constant, group invariant effects corresponding to group independent variables in matrix X, B is a vector, modelled as random variables, that represents the varying and hierarchical, group-specific effects related to group-specific variables in matrix Z, Σ_θ is a variance-covariance matrix parametrized by variance component vector θ and ϵ is the data level error following a normal distribution with zero mean and variance of σ^2. The probability model of Y with the condition that $B = b$ can thus be written as:

$$\begin{aligned} (Y|B = b) &\sim N(X\beta + Zb, \sigma^2 I), \\ B &\sim N(0, \Sigma_\theta). \end{aligned} \tag{2.6}$$

This is an alternative way for formulating an Equation (2.2), and is used here to simplify the presentation of the estimation methods. The matrix X here contains both of the non-hierarchical variable groups x and z, while the matrix Z contains the hierarchical variables u. The reason for this is that the group-specific, but not hierarchical variables z from the earlier formulation can be specified as similar constant effect variables as the ones in x, by copying each variable in z separately for each group j, and then setting the variable values to zero for those observations that do not belong into the group j for each column.

2.4.1 Point-estimate methods

For a Hierarchical Linear Model such as (2.5), analytical expressions for finding point-estimates for the parameters β and b can be derived. For parameters β, this can be done by formulating the Best Linear Unbiased Estimator (BLUE), discussed more in detail, for example, by Zmyślony [1980]. As noted

by Puntanen and Styan [1989], this estimator corresponds to the Generalized Least Squares (GLS) estimator that defined with matrix notation as

$$\hat{\beta} = \text{BLUE}(\beta) = (X^\top V^{-1} X)^{-1} X^\top V^{-1} Y, \tag{2.7}$$

where $V = \text{Var}(Y) = Z^\top \Sigma_\theta Z + \sigma^2$.

The values of the parameters b can not be estimated as they are treated as random variables. Thus, we need to find predictions for those values, which can be done by finding such values $\hat{b}$ that are linear in the data Y, unbiased, and minimize the prediction error

$$e = \left[s^\top X\hat{\beta} + r^\top Z\hat{b} - (s^\top X\beta + r^\top Zb) \right]^2 \tag{2.8}$$

with arbitrary vectors s and r. Such predictor is called the Best Linear Unbiased Predictor (BLUP), discussed e.g. by Goldberger [1962] and Robinson et al. [1991], and can be formulated in this case as

$$\hat{b} = \text{BLUP}(b) = \Sigma_\theta X^\top V^{-1} (Y - X\hat{\beta}). \tag{2.9}$$

From Equations (2.7) and (2.9) it can be seen that in order to obtain the estimates for the regression parameters β and b, one must also estimate the variance components σ^2 and Σ_θ. These components can be estimated in multiple different ways, some of which are presented below. These methods can also be used in finding the estimators and predictors for parameters β and b when analytical expressions are not available, which is the case with everything other than linear regression models with normal likelihood. [McCulloch, 1997]

2.4.1.1 Maximum Likelihood

One option for estimating the coefficient values of the model's different parameters is Maximum Likelihood (ML), where such values that maximize the likelihood function of the model are searched. These parameter values represent the values with which the likelihood of the observed data is highest. For example, the likelihood of the general model (2.5) can be defined as

$$\mathcal{L}(\theta, \beta, \sigma | y) = \int_b p(y_i | \theta, \beta, \sigma) f(b | \Sigma_\theta) db, \tag{2.10}$$

where $p(y_i|\theta, \beta, \sigma)$ is the probability mass function of y given the parameters, and $f(b|\Sigma_\theta)$ is the probability density at b. It is good to note that here the likelihood of the observed data y is no longer conditional on the parameters b, as it is marginalized over them by integrating over parameter space of b.

The Maximum Likelihood solution for the model is then obtained by maximizing the logarithm of this function with regards to the necessary parameters, with some nonlinear optimization methods such as Newton's method or BFGS. In practice, this is often a very complex problem and guaranteeing the convergence of the algorithm requires, for example, factorization of the variance-covariance matrix Σ_θ, removing some parameters from the objective function by determining analytical expressions for their maximizers, as well as other steps to simplify the objective function.

2.4.1.2 Restricted Maximum Likelihood

A commonly known problem with Maximum Likelihood estimation is that it does not take the degrees of freedom of estimating the mean into account and is thus yields biased estimates for the variance. (e.g. Harville [1977]) This is usually not a very significant problem in non-hierarchical regression models, but with hierarchical models, the effect of this bias can become significant. This problem can be addressed by maximizing a restricted version of the log-likelihood function (REML), developed and discussed more in detail by Patterson and Thompson [1971]. The approach for obtaining the parameter estimates with REML is similar to Maximum Likelihood estimates, as the parameter values maximizing the function must be determined with numerical methods, such as nonlinear optimization. Also, similarly as with Maximum Likelihood, different matrix factorizations and modifications to the expression are often necessary to guarantee convergence of the optimization method due to the complexity of the objective function.

2.4.1.3 Maximum a Posteriori

If the problem is defined in a Bayesian framework, such that all of the coefficients are modelled as random variables, point estimates for them can be obtained using a Maximum a Posteriori (MAP) estimate. [Bassett and Deride, 2019] The concept of MAP estimation is similar to Maximum Likelihood, but instead of finding the maximizing parameter values for the likelihood, the parameter values that maximize the marginal posterior distributions of the parameters are searched. In general, the MAP estimator can be derived

as

$$\begin{aligned}
\text{MAP}_\Phi &= \arg\max_\Phi P(\Phi|Y) \\
&= \arg\max_\Phi \frac{P(Y|\Phi)P(\Phi)}{P(Y)} \\
&= \arg\max_\Phi P(Y|\Phi)P(\Phi) \\
&= \arg\max_\Phi \log(P(Y|\Phi)P(\Phi)) \\
&= \arg\max_\Phi \log(P(Y|\Phi)) + \log(P(\Phi)),
\end{aligned} \tag{2.11}$$

where Φ is the collection of the model parameters.

Using the Bayes' theorem, the part of the posterior distribution of a hierarchical model, formulated as in (2.5), that depends on the parameters can be written as

$$\begin{aligned}
P(\beta, B, \sigma|Y) &\propto P(Y|\beta, B, \sigma)P(B, \Sigma_\theta)P(\beta)P(\sigma) \\
&= P(Y|\beta, B, \sigma)P(B|\Sigma_\theta)P(\Sigma_\theta)P(\beta)P(\sigma)
\end{aligned} \tag{2.12}$$

Assuming normal likelihood, normal priors for β parameters, a uniform prior for the data-level standard deviation and adopting a parametrization $\Sigma_\theta = \text{Diag}(\theta)\Omega\text{Diag}(\theta)$ with a half-Cauchy prior for the standard deviations θ and an inverse Wishart prior for the correlation matrix Ω as proposed by Gelman et al. [2013], the components of the distribution proportional to the posterior can be written as

$$\begin{aligned}
P(Y|\beta, \sigma, B) &= \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(Y_i - X_i\beta - Z_i B_{j(i)})^2}{2\sigma^2}\right) \\
P(B|\Sigma_\theta) &= \prod_{j=1}^{J} (2\pi)^{\frac{-K}{2}} \det(\Sigma_\theta)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} B_j^\top \Sigma_\theta^{-1} B_j)\right) \\
P(\Omega) &= \frac{1}{2^{vK/2}} \frac{1}{\Gamma_K\left(\frac{v}{2}\right)} |S|^{v/2} |\Omega|^{-(c+K+1)/2} \exp\left(\frac{1}{2}\text{trace}(S\Omega^{-1})\right) \\
P(\theta) &= \frac{1}{\pi\gamma_\theta\left(1 + \frac{\theta}{\gamma_\theta}\right)^2}, \quad \text{where} \quad \theta \geq 0 \\
P(\beta) &= \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left(\frac{-(\beta_m - \mu_m)^2}{2\sigma_m^2}\right)
\end{aligned}$$

where I is the number of observations, J is the number of groups in the hierarchy, K is the number of hierarchical variables with varying effects, M

is the number of pooled variables with constant effects, and $S, v, \gamma_\theta, \sigma_m, \mu_m$ and γ are predefined parameters for the prior distributions.

From the above equations, it is easy to see that the resulting optimization problem is very high-dimensional and potentially impossible to solve reliably with standard non-linear optimization techniques. However, if we simplify the model and assume that there is no correlation between the parameters of the hierarchically modelled variables Z of the same level, we can model the likelihood of the parameters with univariate normal distributions, as

$$P(B|\theta) = \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\theta_k^2}} \exp\left(\frac{-(B_{j,k})^2}{2\theta_k^2}\right). \tag{2.13}$$

Then, by inserting this, together with the other necessary expressions to (2.12), but ignoring the prior of the data-level standard deviation as it does not affect the parameter values in the optimum, the MAP estimator $\hat{\Phi}$ for the coefficients can be written and simplified as follows.

$$\begin{aligned}
\hat{\Phi} =& \arg\max_{\beta,B,\theta} \quad \log(\prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(Y_i - X_i\beta - Z_iB_{j(i)})^2}{2\sigma^2}\right)) \\
&+ \log(\prod_{j=1}^{J}\prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\theta_k^2}} \exp\left(\frac{-(B_{j,k})^2}{2\theta_k^2}\right)) + \log(\prod_{k=1}^{K} \frac{1}{\pi\gamma_\theta\left(1+\frac{\theta_k}{\gamma_\theta}\right)^2}) \\
&+ \log(\prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left(\frac{-(\beta_m - \mu_m)^2}{2\sigma_m^2}\right)) \\
=& \arg\max_{\beta,B,\theta} \quad -\sum_{i=1}^{I} \frac{(Y_i - X_i\beta - Z_iB_{j(i)})^2}{2\sigma^2} - \sum_{j=1}^{J}\sum_{k=1}^{K}(-\log(\theta_k) + \frac{-(B_{j,k})^2}{2\theta_k^2}) \\
&- \sum_{k=1}^{K} \log\left(1+\frac{\theta_k}{\gamma_\theta}\right) - \sum_{m=1}^{M}\left(\frac{-(\beta_m-\mu_m)^2}{2\sigma_m^2}\right) \\
=& \arg\min_{\beta,B,\theta} \quad \sum_{i=1}^{I}(Y_i - X_i\beta - Z_iB_{j(i)})^2 + \frac{1}{\sigma^2}\sum_{j=1}^{J}\sum_{k=1}^{K}(-\log(\theta_k) + \frac{-(B_{j,k})^2}{2\theta_k^2}) \\
&+ \frac{1}{\sigma^2}\sum_{k=1}^{K} \log\left(1+\frac{\theta_k}{\gamma_\theta}\right) - \frac{1}{\sigma^2}\sum_{m=1}^{M}\left(\frac{(\beta_m + \frac{1}{\sigma^2}\mu_m)^2}{2\sigma_m^2}\right).
\end{aligned} \tag{2.14}$$

From this formulation, it can be seen that the MAP estimator reduces to a version of penalized Maximum Likelihood estimator with a squared sum of

the residuals on the leftmost term and a penalty term depending only on the parameter values after that. Thus, the MAP estimator can be considered as a type of regularized regression also in the case of hierarchical models.

2.4.2 Distribution estimation methods

One of the major advantages of the Bayesian framework is that it allows easy and intuitive inference on the uncertainty of the estimates when the complete posterior distribution is estimated. The ability to estimate and perform inferences on the posterior distribution is at the very core of Bayesian data-analysis and is often mentioned as one of the defining attributes for it. In fact, the MAP estimation discussed in the previous section is often not considered as Bayesian, as it does not yield any information on the actual posterior distribution. However, deriving analytical expressions for the posterior densities is often extremely difficult, in every other case than the very simplest ones, as the evidence $P(Y)$ often includes intractable integrals. Thus, either an approximative expression must be constructed, or the posterior density must be approximated with some numerical method. In the next two sections, we present and discuss a widely used numerical sampling method Markov chain Monte Carlo (MCMC), and especially the Hamiltonian Monte Carlo (HMC) variant of it, as well as a Variational Bayes method for approximating the analytical posterior density with optimization methods.

2.4.2.1 Markov chain Monte Carlo

In Markov chain Monte Carlo (MCMC) methods, the idea, in the context of Bayesian data analysis, is to construct a Markov chain with an equilibrium distribution that can be used to approximate the posterior distribution of the model. The posterior can then be approximated with the Markov chain by running the chain until it converges to the equilibrium and then drawing a sufficient amount of samples from the converged chain to be used as the approximative posterior distribution. The feature distinguishing MCMC methods from plain Markov chain sampling is that in MCMC, the draws are "corrected" after each iteration with different methods to better approximate the posterior distributions [Andrieu et al., 2003]. One of the most used MCMC methods is the Metropolis-Hastings algorithm, developed by Metropolis et al. [1953] and Hastings [1970], where each step of the Markov chain can either be accepted or rejected, with a probability defined as the ratio of the likelihood of the proposed next state and likelihood of the current state of the chain. Metropolis-Hastings algorithm is relatively simple, and it allows omitting the normalizing constant (i.e., the evidence), from the

expression of the posterior distribution, similarly as in the point-estimate methods. However, one must be careful in choosing the proposal distribution, from which the suggested Markov chain steps are drawn, as it has a very significant effect on the results.

Another popular MCMC variant is Hamiltonian Monte Carlo (HMC), where Hamiltonian dynamics are used to generate such proposal steps from the proposal distribution, which are accepted with higher probability [Neal et al., 2011]. The method was originally developed by Duane et al. [1987] and called Hybrid Monte Carlo, but in later literature, it is more often referred to with the more descriptive name, Hamiltonian Monte Carlo. Currently, HMC is considered one of the most progressed MCMC methods, and for example, Stan, a widely used platform for statistical computing, is built around HMC Carpenter et al. [2017]. However, with applications such as demand forecasting in a retail context, sampling methods are rarely feasible in terms of computational complexity. The reasons for this are the high model dimensionality often needed in a retail context, a high number of observations for each model, and the need to estimate a large number of these complex models repeatedly. Due to this limitation, this thesis focuses on other ways to estimate the model parameters.

2.4.2.2 Variational Bayes

The idea of the Variational Bayes (VB) method is to search for such distribution $Q(\Phi)$, that minimizes the Kullback-Leibler divergence between the variational distribution Q and the true posterior distribution $P(\Phi|Y)$ and thus works as an approximate of the posterior distribution [Jordan et al., 1999]. The KL divergence between the distributions is defined as

$$D_{KL}(Q|P) = \int_{-\infty}^{\infty} q(\Phi) \log \frac{q(\Phi)}{p(\Phi|Y)} d\Phi, \tag{2.15}$$

which can be modified to

$$\begin{aligned} \log(p(Y)) &= D_{KL}(Q|P) - \int_{-\infty}^{\infty} q(\Phi)(\log q(\Phi) - \log p(\Phi, Y)) d\Phi \\ \log(p(Y)) &= D_{KL}(Q|P) - \mathcal{L}(Q), \end{aligned} \tag{2.16}$$

and because $\log p(Y)$ is fixed, we can obtain the minimum KL-divergence by maximizing $\mathcal{L}(Q)$, also known as Evidence Lower Bound (ELBO), as also discussed by e.g., Blei et al. [2017]. Now, as $p(\Phi, Y)$ can be defined as the product of the likelihood and priors of the model which are already known, we only need to decide a suitable variational distribution q to be able to approximate the posterior with variational inference and then find the correct

parameters for the distribution utilizing some optimization method. There exists a few methods to determine the variational distributions, but one of the most common ones is the mean-field approximation, where the latent variables are assumed to be independent, and thus the total distribution can be presented as a factorization of the marginal distributions or variational factors of the latent variables. The mean-field approximation can be formulated as

$$q(\Phi) = \prod_{i=1}^{C} q_i(\Phi_i), \tag{2.17}$$

where C is the total number of variables in the model. The parametric form of each variational factor q_i can be specified differently for each variable. For example, a common choice for a continuous variable is a Gaussian factor. More discussion on the mean-field approximation can be found, for example, in Blei et al. [2017]. The parameters of the variational family can be estimated with different methods, such as coordinate ascent or gradient-based optimization methods. However, deriving the update rules for coordinate ascent or the analytical forms of gradients is often challenging even for simple models, as noted by Ranganath et al. [2014], which makes these methods inconvenient when the objective is to search for the best alternative from a group of different models. For this reason, also some methods requiring less manual work has been developed. Two good examples of these less manual methods are stochastic optimization based Black Box Variational inference by Ranganath et al. [2014] and Automatic Differentiation Variational Inference presented by Kucukelbir et al. [2017]. The basic idea in Black Box Variational inference is to use noisy but unbiased estimates of the gradients, derived as an expectation with respect to the variational factors, to update the parameters of the variational factors. The actual approximations of the gradients are obtained by sampling the derived expectation. The Automatic Differentiation Variational Inference (ADVI), on the other hand, relies on transforming the problem into a common space, where a large class of inference problems can be solved in a similar manner.

Chapter 3

Methods

3.1 Forecast models

The idea of the thesis is to compare the predictive performance of differently specified and estimated hierarchical models to the performance of non-hierarchical pooled and non-pooled models to obtain information on the benefits of hierarchical modelling of promotions in a retail context. The analysis is performed on a dataset from a European retailer, described more in detail in Chapter 4. The hierarchy structure used in the models consists of the two lowest levels, product, and product-store, of the product hierarchy, specified in Figure 1.1. This means that the hierarchical models will be estimated separately for each product, and the hierarchical and separate parameters are allowed to vary between the different product-store combinations of that product.

As the goal is to predict future demand, which can be interpreted as count data, the natural choice for the error distribution of the generalized linear models would be Poisson distribution. However, to better understand the trade-off in computational and predictive performance between different distribution choices, linear regression models are also experimented. The linear regression models are experimented both with an untransformed response variable, and with the response variable transformed to a logarithmic scale, leading to a so-called log-linear regression model. With the log-linear model, some additional steps are needed to obtain a good and unbiased fit. This is because the log-transform of the response variable also transforms the error distribution such that the mean of the distribution becomes non-zero, as explained more in detail, for example, by Teekens and Koerts [1972] and Miller [1984]. As discussed in these studies, the amount of this bias is related to the variance of the error distribution and, in this case, is equal to

$\frac{\sigma^2}{2}$, where σ^2 is the variance of the model. Thus, as explained by Beauchamp and Olson [1973], the bias can be removed by either adding $\frac{\hat{\sigma}^2}{2}$, where $\hat{\sigma}^2$ is the estimated variance of the mode, to the obtained forecasts before transforming them back to the response scale or by multiplying them with $\exp(\frac{\hat{\sigma}^2}{2})$ after the transformation, which is what is done here. In addition, to allow zero values for the response variable, a quantity of 1 is added to each of the response variable value before the transformation, and the same amount is subtracted from the fitted and forecasted values after transforming them back to the original scale, but before applying the bias correction. It is acknowledged that this affects the estimated parameter values of the models, but it is acceptable as the parameter values are not the subject of interest in this study.

The underlying hypothesis is that the predictive performance of the hierarchical models should be similar to the performance of non-pooled models for product-store combinations with a large amount of promotional data, but the hierarchical models should outperform the non-pooled models for combinations for which there is only little promotional data in the training set. The pooled benchmark models can be expected to perform relatively well for product-stores with little promotion data, but some improvements should be expected from the hierarchical models also in comparison to them.

3.1.1 Benchmark models

The linear predictor of the non-hierarchical GLMs used as a benchmark in the analysis are formulated as follows:

$$\lambda = X\beta + C\kappa, \tag{3.1}$$

where X is a matrix with k columns, which consists of the regressors related to explaining non-promotion related variance, β is a column vector of length k, containing the coefficients for the regressors in X, C represents matrix with m columns, containing separate indicator variable column for each of the m different promotion types and κ is a vector of length m containing the coefficients for each parameter. In the pooled model, one model per product is estimated, and all the estimated parameters of the model will be equal for all product-stores of the product. This means that the parameter matrices X and C contain all observations of all the product-stores within the product for which the model is estimated for. In the case of the non-pooled models, a separate model is estimated for each of the product stores. This means that the X and C of one model contain only observations of one product-store, leading to separate parameters for each of the product-stores. For the pooled

estimation, only one model containing all four promotion type indicators is estimated, whereas, for the non-pooled, separate version, two different models are fitted. The first model contains only one general promotion indicator variable in C, and the second one contains all four promotion type indicators present in the dataset.

3.1.2 Hierarchical models

Three different types of hierarchical modelling methods are considered and tested against the benchmark models. The tested modelling methods are GLMM, Bayesian GLMM, and Empirical Bayes. GLMM:s and their Bayesian counterparts are models with the hierarchy built within, as explained more in detail in Chapter 2, whereas Empirical Bayes models are single-level models, where the hierarchical structure is implemented through two-stage estimation and prior parameters. Three differently specified models are estimated with each of the three modelling methods for every product in the dataset.

3.1.2.1 Generalized Linear Mixed Models

The GLMMs used here are specified so that one model is estimated per product, which acts as the higher hierarchy level. The different stores where the product is sold are separated with different group indicators, and the parameters are allowed to vary between the different stores. The linear predictor of the used models is specified as

$$\begin{aligned} \lambda_i &= \sum_{n=1}^{N} \beta^n_{j(i)} x^n_i + \sum_{m=1}^{M} \kappa^m_{j(i)} c^m_i \\ \kappa^m_j &= \mu_{\kappa^m} + \xi^m_j, \qquad \xi^m \sim N(0, \Sigma_\theta), \end{aligned} \tag{3.2}$$

where N is the number of varying (but not hierarchical), non-campaign related regressor coefficients $\beta^n_{j(i)}$, where $j(i)$ denotes the store j to which the observation i belongs to. M is the number of hierarchically modelled campaign type indicators with coefficients $\kappa^m_{j(i)}$, which are composed of the store invariant "constant effects" μ_{κ^m} and the store-specific "varying effects" ξ^m_j that are modelled as random variables following a multi-normal distribution with zero-means and variance-covariance matrix Σ_θ. The variance-covariance matrix Σ_θ is parametrized with a vector θ, which is a vector composed of the standard deviations of the random effects. If the coefficients of the M different campaign type indicators are assumed to be independent, Σ_θ can simply be written as a diagonal matrix with elements of θ^2 on the diagonal, i.e. $\Sigma_\theta = \text{diag}(\theta) I \text{diag}(\theta)$, where I is an identity matrix. Three different models

of this form are estimated and evaluated. The first one is again with only one campaign indicator (i.e., M = 1), and the latter two are specified to contain all four promotion type indicators (i.e., M = 4). In the second model, no correlations between the campaign type variables are estimated (Σ_θ is diagonal), and in the last one, correlations between the types are estimated.

3.1.2.2 Bayesian Generalized Linear Mixed Models

The Bayesian GLMMs are specified very similarly as the frequentist GLMMs, with the difference that all of the parameters are modelled as random variables with separate probability models, or priors, and hyperparameters. In addition, the property of interest is not the maximum likelihood of the parameters but the marginal posterior densities of them. A general form of the linear predictor of these models is specified as

$$\begin{aligned} \lambda_i =& \sum_{n=1}^{N} \beta^n_{j(i)} x^n_i + \sum_{m=1}^{M} \kappa^m_{j(i)} c^m_i \\ \kappa^m_j =& \mu_{\kappa^m} + \xi^m_j, \qquad \xi^m \sim N(0, \Sigma_\theta), \\ \beta^n_{j(i)} \sim& N(\mu_n, \sigma^2_n), \\ \mu_{\kappa^m} \sim& N(\mu_m, \sigma^2_m), \\ \Sigma_\theta \sim& N(0, \Sigma), \end{aligned} \tag{3.3}$$

where $N, \beta^n_{j(i)}, j(i), M, \kappa^m_{j(i)}, \mu_{\kappa^m}, \xi^m_j$ and Σ_θ are as in Equation 3.2, and μ_n, σ_n, μ_m, σ_m and Σ are the hyperparameters of the prior distributions. The prior hyperparameters are specified as weakly informative in order to obtain regularization effect but not to restrict the estimation too much. The values for the hyperparameters are chosen separately in each model, based on the variances of the target variable y and the corresponding regressor, as proposed by Gelman et al. [2020]. Three differently specified models are tested, similarly as with the GLMMs. The first one again includes only one general promotion indicator. The second one includes four promotion type indicators but no correlations between their parameter values. The last one includes the four type indicators and estimates correlations between the parameter values. The main difference with these models compared to the GLMMs is that also the non-hierarchical parameters have prior distributions. As the priors are specified as weakly informative, they act purely as regularization. This most likely brings some advantages in situations where there is a risk of overfitting the parameters due to little data for these non-hierarchical parameters, but it also increases the model's computational complexity.

3.1.2.3 Empirical Bayes

The third type of models tested are Empirical Bayes models, where the basic idea is to obtain the prior distributions, or the hyperparameters of predefined prior distributions, from the dataset. This is slightly against the principles of Bayesian data analysis because the prior distributions should represent prior beliefs of the data and should not generally be estimated from it. However, these kinds of models have proved to be useful in many real-life applications for introducing regularization into the models as well as in modelling hierarchical structures, as the hierarchical model estimation process can be simplified with a two-stage approach. In this context, the Empirical Bayes is implemented as a two-stage estimation process of models, where we first fit a pooled GLM model on a product level to obtain the parameter estimates and the standard deviations of them that are common for all product-stores within that product, and then use those as the hyperparameters of prior distributions on the lower-level models. First, a pooled GLM where the linear predictor is of form

$$\lambda_i = \sum_{n=1}^{N} \beta^n x_i^n + \sum_{m=1}^{M} \kappa^m c_i^m, \tag{3.4}$$

is fitted on the product level as a GLM, i.e., without any prior specification for the parameters. This approach, instead of the Bayesian one, was chosen as it can be assumed that there should exist enough data on the product level to reliably estimate the parameters by simply maximizing the likelihood. The reason for this is that with a lot of data, the weight of the prior terms in the posterior, with comparison to the likelihood portion, is minimal and thus relatively insignificant. After obtaining the maximum likelihood estimates $\hat{\beta}^n, \hat{\kappa}^m$, and the standard deviations $\hat{\sigma}_{\beta^n}, \hat{\sigma}_{\kappa^m}$ of the parameter estimates of the pooled GLM, they are used as hyperparameters for the prior distributions of a non-pooled Bayesian GLM, estimated at the lowest hierarchy level. The linear predictor of this Bayesian GLM is formulated as

$$\begin{aligned} \lambda_i &= \sum_{n=1}^{N} \beta^n x_i^n + \sum_{m=1}^{M} \kappa^m c_i^m, \\ \beta^n &\sim N(\hat{\beta}^n, \hat{\sigma}^2_{\beta^n}), \\ \kappa^m &\sim N(\hat{\kappa}^m, \hat{\sigma}^2_{\kappa^m}). \end{aligned} \tag{3.5}$$

Similarly, as with GLMMs and Bayesian GLMMs, three differently specified Empirical Bayes models are estimated. The first one again has only one general promotion indicator. The second has four promotion type indicators, but no correlation is modelled between their parameter estimates. The third is a

model where the parameters of these promotion type indicators are modelled to have a correlation between them, i.e., the β^n and κ^m parameters have a multivariate normal distribution with means and covariance matrix obtained from the estimation of the pooled model. In practice, this formulation of Empirical Bayes models resembles closely the GLMMs specified in 3.1.2.1, but with the difference that the parameters of different levels are estimated separately. This is expected to provide faster estimation times as the models are less complicated. Another difference here compared to other models is that also the non-promotion related parameters are estimated as hierarchical ones, even though they are separated in the model specification to make the notation more consistent and more comfortable to follow. This is done because using the pooled estimates for hyperprior specifications also for those parameters does not increase the model's computational complexity, as some default hyperparameters should be specified for the parameters in any case. However, as there is a lot of data for estimation of the non-promotion related parameters also in the product-store level, the empirically specified priors should not have a large impact on the final estimates of the product-store level parameters unless the standard deviations of the pooled estimates are very small.

3.2 Estimation algorithms

The parameters and variance components of the Generalized Linear Mixed Models specified in Section 3.1.2.1 are estimated by maximizing a REML criterion with the L-BFGS method by utilizing the implementation in R package lme4 [Bates et al., 2015]. The implementation of the algorithm uses sophisticated computational methods to cope with the possible singularity of the variance-covariance matrix. These include composing a sparse Cholesky factor of the relative covariance factor, in order to ensure the positive-definity of the variance-covariance matrix, as well as marginalizing over certain parameters that are not of interest to simplify the REML criterion to a profiled deviance function, which is easier to maximize. A detailed description on the implementation is provided in Bates [2010].

The parameters of Bayesian GLMMs specified in Section 3.1.2.2 and the parameters of the product-store level part of the Empirical Bayes models specified in Equation (3.5), are estimated in two different ways. The first, fully Bayesian estimation method is ADVI, implemented using RStan [Stan Development Team, 2020]. Approximations of the whole marginal posterior distributions are obtained this way, but the forecast is calculated using only the means of the marginal posterior distributions as the parameter estimates.

The second way is to obtain point estimates of the parameters by performing MAP, which is implemented using RStan:s L-BFGS implementation. The speed and accuracy of the two ways to estimate the models are evaluated and discussed in order to find out which method is preferred. The most interesting comparison here is to see how large difference there is between the mean (ADVI) and mode (MAP) estimates of the posterior and whether either proves to be better in demand forecasting. Another interesting comparison for the two estimates is to analyze how much more computationally complex the mean estimation is. It is good to note that the ADVI implementation of Stan is still on an experimental stage, and it is not entirely stable, which is acknowledged when evaluating the models. The benchmark GLM models, as well as the pooled parts of the Empirical Bayes models, specified in Equation (3.4), are estimated using the glm-function of the stats package of R [R Core Team, 2020].

3.3 Evaluation of models and algorithms

The different models and estimation algorithms are evaluated in terms of forecasting accuracy, estimation time per product, and the number of excessive errors produced, i.e., reliability of the model. All three measures are needed, as in order to be able to utilize the models in continuous demand forecasting of large retail chains, the models need to be efficient to estimate and provide as accurate predictions as possible without the risk of getting unrealistic forecasts very often. To obtain reliable information on forecast accuracy, the dataset is divided into two sets, a training set and a validation set, so that the training set contains the first four years of sales data, and the validation set contains the last year. The model parameters are estimated based on only the training set, and the validation set is used to obtain predictions and calculate forecast error metrics. In general, forecast accuracy can be measured in many ways, and the most suitable method always depends on the application at hand. These different ways are discussed comprehensively, for example, by Hyndman [2014]. They cover the strengths and drawbacks of scale-dependent (absolute) measures such as Mean Absolute Deviation (MAD), scale-independent (percentage) measures such as Mean Absolute Percentage Error (MAPE) as well as scaled measures such as Mean Absolute Scaled Error (MASE). They note that the scale-dependent measures are useful when the forecasts are on the same scale, and that percentage errors are useful when the multiple forecast series with different scales are compared. One major drawback with the percentage metrics is that they are undefined for periods with 0 sales and are not very robust if the sales are

low. They present a scaled error (MASE), where absolute forecast errors are scaled with a forecast error from a naive forecast model, which addresses the main problems in both absolute and percentage errors. However, with the hierarchical models, comparison to a naive forecast is not very meaningful, and thus, it is not used here. The main forecast accuracy metrics used here are the average of product level Weighted Absolute Percentage Errors (WAPE) and the average of product level Weighted Percentage Errors (WPE), defined for product p as:

$$\text{WAPE}_p = \frac{\sum_{s=1}^{S}\sum_{t=1}^{T}|y_{pst} - \hat{y}_{pst}|}{\sum_{s=1}^{S}\sum_{t=1}^{T} y_{pst}} \tag{3.6}$$

$$\text{WPE}_p = \frac{\sum_{s=1}^{S}\sum_{t=1}^{T}(y_{pst} - \hat{y}_{pst})}{\sum_{s=1}^{S}\sum_{t=1}^{T} y_{pst}}, \tag{3.7}$$

where y_{pst} is the observed sales quantity for product p in store s on a day t and $\hat{y}_{pst}$ is the forecasted amount for the same product-store-day. These metrics allow a meaningful comparison between forecast accuracies of the models with a dataset including products with a different level of sales, so that also the accuracies of forecasts of products with lower sales have an effect, without having problems with undefined or unreliable percentage measures. In addition to the mean WAPE, also product-store-day level MAD of the forecast models is analyzed in order to allow comparison of the absolute forecast errors. As the main point of interest here is the campaign forecasting ability of the models, only days with promotional activity are included in the data when calculating the error metrics.

In this application, one of the mains benefits sought by hierarchical campaign forecasting models is that they can often provide reasonable forecast also in cases where little or no campaign data is present in the past on some of the lowest hierarchy levels, without a separate process for them. For this reason, the prediction accuracy of models in these cases is of particular interest. However, the dataset used in the analysis contains so many promotion days that analyzing this effect directly would be difficult. Therefore, the dataset is modified so that for some stores, information of promotion days in the training set is removed by setting the promotion related indicator variables to zero but adding a new control variable to those periods to control the variation in sales caused by the promotions. The value of this control variable is always zero in the validation set, making sure that no unwanted information is passed from the model estimation phase to the forecasting phase in these situations. Different amounts of past promotion data are removed for different stores to obtain information on how the models behave with a different

number of data points. The exact percentages of the removed promotion data per store are presented in Chapter 4.

Retail demand forecasts are often used to automate the replenishment of retail chains, which requires reliability for the used forecast models. For this reason, also the number of weeks with excess forecast errors are evaluated for each model by counting the weeks for which the percentage error of the total forecast exceeds 1500%. These weeks are labeled as "outliers", and the total amount of these outlier weeks is counted and analyzed per forecast model. As the percentage errors can be quite large if the sales of the week are very low, we do not want to count a week as an outlier in those cases. Thus, weeks where the total sales are less than 10% of the average weekly sales of the product-store in question, are excluded from this count.

Chapter 4

Data

The data for testing and comparing the methods of this thesis is obtained from a large European retailer. It consists of 1000 products sold in 11 different stores, yielding a total of 11 000 product-store combinations. All of the stores are located in the same country within a reasonably small geographic area. This allows us to model the sales hierarchically across different stores, as the sales behaviour and effect of the promotions can be assumed to be similar in different stores. For each product-store combination, there are five years of sales data recorded in recent history. Each row of the dataset describes the number of units sold of a product in a specific store on a particular day, together with the price with which the units have been sold. In addition to the sales quantities, the dataset includes indicators for four different types of promotional activities for each day and information on different categorical groupings for the products. Two extremely outlying observations were removed from the dataset, as they were clearly caused by errors in some part of the data gathering process and are something that should not be taken into account when assessing the performance of the models.

The number of promotion days of a single product-store combination during the 5-year period ranges from 34 to 1827, with a sample mean of 619.4 and a median of 483 days. The daily sales quantity in units ranges from 0 to 9536, with a sample mean of 4.24 and a median of 2. Due to the characteristics of the customer's products, there is also a high number of zero-sales days in the dataset, a sample mean of the number of zero-sales days for one product-store combination during the recorded period is 497, and a median is 504 days.

A part of the histogram of the daily sales numbers, with x-axis range limited from the right-hand side for clarity, is presented in Figure 4.1. As can be seen from the figure, the distribution of sales is relatively close to a log-normal, Poisson, or negative-binomial distributions with the right param-

eters. Also, some example time series of the products are presented in Figure 4.2, where the daily sales are presented for a duration of one year for four different products, each in three different stores. Each subfigure represents one product, and the different time-series in them represent different stores. The promotion periods are illustrated with a light blue highlight. From the sales graphs, it can be seen that the daily sales numbers vary significantly from one day to another and that the promotional activities have a significant effect on the level of sales. It also illustrates that the sales behaviour of the products is similar, but not identical, across the different stores, which justifies the hierarchical approach. Of course, these are only a few examples of many product-stores, only to demonstrate how the time series can look like, so no general conclusions about the behaviour of the PL:s can be drawn from these figures.

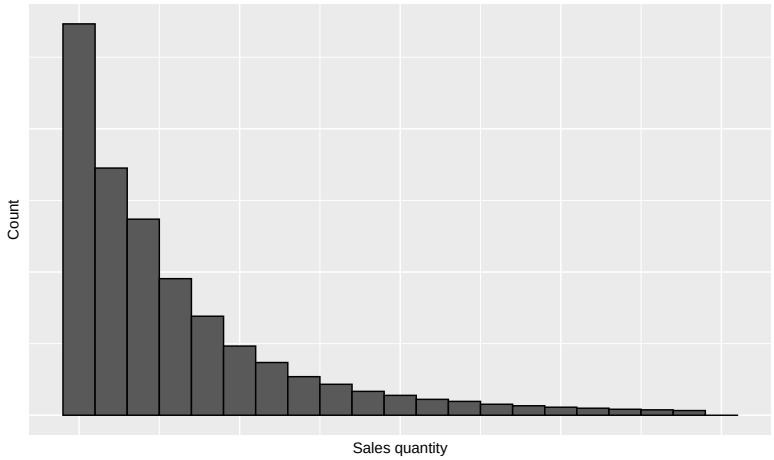
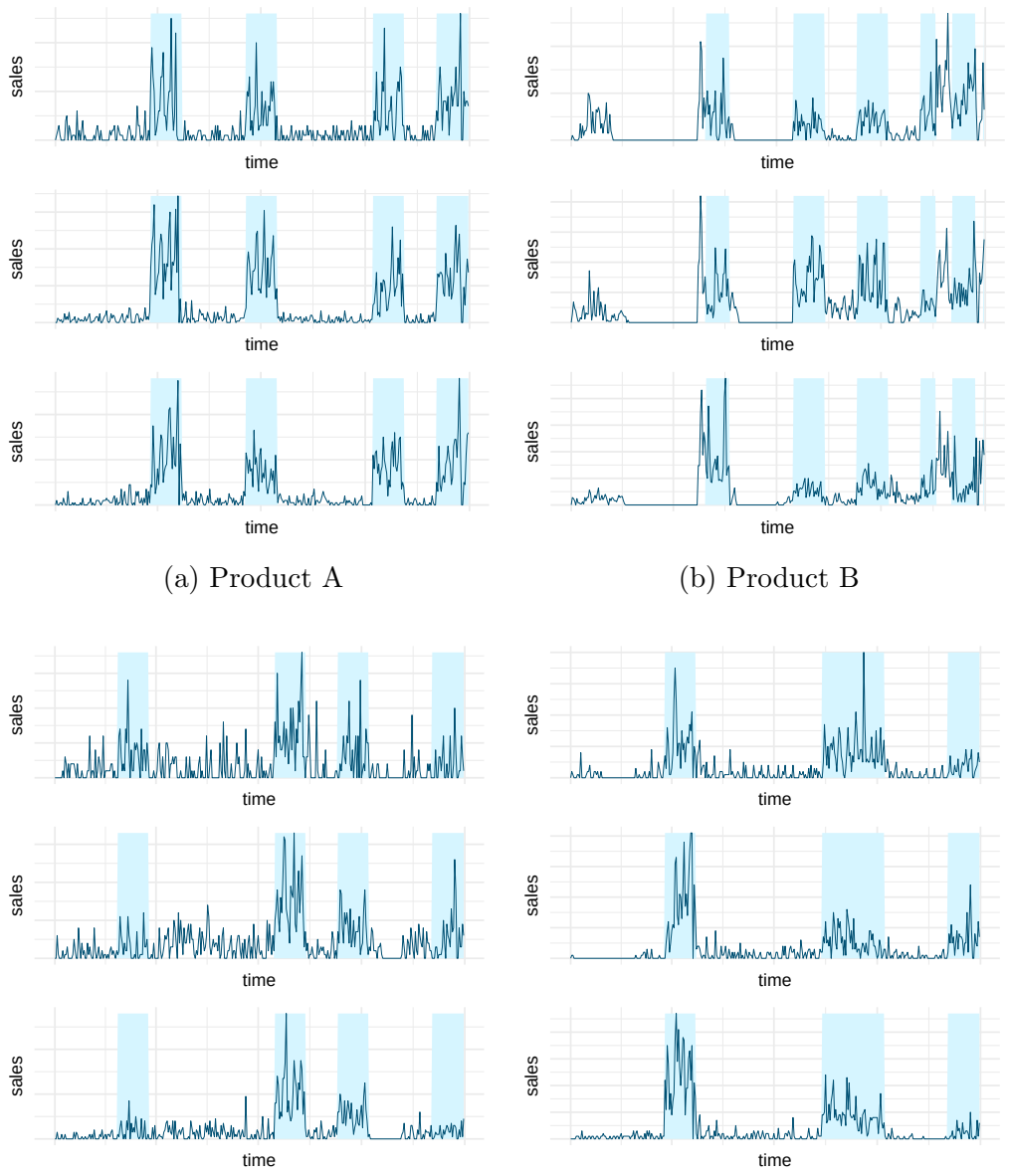


Figure 4.1: Histogram of daily sales. Axis scales are removed to preserve anonymity of the data



(a) Product A

(b) Product B

(c) Product C

(d) Product D

Figure 4.2: Time series presenting the daily sales of one calendar-year, for four different products, each in three different stores. The promotion periods are illustrated with a light blue highlight. Axis tick labels are removed to preserve anonymity of the data.

The data is divided into training and validation sets by assigning the first four years to act as a training set and using the last year as a validation set. Section 3.3 discussed the need to remove some of the promotion data

from the training set. The percentages of promotion data removed from the training set for each store are presented in Table 4.1. The data is removed by randomly chosen the specified percentage of days with promotion activity from each product-store combination and then setting the promotion indicator variables to zero for those days. Another variable is set to 1 for those days to control the variation caused by the promotional activities. This control variable always has the value 0 in the validation set.

Table 4.1: Percentages of promotion data removed from the training set per store.

Store id:s	% of promotion data removed
1-7	0
8	80
9	90
10	95
11	100

Chapter 5

Results

In order to understand the advantages and disadvantages of different approaches of hierarchical models, 15 different forecast models were experimented with three different regression types by estimating the model parameters with the training set and calculating forecasts and forecast accuracy metrics for the succeeding year using the validation set. Each of these models was described in detail in Chapter 3, but a recap of the model properties and the shorthands used in this chapter are presented in Table 5.1. In general, the shorthand GLM refers to the non-hierarchical Generalized Linear Models (Equation (3.1)) used as benchmarks for the hierarchical models. GLMM refers to the Generalized Linear Mixed models from Equation (3.2), BGLMM to the Bayesian Generalized Mixed Models from Equation (3.3), and the shorthand EB refers to the Empirical Bayes models presented in (3.4) and (3.5). For the benchmark models, P refers to the pooled model where one model is estimated per product, and S refers to the separate models where each product-store gets its own model. The way the promotions are categorized is indicated with a number after the model. Number 1 refers to the models where promotions are handled with only one, generic promotion indicator variable, i.e., no differentiation between the types of promotions are done, number 2 refers to models where the promotions are categorized into four different promotion types, but no correlations between the parameter estimates of the promotion types are estimated. Number 3 refers to models where the promotions are again categorized into four different types, but now correlations are estimated between the parameter estimates of the types. For Bayesian models, the estimation algorithm is also specified, MAP referring to models that are estimated with a Maximum a Posteriori approach and VI to models estimated with Automatic Differentiation Variational Inference procedure.

Table 5.1: Properties of tested models

Model	Reference	Hierarchical	Bayesian	M	Correlations
GLM P	(3.1)			4	
GLM S1	(3.1)			1	
GLM S2	(3.1)			4	
GLMM 1	(3.2)	X		1	
GLMM 2	(3.2)	X		4	
GLMM 3	(3.2)	X		4	X
BGLMM 1	(3.3)	X	X	1	
BGLMM 2	(3.3)	X	X	4	
BGLMM 3	(3.3)	X	X	4	X
EB 1	(3.4) & (3.5)	X	X	1	
EB 2	(3.4) & (3.5)	X	X	4	
EB 3	(3.4) & (3.5)	X	X	4	X

5.1 Initial results

At first, the models were estimated using the sales data from the years 2012 - 2015 as a target variable. However, this approach resulted in significant overforecasting with all models. The Figure 5.1 presents the aggregate sum of sales of all products within the scope of this thesis, for both training and validation periods, together with aggregate sums of the fits and forecasts with three example models. From this figure, the reason for the consistent bias in the forecasts is easy to see, as there is a change in the aggregate level trend within the last three years, making it impossible to fit the models properly to this sort of data.

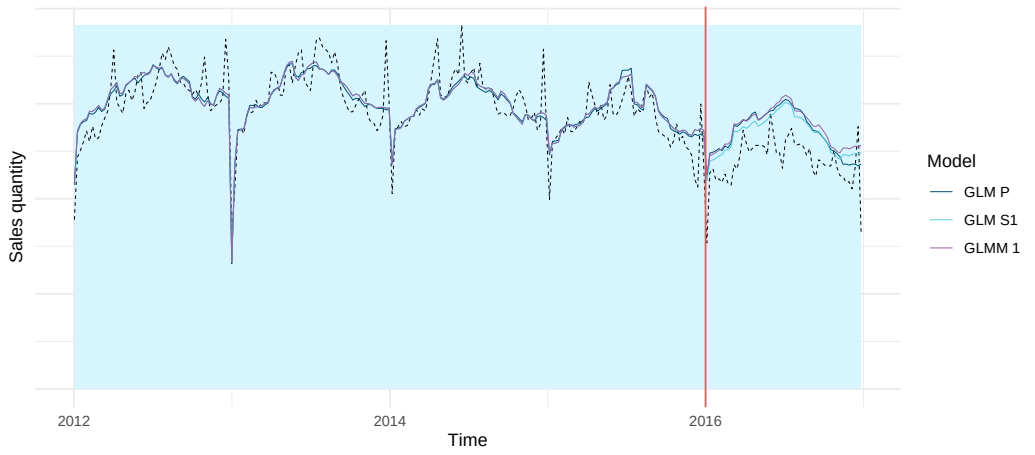


Figure 5.1: The aggregate sum of sales and three example forecast models. The red vertical line illustrates the division to training and validation sets

This difference between the sales pattern of the validation set and the sales pattern of the training set makes a meaningful comparison of the different models very hard. Thus, the first year of sales was decided to be dropped from the training set to make the descending trend more evident to the models and to align better the sales pattern of the validation set with the one in the training set. All further analysis is made with this modified training set.

5.2 Estimation time

One of the objectives of the thesis was to analyze the estimation times of differently specified and estimated hierarchical models, to understand which models are feasible to use in a continuous large scale demand forecasting setup and to get a sense of how the model complexity affects the estimation times in different cases. The times are reported in seconds and calculated as a mean time taken to estimate one product-level model, or all store-specific models of one product, depending on the model type. The mean times across all of the products in the dataset are presented in Table 5.2. The GLMM and Bayesian GLMM models proved to be too complex to estimate as a Poisson regression in a feasible time, so the results of those models are not presented. Similarly, estimating the parameters for Bayesian GLMM models with the Variational Inference approach also turned out to be too complex to estimate for this scope in a reasonable time, which is why those are not presented either.

Table 5.2: Average estimation times per model in seconds

Model	Linear	Log-linear	Poisson
GLM P	0.03	0.03	0.12
GLM S1	0.37	0.39	1.06
GLM S2	0.56	0.56	2.03
GLMM 1	2.38	2.63	NA
GLMM 2	7.64	7.72	NA
GLMM 3	27.17	26.7	NA
BGLMM 1 (MAP)	422.24	546.95	NA
BGLMM 2 (MAP)	582.17	1881.57	NA
BGLMM 3 (MAP)	1223.2	2763.31	NA
EB 1 (MAP)	2.22	1.73	1.95
EB 2 (MAP)	4.61	3.79	3.41
EB 3 (MAP)	6.5	6.04	6.14
EB 1 (VI)	77.19	114.82	154.64
EB 2 (VI)	85.54	128.65	174.31
EB 3 (VI)	117.23	263.89	315.46

As can be seen from the table, the average estimation times vary significantly between the different models and algorithms. As expected, the benchmark models are the fastest to estimate as they are also the simplest ones. From the hierarchical models, the Empirical Bayes models estimated with MAP are the most efficient, with the estimation time increasing with the more complex models but varying only little between the different regression types. The GLMMs are also relatively fast, but their running time seems to increase significantly with the more complex models. The Variational Inference estimation of the Empirical Bayes models, and especially the Bayesian GLMMs are significantly slower than the other options, and due to the slow estimation times, these models can not be used in large scale retail demand forecasting. However, the forecast accuracy of them is still interesting to compare to other models to understand each approach's benefits. Based on the estimation times, the most promising approach seems to be the Empirical Bayes models with MAP. Another feasible approach would be a simple linear or log-linear GLMM, but anything else seems to be too heavy to use in practice. Another version of this table, with also the maximum and minimum estimation times for each forecast model is presented in Appendix A.1. That table indicates that the distribution of the estimation times is not symmetric, as the minimum times are always much closer to the point estimates than the maximum times. This means that the most of the model estimation times are expected to be less than the mean time, but the

estimation of some models might take considerably longer than that.

5.3 Excess errors and reliability

Another subject of interest was the reliability and excess errors produced by the forecast models and fitting algorithms, which is evaluated here by assessing the number of weeks where the forecast is considerably higher than the sales of that week and can thus be considered an outlier. These weeks were defined to be the weeks where the percentage error of sum of the daily forecasts of that week is at least 1500%, and the sales of that week were at least 10 % of the average weekly sales of that product-store. The total number of these weeks across the whole research scope during the validation period is presented in Table 5.3, separately for each of the tested forecast models. The results in the table indicate that the number of outlier forecast weeks varies quite significantly between the different regression types. However, it is rather similar between the different models within the regression types, except for the models estimated with the Variational Inference approach. Linear regression seems to produce models with the most reliable performance, while especially the log-linear models tend to have more outlying forecasts. An interesting result is that with the log-linear and Poisson regression models, the EB models estimated with Variational Inference seem to be more reliable than with linear regression, but the benchmark GLMs seem to have more outliers, especially with Poisson regression. In addition, the Empirical Bayes models that are estimated with MAP seem to again stand out by having the lowest outlier counts within each of the regression types. However, in this case, the other hierarchical models, apart from the VI estimates, do not seem to fall much behind.

Table 5.3: Number of weeks with outlier valued forecasts in the validation set.

Model	Linear	Log-linear	Poisson
GLM P	25	84	28
GLM S1	6	50	90
GLM S2	6	41	105
GLMM 1	6	50	NA
GLMM 2	6	39	NA
GLMM 3	6	45	NA
BGLMM 1 (MAP)	6	61	NA
BGLMM 2 (MAP)	6	54	NA
BGLMM 3 (MAP)	8	48	NA
EB 1 (MAP)	3	25	4
EB 2 (MAP)	4	25	3
EB 3 (MAP)	21	75	24
EB 1 (VI)	104	31	3
EB 2 (VI)	348	28	5
EB 3 (VI)	173	79	30

5.4 Forecast accuracy

One of the most interesting criterion when evaluating forecasting models is forecast accuracy. Here, the primary forecast accuracy metric used is calculated as the mean of the product level WAPEs, which are calculated on a daily level for the duration of the validation period, only taking the days when the product has been in promotion into account, as explained previously in Section 3.3. In addition to the WAPEs, also the mean absolute daily deviations (MAD) and mean daily weighted percentage errors (WPE) of the forecast models are calculated and analyzed to understand how the models work for products with different scales of sales and to find out about any possible biases in the forecasts.

The mean values of WAPEs for each forecast model are presented in Table 5.4. The table is ordered so that it is easier to compare different implementations of the same base model and the respective benchmark, and the best value for each regression type is bolded. The WAPE values range from 78.5 % to 84.5 %, which can seem high, but considering that the metric is calculated as an average daily error of the promotion days of the whole one-year forecast horizon, the numbers can be considered very reasonable. As shown in the table, the Empirical Bayes (MAP) models seem to again perform the best,

with GLMMs often being the close second in their respective categories. In fact, these two model types are the only ones to systematically obtain better forecast accuracy than the respective benchmark models, as the other model types do not seem to offer any advantage over the non-hierarchical benchmark models regarding this metric. In addition, linear regression seems to be roughly on par with the other regression types, and no significant advantage seems to be obtained by allowing the correlations between the variables.

Table 5.4: Mean values of product-level WAPES over the validation set

Model	Linear	Log-linear	Poisson
GLM P	0.8328	0.8342	0.8313
GLM S1	0.8296	0.8419	0.8451
GLMM 1	0.8317	0.8384	NA
BGLMM 1 (MAP)	0.8333	0.8445	NA
EB 1 (MAP)	0.8207	0.8276	0.8334
EB 1 (VI)	0.86	0.8314	0.8349
GLM S2	0.8008	0.8326	0.8201
GLMM 2	0.7965	0.8187	NA
BGLMM 2 (MAP)	0.8117	0.8656	NA
EB 2 (MAP)	**0.7852**	**0.8051**	**0.7956**
EB 2 (VI)	0.8443	0.8093	0.7974
GLMM 3	0.7970	0.8200	NA
BGLMM 3 (MAP)	0.8086	0.8602	NA
EB 3 (MAP)	0.818	0.8244	0.8215
EB 3 (VI)	0.8803	0.8256	0.8265

In addition to WAPEs, to make sure that the models perform well also on fast-moving products, the Mean Absolute Deviation of the forecasts was also calculated as that provides information on the average daily forecast errors across all of the products and is more sensitive to the errors of high-selling products. The daily MAD-values for each forecast model are presented in Table 5.5. The values vary around 3.5 units per day, and as can be seen from the table, the GLMM and EB (MAP) models again systematically outperform the respective benchmark models, whereas the BGLMM and EB (VI) models outperform the benchmarks only occasionally. With linear and Poisson regression, the EB (MAP) models perform the best, with GLMMs being the close second. With log-linear regression, the order of the two models is reversed. In general, the difference between GLMMs and EB (MAP) models is a lot closer when measured with MAD than with WAPE, which indicates that GLMMs might perform slightly better for higher selling products, whereas

EB (MAP) models would then work better for slow-movers. Otherwise, the models' relative performance is somewhat similar with regards to MAD as with regards to WAPE, indicating that no systematic differences regarding the pure sales quantity of product should exist with these models.

Table 5.5: Daily MAD values over the validation set

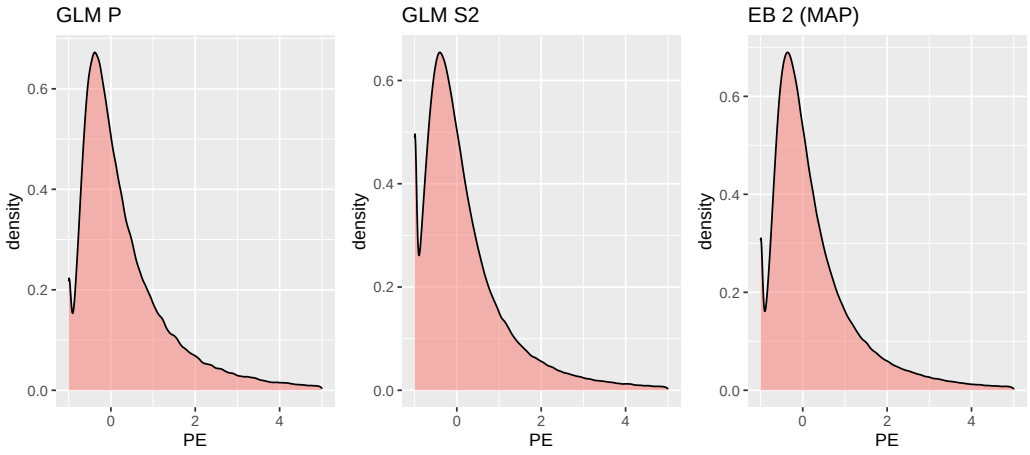
Model	Linear	Log-linear	Poisson
GLM P	3.5146	3.9429	3.5638
GLM S1	3.3783	3.8206	3.5096
GLMM 1	3.3605	**3.7955**	NA
BGLMM 1 (MAP)	3.3634	3.8528	NA
EB 1 (MAP)	3.3053	3.8211	3.4277
EB 1 (VI)	3.5795	3.8367	3.4357
GLM S2	3.3272	3.8775	3.4731
GLMM 2	3.2855	3.8098	NA
BGLMM 2 (MAP)	3.4421	3.9981	NA
EB 2 (MAP)	**3.2141**	3.7986	**3.3135**
EB 2 (VI)	3.6998	3.8245	3.3207
GLMM 3	3.2856	3.8213	NA
BGLMM 3 (MAP)	3.4722	3.9295	NA
EB 3 (MAP)	3.4317	3.8861	3.5031
EB 3 (VI)	4.2433	3.8873	3.5159

To complement the forecast accuracy measures, it is also good to analyze the possible biases in the forecasts as obtaining non-biased forecasts is also very important in demand forecasting. The mean daily bias-%, calculated as a mean WPE across all products, is presented in Table 5.6 for each of the tested forecast models. Surprisingly, the least biased forecast based on this measure is obtained the benchmark model 3, with a bias-% of roughly 13.6% for linear and log-linear regressions and 16.6% for Poisson regression. The positive percentages here indicates that the models tend to overforecast the promotions on an aggregate level, which is most likely related to the shift in sales pattern in the validation set discussed in Section 5.1, despite the efforts to minimize the effect. From the hierarchical models, the least biased ones seem to be EB (MAP) 2 and 3, depending on the regression type. It is also interesting to note here that the log-linear regression seems to produce the least biased forecasts here, even though that regression type performed the worst regarding the forecast accuracy measures.

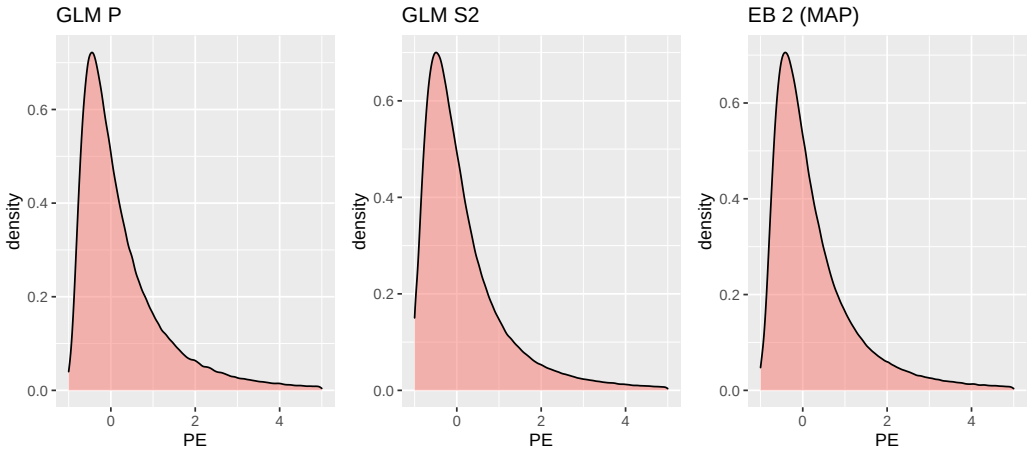
Table 5.6: Mean values of product-level WPE:s over the validation set

Model	Linear	Log-linear	Poisson
GLM P	0.1964	0.1444	0.1931
GLM S1	0.2124	0.1771	0.233
GLMM 1	0.2731	0.2251	NA
BGLMM 1 (MAP)	0.2749	0.2252	NA
EB 1 (MAP)	0.2476	0.2225	0.2875
EB 1 (VI)	0.2854	0.2264	0.2875
GLM S2	**0.1355**	**0.1363**	**0.1656**
GLMM 2	0.1901	0.1759	NA
BGLMM 2 (MAP)	0.1992	0.2052	NA
EB 2 (MAP)	0.1677	0.1715	0.2106
EB 2 (VI)	0.2267	0.174	0.2097
GLMM 3	0.1901	0.1777	NA
BGLMM 3 (MAP)	0.2005	0.2008	NA
EB 3 (MAP)	0.1843	0.1439	0.1951
EB 3 (VI)	0.2493	0.1449	0.1988

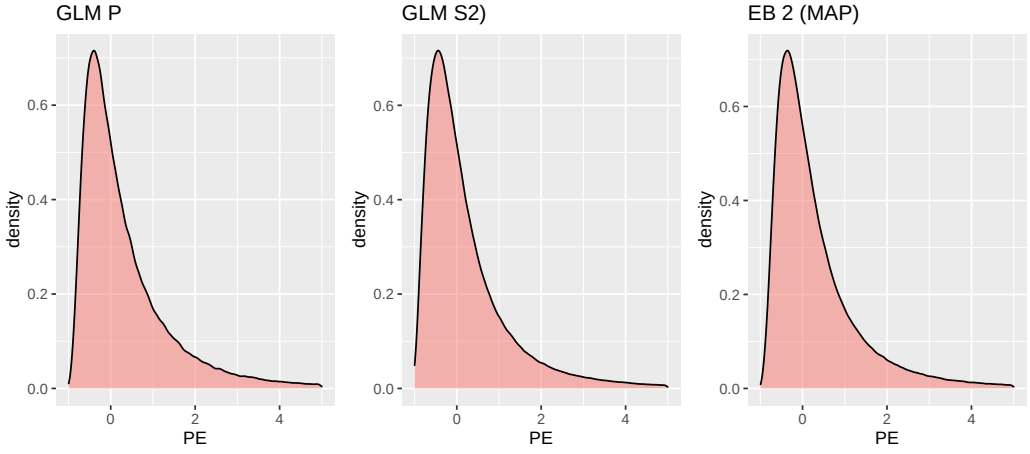
While the bias-% of roughly 15-25% might seem relatively high, it is good to remember that the metric values were calculated as a sample mean from the forecasts of only campaign days and that some promotion data was removed from the training set from certain stores. This removal explains, at least to some extent, the smaller bias of GLM 2 and 3 models, as, for example, without campaign data in the training set, campaigns of the future are not forecasted, causing significant underforecasting for some product-store combinations and thus decreasing the total overforecasting bias. Thus, it is also essential to look at the distribution of the forecast errors and the store-specific forecast accuracy measures. The approximated densities of the error distributions on a total level of all regression types and three different example forecast models are presented in Figure 5.2, while the density graphs of all models are presented in Appendix A.2, in Figure A.1. The x-axis of the figures is restricted to 5 for clarity, as the outlying observations were analyzed already in the earlier parts. The figure illustrates rather clearly that there are no large differences between the error distributions of different forecast models within the same regression type. Also, the different regression types result in somewhat similar error distributions, with the exception that the linear regression results in more forecasts of 0 (PE of -1), as all negative forecasts are capped to zero. The density curves also indicate that the forecast error is actually negative for most of the days, and the outlying forecasts cause the average percentage errors to be positive for all of the models.



(a) Linear regression



(b) Log-linear regression



(c) Poisson regression

Figure 5.2: The approximated density of the daily percentage error distributions of different forecast models and regression types.

As there are so many different models and regression types, reporting store-specific results for all of them would be rather heavy and possibly confusing, and as the linear regression has proved to be a competitive option for the other regression models, only the results of the linear regression are reported here. Also, the store-specific results of other regression models were analyzed, and the conclusions and insights found from the linear regression results also apply to them, with similar differences in the error levels as in the total level results. The store-specific results of the other regression models can be found from Appendix A.3. Table 5.7 presents the MAD values for each of the stores. The values of stores 1 to 7 are averaged together as no promotion data was removed from those stores, and thus there should not be any major differences between them. The results for the different stores seem to be roughly similar to the total level numbers, the main difference being that the pooled benchmark model GLM P performs better than the separate benchmarks GLM S1 and S2 for stores with little campaign data, as separate models naturally suffer the most from the lack of store-specific data. In contrast, the hierarchical and pooled models utilize the campaign data from other stores. It is also notable that in store 11, most of the hierarchical models outperform the benchmark models, but in all other stores, the results are very similar to the total level numbers, where the GLMM and EB (MAP) models are the only ones to systematically provide more accurate forecasts than the benchmarks.

Table 5.7: Mean Absolute (daily) Deviation of each store group.

Stores	1-7	8	9	10	11
GLM P	3.5029	3.3258	4.1694	3.2768	3.3638
GLM S1	3.3168	2.8412	4.614	3.0293	3.4376
GLMM 1	3.3145	2.8424	4.5514	3.007	3.3434
BGLMM 1 (MAP)	3.317	2.8435	4.5521	3.0239	3.3394
EB 1 (MAP)	3.2598	2.9536	4.2059	3.1119	3.2541
EB 1 (VI)	3.3124	3.8625	4.5327	4.5176	3.2599
GLM S2	3.2506	2.8182	4.5226	3.0454	3.4376
GLMM 2	3.2431	2.8097	4.399	2.9553	3.2562
BGLMM 2 (MAP)	3.3948	2.9325	4.5759	3.1757	3.3968
EB 2 (MAP)	**3.1636**	2.9063	**4.0984**	3.0175	**3.175**
EB 2 (VI)	3.7047	3.0133	4.2422	3.1602	4.3237
GLMM 3	3.2428	**2.8059**	4.4247	**2.9389**	3.254
BGLMM 3 (MAP)	3.4253	2.9982	4.5673	3.1564	3.4772
EB 3 (MAP)	3.3858	3.2811	4.1525	3.2414	3.3657
EB 3 (VI)	4.1357	3.3122	4.1493	6.8273	3.3846

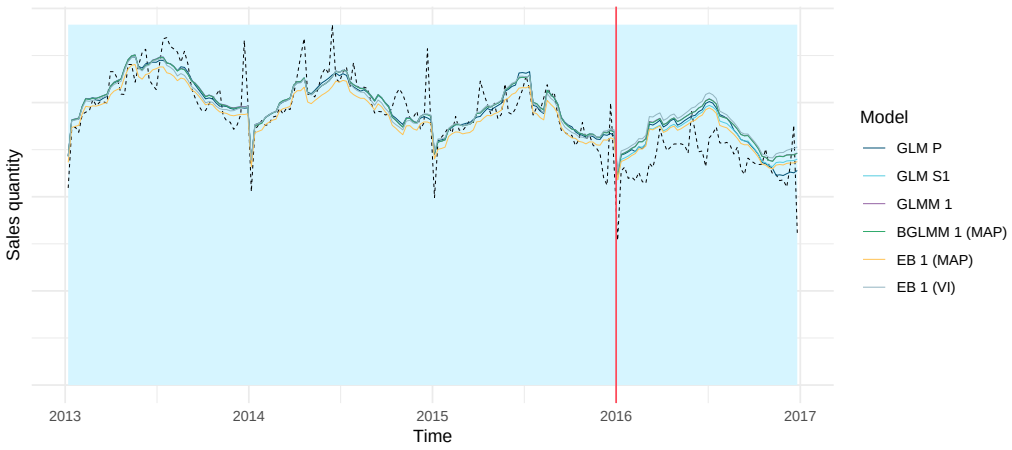
The product-store level WPE values are presented in 5.8. Also regarding this metric, the results are rather similar to the total level ones, apart from store 11. In store 11, the GLM S1 and S2 models show a negative bias, which is understandable because those models do not have any information about the promotions in the training set and can not forecast the effects of them in any way. Interestingly, the absolute value of the bias for the separate models is lower than the ones of the pooled and hierarchical models, indicating that less bias is present in the forecast if the campaigns are not acknowledged in any way. However, this finding is again most likely due to the small shift in sales that remains in the dataset when moving from the training set to the validation set, even though the dataset was modified to mitigate this effect. Another possible reason for that is the fact that the percentage error metrics tend to favor underforecasting models as the underforecasting error is capped to -100%, while there are no limits for overforecasting error. No major difference between the hierarchical models with the different amounts of promotion data can be found, at least based on these metrics, so it can be assumed that the conclusions drawn from the total level results should be quite well generalizable to models with different amounts of data.

Table 5.8: Product-store level WPE of each store group.

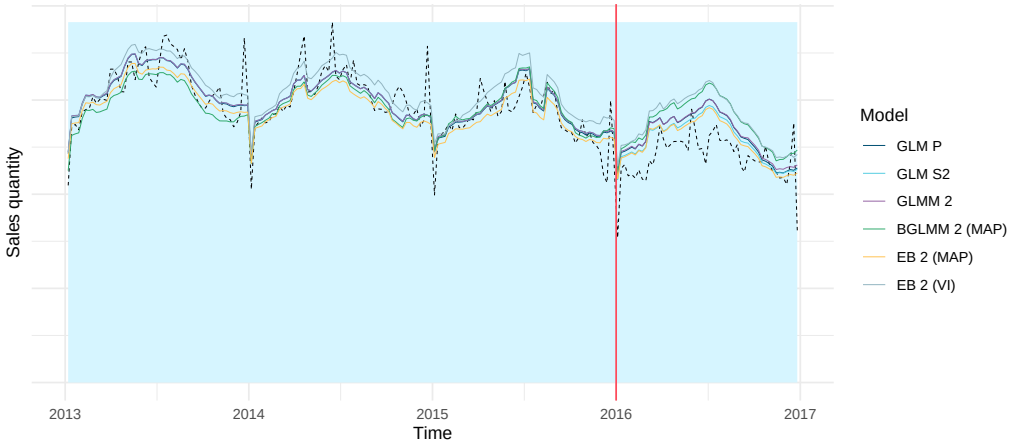
Stores	1-7	8	9	10	11
GLM P	0.5316	1.0884	0.7663	0.5678	0.4432
GLM S1	0.3522	0.3812	1.3815	0.3552	-0.2914
GLMM 1	0.3552	0.4123	1.3413	0.3868	0.4984
BGLMM 1 (MAP)	0.3598	0.4361	1.3346	0.4042	0.5054
EB 1 (MAP)	0.4674	0.5681	1.0435	0.5592	0.4812
EB 1 (VI)	0.4792	1.037	1.0574	0.5758	0.4799
GLM S2	**0.2544**	**0.2957**	0.8152	**0.243**	**-0.2914**
GLMM 2	0.2703	0.3256	0.807	0.2867	0.3564
BGLMM 2 (MAP)	0.2906	0.3609	0.7769	0.3272	0.3161
EB 2 (MAP)	0.3291	0.4782	0.8131	0.453	0.3612
EB 2 (VI)	0.3667	0.4968	0.8491	0.4781	0.7481
GLMM 3	0.2669	0.3062	0.8326	0.2721	0.353
BGLMM 3 (MAP)	0.3035	0.3797	0.8204	0.3236	0.3262
EB 3 (MAP)	0.4687	1.0234	**0.7642**	0.5498	0.4478
EB 3 (VI)	0.5151	1.0345	0.7667	0.9073	0.4516

In addition to the accuracy metric values, it is also interesting to analyze the time-series graphs for the forecast and sales. The forecasts obtained with linear regression are presented in Figure 5.3. From the plots, it can be seen

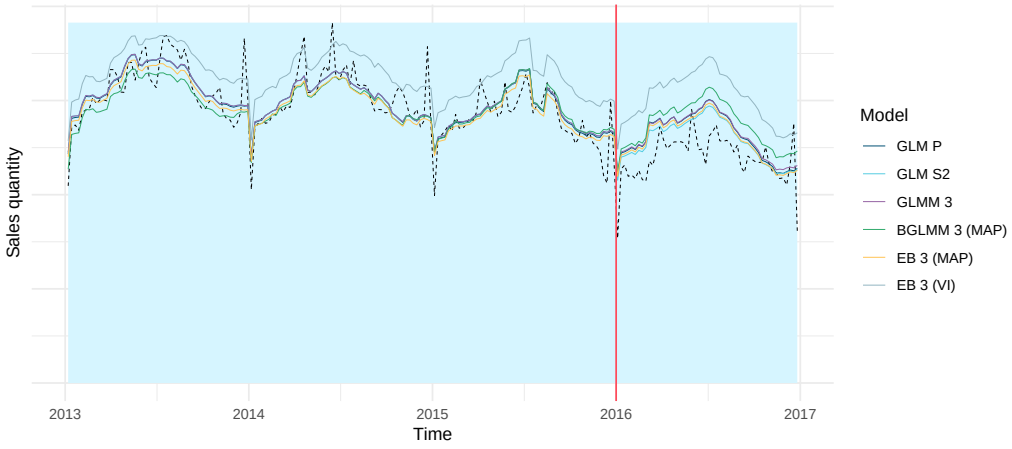
that the reason for the BGLMMs and EB (VI) models struggling to beat the benchmarks is that they produce significantly higher forecasts throughout the validation period than the other models, especially with log-linear and Poisson regression. The GLMMs and EB (MAP) models, on the other hand, are roughly on the same level as the benchmarks, with the GLMMs closer to the pooled models and EB (MAP) models closer to the separate models. However, it must be acknowledged that these plots are aggregated and do not necessarily reflect how the individual product or product-store-specific forecasts behave.



(a) Models 1



(b) Models 2



(c) Models 3

Figure 5.3: The total level aggregate sum of sales and forecasts as a time series graph. The red vertical line illustrates the division to training and validation sets

5.5 Discussion

Based on the results presented above, the models' main differences are related to the estimation time and reliability of the forecast models, while there is little difference in the forecast accuracy metric values for different model specifications. This is not surprising as many of the models are rather similar. The differences in the specifications are mainly related to the estimation procedures and to the behaviour of the models in situations when only little data is available for fitting. In addition, the slight shift in the seasonal pattern and trend from the training set to the validation set affects the forecast accuracy measures, making it challenging to draw any concrete conclusions when the accuracy metric values are so close to each other. It is also good to keep in mind that all of the results presented in this thesis are aggregate level results. While these results have been verified but not reported here in order to keep the results section easier to follow, by analyzing a few randomly chosen product level models, some of the individual product level models might include some systematic behaviour invisible in the aggregate level results and not caught by the sanity checks performed on the product level models. On a general level, all of the models were able to forecast the campaign effects rather well on all analyzed situations, and the small amount of campaign data in the training set for some stores did not seem to have a considerable negative effect on the performance. This indicates that the hierarchical models offer a promising solution to problems discussed earlier in this thesis. Not all of the models were efficient enough to estimate to be used to support the decision making in the replenishment of large retail chains, but GLMM and Empirical Bayes models seemed promising on that front. An interesting thing to note is that even though there are large differences in the outlier counts of different models, that difference does not transfer into forecast accuracies to such extent. There might be many reasons for this, but the most likely one is that most of the outliers observed are coming from the same few products and thus contribute only little to the accuracy metrics. This would indicate that the problems are more in the estimation phase and related to the convergence of the algorithms rather than to the actual instability of the model specification.

A few main observations seem to stand out from the presented results consistently. Firstly, the MAP estimated Empirical Bayes models seem to stand out from the tested models by yielding either the best results or results that are on par with the best ones with regards to every tested attribute. One possible reason for this is that dividing the estimation procedure into two relatively well-optimized and straightforward steps ensures the convergence of

the algorithm to a good solution better than the models where the estimation is simultaneous. The second reason for this difference could be that the EB models are the only ones where also the other parameters than campaign type indicators are modelled as hierarchical, as it would be computationally too heavy with any other models. With continuous variables, the effect of the prior distribution should be very insignificant, as there is so much data for them that the likelihood portion of the model outweighs the priors. However, it might be that the regularization that is dependent on the pooled model directs the optimization into better models regardless. This means that there should be no need to use the models that are more complex to estimate when the data is similar to the one used here, as no significant advantages can be obtained with them.

Secondly, the fully Bayesian estimation with Variational Inference does not seem to offer significant advantages over the MAP estimation of the Bayesian models in this application, indicating that the maximum of the posterior is at least as good of an estimate as the mean of the posterior is. This means that unless information about the marginal posterior distributions of the estimated parameters is specifically needed, using the more straightforward MAP estimation procedure should be sufficient, and thus a preferred option. The reasons for not getting advantages from estimating the full posterior might be that the mode of the distribution is very close to the mean, meaning that the posteriors are symmetric, and thus there is not much difference in the two point estimates, or that estimating the full posteriors is such a complex problem that the mean estimates are just not reliable. Another consistent finding is that the type 2 models, where the campaigns were divided into four groups, are generally more accurate than the type 1 models with no such grouping. However, no advantages seem to be obtained when the correlations are added into such models. There can be various reasons for this, but the most probable explanation is that there simply are not significant enough correlation between the parameters of the different campaign types to make a difference, or that estimating the correlations between them is not necessary as all the needed information can always be obtained from the other groups within the same hierarchical model.

An important finding is also that despite being a theoretically wrong model for count variables, the linear regression seems to produce at least as accurate forecasts as Poisson and log-linear regression models in this application. This is interesting as linear regression is the most time-efficient option, making it unnecessary to use more complex generalized regression models for this application. The negative forecasts produced by the linear model can be transformed to zero, and the higher absolute value of the negative forecast can be interpreted as a more certain forecast of zero. However,

in order to use and quantify these probabilities, more research into this is needed. There most likely are multiple reasons for the good performance of linear regression compared to the mathematically more accurate models, but one significant contributor is most likely the large number of zero sales days within the dataset. Especially Poisson regression is known to struggle in those cases, as that distribution is not equipped to handle large number of zero observations. In addition, the estimation methods for linear regression are more reliable, meaning that the chance of the algorithms converging to non-optimal parameters is smaller with linear regression than with other regression types.

Regarding the analyses with different amounts of campaign data in the training set, implemented with removing data from different stores, no significant differences were found between the hierarchical models. As expected, the non-pooled benchmark models struggled to forecast future campaign effects if there was only little data in the past, while the pooled benchmark model came very close to the performance of the hierarchical models in those cases. However, the hierarchical models did not seem to offer superior performance in comparison to all of the benchmark model in any case. The main advantage of hierarchical models in campaign forecasting with this dataset seems to be their ability to perform relatively well in all situations, independent of the amount of data available on specific product-stores, eliminating the need to specify rules and heuristics with which the model estimation level of the product-stores is specified.

Chapter 6

Conclusions

This study reviewed different ways to specify and estimate hierarchical campaign forecasting models. After the review, a selected group of them was tested and evaluated with a dataset obtained from a large European food retail chain, especially concentrating on evaluating the ability of the models to generate accurate forecasts in situations where there is not enough campaign data in the past on the lowest hierarchy levels. On a general level, the hierarchical models were able to forecast the campaign effects rather well on all situations, according to the aggregate level results presented in Chapter 5. No significant differences between the experimented models regarding the forecast accuracy were detected, but the estimation times and the reliability of the produced forecasts varied rather significantly. In addition, as no significant improvements in forecast accuracy compared to the best benchmark models were obtained, the main advantage of hierarchical models with this dataset seems to be obtained through their ability to produce good forecasts regardless of the amount of historical data in individual product-stores. This can be very useful in many cases as it eliminates the need to specify separate models for different product-stores and heuristics with which the estimation level is decided. The most promising model proved to be an Empirical Bayes implementation, with the fastest estimation times and often also the best forecast accuracies. Also, simple Generalized Linear Mixed Effect models were feasible, but they did not seem to yield any advantages over the Empirical Bayes models. With the EB models, fully Bayesian estimation of the parameters with Variational Inference did not improve any results compared to simpler Maximum a Posteriori estimation. Thus, when only point estimates of the parameters and forecasts are needed, MAP is a sufficient estimation method. However, if the full posterior distribution is of interest, then the Variational Inference methods seem very useful compared to sampling methods. The high outlier counts indicate that the Stan imple-

mentation of the ADVI algorithm had some problems in convergence, but in general, the results of the algorithm seemed to be in line with other methods. In addition, Poisson regression or log-linear regression did not prove to be better than linear regression in this context.

As the models were only experimented with one dataset and one division to training and validation sets, further research should be conducted on how the estimation times change with regards to the size of the dataset, and the forecast accuracy could be analyzed with different validation sets and forecast horizons. An interesting experiment would be to construct a rolling forecast, where the models are estimated, e.g., monthly with the training and validation set division changing to represent different forecast days, simulating the usage of the models in an operational setting where more data comes to use when time passes. In addition, this study focused only on a two-layered hierarchy. All of the models presented here should work very similarly when applied into hierarchical structures with more layers, but adding more layers will most likely increase the estimation times significantly with little gains in the forecast accuracies, as long as there are enough campaign days on the product-level. However, the magnitude of this effect should be studied more in-depth to better understand how modelling additional hierarchy levels affect the models. The quality and accuracy of the promotion forecasting are always highly dependent on the grouping of the campaign types, regardless of the model used. Here, dividing the promotions into four groups increased the forecast accuracy, but perhaps even more accurate forecasts could be obtained with different grouping. Thus, research into how to perform such grouping that divides the promotions into categories representing the sales behaviour in the best possible way could be conducted.

Bibliography

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50 (1-2):5–43, 2003.

Robert Bassett and Julio Deride. Maximum a posteriori estimators as a limit of bayes estimators. *Mathematical Programming*, 174(1-2):129–144, 2019.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

Douglas M Bates. lme4: Mixed-effects modeling with r, 2010.

John J Beauchamp and Jerry S Olson. Corrections for bias in regression estimates after logarithmic transformation. *Ecology*, 54(6):1403–1407, 1973.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Anthony S Bryk and Stephen W Raudenbush. Application of hierarchical linear models to assessing change. *Psychological bulletin*, 101(1):147, 1987.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Youyi Fong, Håvard Rue, and Jon Wakefield. Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412, 2010.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

Andrew Gelman, Alex Kiss, and Jeffrey Fagan. An analysis of the nypd's stop-and-frisk policy in the context of claims of racial bias. *Columbia Public Law Research Paper*, (05-95), 2006.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and other stories*. Cambridge University Press, 2020.

Arthur S Goldberger. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57(298): 369–375, 1962.

David A Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American statistical association*, 72(358):320–338, 1977.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

David A Hofmann, Mark A Griffin, and Mark B Gavin. The application of hierarchical linear modeling to organizational research. 2000.

Rob J Hyndman. Measuring forecast accuracy. *Business forecasting: Practical problems and solutions*, pages 177–183, 2014.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

Charles E McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437): 162–170, 1997.

Erik Meijer, H Goldstein, Jan Deleeuw, and Jan de Deleeuw. *Handbook of multilevel analysis*. Statistics. Springer, New York, NY, 2008. ISBN 0387731830.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Don M Miller. Reducing transformation bias in curve fitting. *The American Statistician*, 38(2):124–126, 1984.

Carl N Morris. Parametric empirical bayes inference: theory and applications. *Journal of the American statistical Association*, 78(381):47–55, 1983.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135 (3):370–384, 1972.

H Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

Simo Puntanen and George PH Styan. The equality of the ordinary least squares estimator and the best linear unbiased estimator. *The American Statistician*, 43(3):153–161, 1989.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822. PMLR, 2014.

Stephen W Raudenbush. Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, 13(2):85–116, 1988.

George K Robinson et al. That blup is a good thing: the estimation of random effects. *Statistical science*, 6(1):15–32, 1991.

Stan Development Team. RStan: the R interface to Stan, 2020. URL `http://mc-stan.org/`. R package version 2.21.2.

Rudolf Teekens and Johan Koerts. Some statistical implications of the log transformation of multiplicative models. *Econometrica: Journal of the Econometric Society*, pages 793–819, 1972.

Yihua Zhao, John Staudenmayer, Brent A Coull, and Matthew P Wand. General design bayesian generalized linear mixed models. *Statistical science*, pages 35–51, 2006.

Roman Zmyślony. A characterization of best linear unbiased estimators in the general linear model. In *Mathematical Statistics and Probability Theory*, pages 365–373. Springer, 1980.

Appendix A

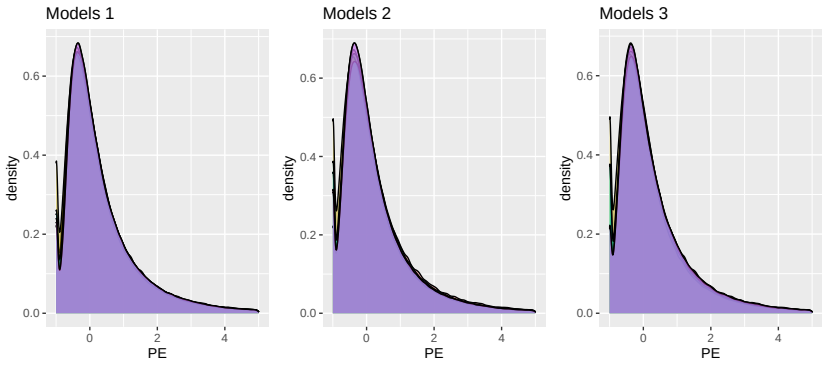
Additional results

A.1 Estimation time ranges

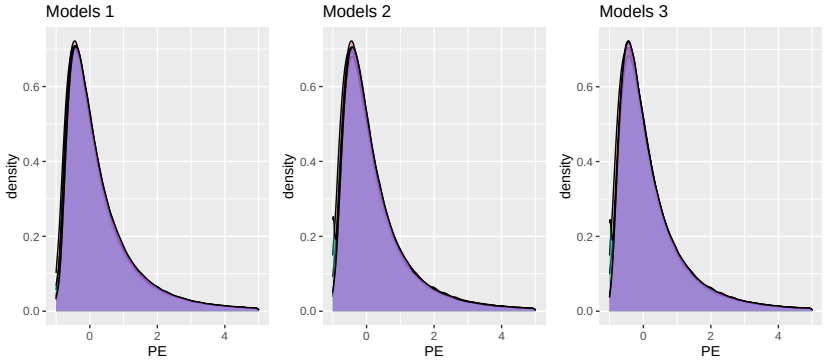
Table A.1: Mean, minimum and maximum estimation times per model in seconds.

Model	Linear	Log-linear	Poisson
GLM P	0.03 (0.02 - 0.18)	0.03 (0.02 - 0.40)	0.12 (0.06 - 7.09)
GLM S1	0.37 (0.27 - 0.82)	0.39 (0.30 - 0.94)	1.06 (0.69 - 4.70)
GLM S2	0.56 (0.37 - 2.42)	0.56 (0.40 - 2.13)	2.03 (1.00 - 11.04)
GLMM 1	2.38 (1.44 - 4.49)	2.63 (1.49 - 7.31)	NA
GLMM 2	7.64 (4.62 - 14.03)	7.72 (4.44 - 13.95)	NA
GLMM 3	27.17 (16 - 92)	26.70 (17 - 62)	NA
BGLMM 1 (MAP)	422.24 (190 - 1770)	546.95 (177 - 3001)	NA
BGLMM 2 (MAP)	582.17 (167 - 3026)	1881.57 (267 - 3080)	NA
BGLMM 3 (MAP)	1223.20 (255 - 3541)	2763.31 (996 - 3334)	NA
EB 1 (MAP)	2.22 (1.28 - 12.89)	1.73 (1.28 - 13.24)	1.95 (1.43 - 8.37)
EB 2 (MAP)	4.61 (2.63 - 16.36)	3.79 (2.52 - 13.34)	3.41 (1.82 - 15.60)
EB 3 (MAP)	6.50 (3.35 - 19.86)	6.04 (3.40 - 13.64)	6.14 (3.29 - 15.86)
EB 1 (VI)	77.19 (59 - 204)	114.82 (71 - 177)	154.64 (118 - 236)
EB 2 (VI)	85.54 (58 - 234)	128.65 (77 - 206)	174.31 (128 - 342)
EB 3 (VI)	117.23 (77 - 341)	263.89 (101 - 399)	315.46 (226 - 599)

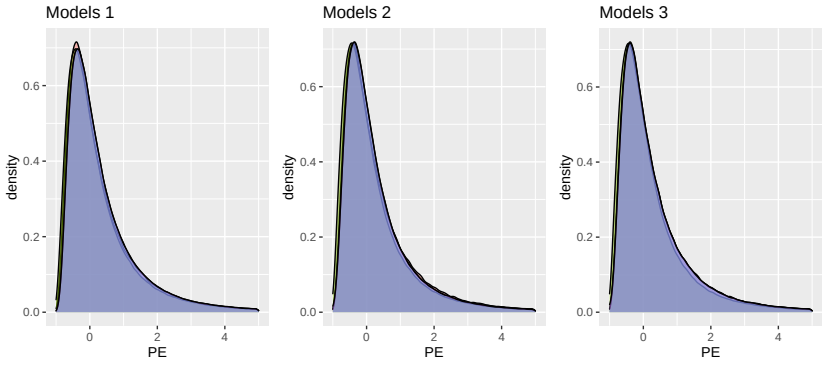
A.2 Density graphs of all models



(a) Linear regression



(b) Log-linear regression



(c) Poisson regression

Figure A.1: The approximated density of the daily percentage error distributions of different forecast models and regression types.

A.3 Store-specific results

Table A.2: Mean Absolute (daily) Deviation of each Store group for log-linear regression.

Stores	1-7	8	9	10	11
GLM P	3.9135	3.8247	4.4648	3.8124	3.8711
GLM S1	3.8214	3.3303	4.8177	3.4598	3.6492
GLMM 1	3.8162	3.2715	4.6798	3.287	3.7796
BGLMM 1 (MAP)	3.8654	3.3596	4.7552	3.3873	3.8035
EB 1 (MAP)	3.8117	3.2692	4.6991	3.5747	3.7863
EB 1 (VI)	3.8314	3.2624	4.7009	3.5935	3.8065
GLM S2	3.857	3.3976	5.0068	3.5846	3.6441
GLMM 2	3.8266	3.3325	4.7291	3.3067	3.7353
BGLMM 2 (MAP)	3.9142	3.6316	4.9796	3.6849	4.2615
EB 2 (MAP)	3.7781	3.3106	4.6601	3.5546	3.7923
EB 2 (VI)	3.7998	3.347	4.6705	3.6219	3.8112
GLMM 3	3.8342	3.352	4.7635	3.3209	3.7404
BGLMM 3 (MAP)	3.8835	3.4862	4.9649	3.5311	4.0363
EB 3 (MAP)	3.8395	3.7914	4.4394	3.7639	3.8699
EB 3 (VI)	3.8387	3.79	4.4341	3.764	3.8953

Table A.3: Mean Absolute (daily) Deviation of each store group for Poisson regression.

Stores	1-7	8	9	10	11
GLM P	3.5498	3.4073	4.1873	3.3477	3.4068
GLM S1	3.5058	2.9878	4.61	3.0292	3.4206
EB 1 (MAP)	3.3933	2.9954	4.442	3.1821	3.3155
EB 1 (VI)	3.3992	3.0011	4.457	3.1963	3.3271
GLM S2	3.462	2.9428	4.5147	3.0767	3.4206
EB 2 (MAP)	3.2754	2.9297	4.2944	3.0572	3.2232
EB 2 (VI)	3.2832	2.9461	4.3022	3.0563	3.2255
EB 3 (MAP)	3.4655	3.3781	4.1582	3.312	3.4224
EB 3 (VI)	3.483	3.391	4.1592	3.3268	3.4118

Table A.4: Product-store level WPE of each store group for log-linear regression.

Stores	1-7	8	9	10	11
GLM P	0.4301	1.0695	0.6644	0.4931	0.3936
GLM S1	0.2581	0.1952	0.9022	0.3101	-0.2872
GLMM 1	0.2612	0.2189	0.8407	0.2946	0.3968
BGLMM 1 (MAP)	0.2739	0.3715	0.8435	0.3132	0.3733
EB 1 (MAP)	0.3843	0.4683	0.88	0.4828	0.4135
EB 1 (VI)	0.4038	0.4899	0.9051	0.4823	0.4185
GLM S2	0.1869	0.1492	0.7825	0.2414	-0.291
GLMM 2	0.1966	0.156	0.6743	0.2263	0.3142
BGLMM 2 (MAP)	0.2165	0.2669	0.6553	0.3422	0.3874
EB 2 (MAP)	0.2895	0.4457	0.7116	0.4173	0.3423
EB 2 (VI)	0.2885	0.4718	0.7048	0.419	0.3426
GLMM 3	0.1959	0.1588	0.6916	0.2243	0.3137
BGLMM 3 (MAP)	0.214	0.2208	0.6678	0.3015	0.3152
EB 3 (MAP)	0.3931	1.0159	0.6601	0.4755	0.3973
EB 3 (VI)	0.3987	1.0428	0.6635	0.4797	0.3945

Table A.5: Product-store level WPE of each store group for Poisson regression.

Stores	1-7	8	9	10	11
GLM P	0.5368	1.1714	0.7283	0.5551	0.4471
GLM S1	0.376	0.4598	1.1041	0.3643	-0.2607
EB 1 (MAP)	0.549	0.7463	1.057	0.5485	0.4831
EB 1 (VI)	0.5375	0.7415	1.0429	0.5496	0.4835
GLM S2	0.2836	0.3733	0.8789	0.2591	-0.2607
EB 2 (MAP)	0.4027	0.6739	0.8246	0.4546	0.3809
EB 2 (VI)	0.4014	0.6706	0.8255	0.4559	0.3818
EB 3 (MAP)	0.5046	1.1329	0.7311	0.5399	0.4521
EB 3 (VI)	0.5093	1.134	0.7305	0.5428	0.4494