RWTH Aachen University

Institut für Statistik und Wirtschaftsmathematik

— Prof. Dr. M. Kateri —

Seminar zur Statistik und stochastischen Modellierung

Sommersemester 2022

Maarja Osi

Dynamic Generalized Linear Models and Bayesian Forecasting

23. September 2022

Contents

Introduction

Linear models are a very popular statistical method to predict random quantities based on a set of known or random variables. For random variables $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ and known vectors $\boldsymbol{x_i} \in \mathbb{R}^p, i = 1, ..., n$, where $p \in \mathbb{N}$ denotes the number of unknown parameters, a linear model is described by

$$\mu_i = \boldsymbol{x_i^T \theta},\ i = 1, ..., n,$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ is an unknown parameter vector.

Over time, models such as logistic and Poisson regression were developed to model more complex non-linear relationships. For example, in the logistic regression model, the relationship

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = \boldsymbol{x_i^T \theta},\ i = 1, ..., n,$$

for $Y_i \sim \mathcal{B}(k, \mu_i), k \in \mathbb{N}$ is studied. In [Nel72], such models were unified under common framework of *generalized linear models*, where the relationship between μ_i and x_i can be described by any function and Y_i are random variables with a distribution from the *exponential family* of distributions introduced in Chapter 1. Since then, generalized linear models have found use in a wide range of applications and have been implemented in most statistical software (see [Lin97, Preface]).

Still, these models are not well suited for modelling time dependent data where the magnitude of the parameters of such relationships might change over time (see [Wes85a, Ch 1.]). For example, in Chapter 5 a data example of advertising on consumer awareness is explored. In this example, the nature of the advertising is changed over the course of time leading to possible change in the effect that advertising has on awareness. In this seminar thesis, we follow [Wes85a], where dynamic extensions of such models are discussed. In Chapters 2 and 3, *dynamic linear* and *dynamic generalized linear models* are defined. In these models, the parameters are calculated sequentially at each time step. Moreover, the use of previously obtained information about parameters in further inference is determined explicitly through linear transformations of such information. In the advertising example from Chapter 5, for example, previous information is mostly dismissed when the nature of advertising is changed. The parameters are updated according to methods of Bayesian statistics. Additionally, means to make predictions about future data are provided. Properties of the exponential family are used to obtain general results applicable to all distributions of this class.

The exponential family of distribution is parameterized over a natural parameter η and scale parameter ϕ. In Chapter 2 and 3, the parameter ϕ is assumed to be constant. In Chapter 4, extensions to the models are presented for which the scale parameter is allowed to vary. This provides the means to check for the goodness of fit for the models. Additionally, the chapter discusses ways to limit the influence of outliers on the results.

Finally, in Chapter 5 applications of the dynamic generalized linear models on real data sets are presented. In Example 5.1, the results obtained from the dynamic and static models on time dependent data are compared to illustrate the superiority of the dynamic models in such cases. In Examples 5.2 and 5.3, it is tested whether the dynamic models are also appropriate to use on time independent data.

1 Preliminaries

Dynamic (generalized) linear models are constructed in the framework of Bayesian statistics. In the following pages, we introduce the most important concepts from Bayesian statistics for our purposes. For this, we follow the first chapter of [Hof09] and the sources given at the respective results. Vectors are denoted in bold.

In Bayesian statistics, probabilities are used to express beliefs about the characteristics of unknown quantities, as opposed to the classical interpretation of probability based on the relative frequency of events. Most importantly, the unknown parameters of a sampling distribution under interest are also treated as random variables. The process of Bayesian inference involves modelling prior beliefs about these parameters and updating them according to the observed data. Such kind of prior beliefs can, for example, be based on expert information or previously observed data.

To formalize the above description, let Y denote a random variable taking values in some set $\mathcal{Y}$. The density of the distribution of Y is denoted by $p(y|\theta)$ for $y \in \mathcal{Y}$. The parameter θ is treated as random variable with possible values in parameter space Θ. Concerning notation, for a given value $\tilde{\theta}$ of θ, the density $p(y|\tilde{\theta})$ denotes the conditional density of Y given $\theta = \tilde{\theta}$. We will use the notation $p(y|\theta)$ in this sense without fixing a concrete value for θ. The prior beliefs about θ are expressed through a *prior distribution* with the density $p(\theta)$.

After data $y \in \mathcal{Y}$ is observed, the beliefs about θ are updated by deriving a *posterior distribution* with the density $p(\theta|y)$ with the help of *Bayes' rule* for densities:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}, \ \theta \in \Theta. \tag{1}$$

The expression in the denominator of (1) is a constant i.e. the left-hand side and the numerator of the right-hand side are proportional to each other. To indicate that $p(\theta|y)$ and $p(y|\theta)p(\theta)$ (or any other two functions) are proportional to each other, we write $p(\theta|y) \propto p(y|\theta)p(\theta)$.

In the context of dynamic (generalized) linear models, we want to make predictions about future data based on currently available information. For this, we introduce the concept of *prior predictive distribution.*

1.1 Definition. [Gel04, p. 8] For a random variable Y with a probability density function $p(y \,|\, \theta)$ and a given prior distribution $p(\theta)$, the *prior predictive distribution* of Y given $p(\theta)$ is defined as

$$p(y) := \int_\Theta p(y, \tilde{\theta}) \; d\tilde{\theta} = \int_\Theta p(y|\tilde{\theta})p(\tilde{\theta}) \; d\tilde{\theta}, \ y \in \mathcal{Y}.$$

In the next example, the posterior distribution is calculated if both the sampling distribution and the prior distribution of the mean of the sampling distribution are normal.

1.2 Example. [Hof09, p. 69-71] Let $\boldsymbol{Y} = (Y_1 \,, \dots, Y_n)$ be a random sample where Y_i are identically distributed random variables each following a $\mathcal{N}(\mu, \sigma^2)$ distribution, with a known and constant $\sigma^2 > 0$. The prior distribution of μ is also assumed to be a normal distribution $\mathcal{N}(\mu_0, \sigma_0^2)$ with known $\mu_0 \in \mathbb{R}$ and $\sigma_0^2 > 0$. Let $\boldsymbol{y} = (y_1 , \dots , y_n)$ be an observed sample with sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Then, applying (1), we obtain a posterior distribution, which is a $\mathcal{N}(\mu_n, \sigma_n^2)$ distribution with

$$\sigma_n^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

$$\mu_n = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\sigma_n^2}}$$

(the exact derivation can be found in the source). We denote this result as

$$(\mu|\boldsymbol{Y} = \boldsymbol{y}) \sim \mathcal{N}(\mu_n, \sigma_n^2).$$

This kind of closure property, where the obtained posterior distributions are of the same distribution family as the prior, is called conjugacy.

1.3 Definition. ([Gel04, p. 41]) A class of prior distributions $\mathcal{P}$ for θ is called *conjugate* for a class of sampling distributions $\mathcal{F}$ if for all $f(\cdot|\theta) \in \mathcal{F}$ and all $p(\theta) \in \mathcal{P}$ the corresponding posterior distribution belongs into $\mathcal{P}$.

Conjugate priors are mostly of algebraic convenience, reducing the complexity of the calculations. We will now address the exponential family of distributions which have prior distributions of general form. In generalized linear models, the data is assumed to follow a distribution from the exponential family as we will see later.

1.4 Definition. ([Wes85a]) A random variable Y, with values in $\mathcal{Y}$, has a distribution belonging into the *exponential family* if its density has the form

$$p(y \mid \eta, \phi) = \exp\left[\phi\{y\eta - a(\eta)\}\right] b(y, \phi), \ y \in \mathcal{Y}. \tag{2}$$

for parameters $\eta \in \Theta \subset \mathbb{R}, \phi > 0$ and some functions $a : \Theta \to \mathbb{R}$ two times differentiable, $b : \mathcal{Y} \times \mathbb{R}_{>0} \to \mathbb{R} \setminus \{0\}$. We write $Y \sim \text{EXPF}(\eta, \phi, a, b)$.

1.5 Remark. ([Wes85a]) The parameter η in Definition 1.4 is called the *natural parameter* and it determines the expectation of Y through the relation

$$\mathbb{E}[Y \mid \eta, \phi] = a'(\eta). \tag{3}$$

The parameter ϕ is called the *scale parameter* and it determines the variance of Y through the relation

$$\mathbb{V}[Y \mid \eta, \phi] = a''(\eta)/\phi$$

In many sources concerning GLMs, $1/\phi$ is used instead of ϕ in the parametrization (2). In the Bayesian setting the parametrization given in Definition 1.4 is preferred, since in many cases conjugate priors can be found for ϕ but not for $1/\phi$.

1.6 Lemma. ([Wes85a]) Let Y be a random variable with a probability distribution from the exponential family. For a fixed ϕ, the natural parameter η has a conjugate prior $\text{CP}(\alpha, \beta)$ with density

$$p(\eta) = c(\alpha, \beta) \exp[\alpha\eta - \beta a(\eta)] \tag{4}$$

for some function $c(\cdot)$ and parameters α, β. The corresponding posterior distribution is $CP(\alpha + \phi \cdot y, \beta + \phi)$.

The prior predictive distribution for Y is given by

$$p(y) = \frac{c(\alpha, \beta)}{c(\alpha + \phi y, \beta, +\phi)} \cdot b(y, \phi). \tag{5}$$

Proof. Denote by Θ the parameter space of η. By Definition 1.1 of $p(y)$, we obtain (5)

$$\begin{aligned}
p(y) &= \int_{\tilde{\eta}\in\Theta} p(y \mid \tilde{\eta}, \phi) \cdot p(\tilde{\eta}) d\tilde{\eta} \\
&= \int_{\tilde{\eta}\in\Theta} \exp\left[\phi\{y\tilde{\eta} - a(\tilde{\eta})\}\right] \cdot c(\alpha, \beta) \exp\left[\alpha\tilde{\eta} - \beta a(\tilde{\eta})\right] b(y, \phi) d\tilde{\eta} \\
&= c(\alpha, \beta) b(y, \phi) \cdot \int_{\tilde{\eta}\in\Theta} \exp\left[(\alpha + \phi y)\tilde{\eta} - (\beta + \phi) a(\tilde{\eta})\right] d\tilde{\eta} \\
&= \frac{c(\alpha, \beta)}{c(\alpha + \phi y, \beta + \phi)} \cdot b(y, \phi) \cdot \int_{\tilde{\eta}\in\Theta} c(\alpha + \phi y, \beta + \phi) \exp\left[(\alpha + \phi y)\tilde{\eta} - (\beta + \phi) \cdot a(\tilde{\eta})\right] d\tilde{\eta} \\
&= \frac{c(\alpha, \beta)}{c(\alpha + \phi y, \beta + \phi)} \cdot b(y, \phi).
\end{aligned}$$

In the last step, we used that the term under the integral is the density of $\mathrm{CP}(\alpha + \phi \cdot y, \beta + \phi)$. Now, we can confirm the parameters and the conjugate form of the posterior distribution by applying the Bayes' rule to calculate the posterior density

$$\begin{aligned}
p(\eta \,|\, y) &= \exp\left[\phi\{y\eta - a(\eta)\}\right] b(y, \theta) \cdot c(\alpha, \beta) \exp[\alpha\eta - \beta a(\eta)] \cdot \frac{c(\alpha + \phi y, \beta + \phi)}{c(\alpha, \beta) b(y, \phi)} \\
&= c(\alpha + \phi y, \beta + \phi) \exp\left[(\alpha + \phi y)\eta - (\beta + \phi) a(\eta)\right],
\end{aligned}$$

which is the density of $\mathrm{CP}(\alpha + \phi y, \beta + \phi)$. □

1.7 Example. Let a single random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$ be given. We can express the density of Y in the exponential family form by setting (see [McN89, p. 30])

$$\begin{aligned}
\eta &:= \mu, \quad \phi := \frac{1}{\sigma^2} \\
a(\eta) &:= \frac{\eta^2}{2}, \quad b(y, \phi) := \exp\left[-\frac{1}{2}\left(y^2\phi + \ln\left(\frac{2\pi}{\phi}\right)\right)\right].
\end{aligned}$$

As noted in Remark 1.5, we have chosen $\phi = \frac{1}{\sigma^2}$ instead of $\phi = \sigma^2$ (as is done in [McN89]). An appropriate conjugate prior for η is a normal prior $\mathcal{N}(\mu_0, \sigma_0^2)$, as we have seen in Example 1.2. We can write this in the form given in (4) by setting

$$\begin{aligned}
\alpha &= \frac{\mu_0}{\sigma_0^2}, \quad \beta = \frac{1}{\sigma_0^2} \\
c(\alpha, \beta) &= \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{1}{2}\frac{\alpha}{\beta}\right).
\end{aligned}$$

Then, according to Lemma 1.6, we obtain $(\eta \,|\, Y = y) \sim \mathrm{CP}[\alpha + \phi \cdot y, \beta + \phi]$. The distribution $\mathrm{CP}[\alpha + \phi \cdot y, \beta + \phi]$ is a normal distribution $\mathcal{N}(\mu_1, \sigma_1^2)$ with

$$\begin{aligned}\sigma_1^2 &= \frac{1}{\beta + \phi} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}} \\ \mu_1 &= \frac{\alpha + \phi \cdot y}{\beta + \phi} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{y}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}.\end{aligned}$$

This corresponds to the result in Example 1.2 since $n = 1$.

2 Dynamic Linear Model

For random variables $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ and known vectors $\boldsymbol{x_i} \in \mathbb{R}^p$, $i = 1, ..., n$, where $p \in \mathbb{N}$ denotes the number of unknown parameters, a linear model is described by

$$\mu_i = \boldsymbol{x_i^T \theta},\ i = 1, ..., n, \tag{6}$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ is an unknown parameter vector. For convenience, the equations (6) are often summarized into one equation

$$\boldsymbol{\mu} = X^T \boldsymbol{\theta}, \tag{7}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T \in \mathbb{R}^n$ and $X = (\boldsymbol{x_1}, .., \boldsymbol{x_n}) \in \mathbb{R}^{p \times n}$.

For linear models the means μ_i are not directly the parameters of interest, but rather the parameter vector $\boldsymbol{\theta}$. Linear models can be generalized to the Bayesian setting by treating $\boldsymbol{\theta}$ as a random variable with a prior distribution. This is illustrated in the following example.

2.1 Example. For $i = 1, ..., n$ let $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ be random variable with a known variance $\sigma^2 > 0$, and $\boldsymbol{x_i} \in \mathbb{R}^p$ be a known vector. Let $\boldsymbol{\theta}$ be a random vector of parameters with a prior distribution $\mathcal{N}_p(\boldsymbol{\theta_0}, \Sigma_0)$, where the mean vector $\boldsymbol{\theta_0} \in \mathbb{R}^p$ and the regular covariance matrix $\Sigma_0 \in \mathbb{R}^{p \times p}$ are known.

Using the notation from beforehand, let's assume a linear model relating $\boldsymbol{Y}$ to X through $\boldsymbol{\theta}$ i.e.

$$\boldsymbol{\mu} = X^T \boldsymbol{\theta}.$$

For an observed value $\boldsymbol{y}$ of $\boldsymbol{Y}$, the posterior density for $\boldsymbol{\theta}$ is a multivariate normal distribution $\mathcal{N}_p(\boldsymbol{m}, \Sigma)$, where (see [Hof09, p. 154 - 155])

$$\boldsymbol{m} = (\Sigma_0^{-1} + \frac{1}{\sigma^2} X X^T)^{-1} \cdot (\Sigma_0^{-1} \boldsymbol{\theta_0} + \frac{1}{\sigma^2} X \boldsymbol{y}) \tag{8}$$

$$\Sigma = (\Sigma_0^{-1} + \frac{1}{\sigma^2} X X^T)^{-1}. \tag{9}$$

In order to model time dependent data more accurately, the dynamic linear model doesn't assume that $\boldsymbol{\theta}$ is independent of time. Instead, in equation (6) the parameter $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\theta_i}$ for each i (or in later notation by $\boldsymbol{\theta_t}$, where $t \in \mathbb{N}$ denotes the time). Thus, the equation (7) can't be used anymore.

As can be seen in equations (8) and (9), each entry of posterior mean and covariance of $\boldsymbol{\theta}$ is influenced by each of the observed values y_i and the vectors $\boldsymbol{x_i}$. In the dynamic linear model, dependency on previously observed data is modelled more explicitly. For time t, the prior distribution for $\boldsymbol{\theta_t}$ is derived by through transforming the posterior distribution of $\boldsymbol{\theta_{t-1}}$ through $G_t\boldsymbol{\theta_{t-1}} + \boldsymbol{w_t}$. The matrix G_t is called the *transition matrix*. It can be chosen by the modeller to represent how information available at time $t-1$ is used to construct a prior distribution for θ_t. The random variable $\boldsymbol{w_t}$ is responsible for modelling increasing uncertainty in time. We will index the dynamic linear model over $\mathbb{N}$ but any discrete set of values might be used. In this Chapter as well as in Chapter 4, the dispersion parameter ϕ is assumed to be known and fixed over time. Models where ϕ is a random variable are discussed in Chapter 4.

2.2 Definition. ([Wes85a]) For $t \in \mathbb{N}$, let

- $Y_t \sim \mathcal{N}(\mu_t, \phi^{-1})$, for a known and constant $\phi > 0$,
- $\boldsymbol{x_t} \in \mathbb{R}^p$ be a known vector,
- $\boldsymbol{\theta_t}$ be a random variable taking values in $\mathbb{R}^p$,
- $G_t \in \mathbb{R}^{p\times p}$ be the transition matrix,
- $W_t \in \mathbb{R}^{p\times p}$ be a symmetric, positive-definite matrix.

Denote by $I_t := \{Y_t, \boldsymbol{x_t}, G_{t+1}, W_{t+1}\}$ the information available at time t but not $t-1$ and set $D_0 := \{\boldsymbol{m_0}, C_0, G_1, W_1\}$, $D_t := D_{t-1} \cup I_t$, which denote the total information available at time 0 and t respectively. Let $\boldsymbol{w_t} \sim \mathcal{N}_p(\boldsymbol{0_p}, W_t)$ be independent of $(\boldsymbol{\theta_{t-1}} \,|\, D_{t-1})$, and $\boldsymbol{\theta_0} \sim \mathcal{N}_p(\boldsymbol{m_0}, C_0)$ for $C_0 \in \mathbb{R}^{p\times p}$ and $\boldsymbol{m_0} \in \mathbb{R}^p$.

The *dynamic linear model* (DLM) $((G_t)_{t\in\mathbb{N}}, (W_t)_{t\in\mathbb{N}}, \boldsymbol{m_0}, C_0)$ for the process $(Y_t \,|\, \mu_t, \phi)_{t\in\mathbb{N}}$ at time $t \in \mathbb{N}$ is specified by the linear model

$$\mu_t = \boldsymbol{x_t^T}\boldsymbol{\theta_t}, \tag{10}$$

and the *state evolution equation*

$$(\boldsymbol{\theta_t} \,|\, D_{t-1}) = (G_t\boldsymbol{\theta_{t-1}} + \boldsymbol{w_t} \,|\, D_{t-1}). \tag{11}$$

We proceed with the analysis of the prior and posterior distributions for $\boldsymbol{\theta_t}$.

2.3 Lemma. ([Ata08, p. 21]) Denote by U_n the n-dimensional identity matrix for $n \in \mathbb{N}$. For matrices $A \in \mathbb{R}^{n\times p}$, $B \in \mathbb{R}^{p\times n}$, such that $U_n + AB$ and $U_p + BA$ are regular, there holds

$$(U_n + AB)^{-1} = U_n - A(U_p + BA)^{-1}B. \tag{12}$$

2.4 Proposition. ([Wes85a]) Let a dynamic linear model $\{(G_t)_t, (W_t)_t, \boldsymbol{m_0}, C_0\}$ for a normal process $(Y_t \,|\, \mu_t, \phi)_{t\in\mathbb{N}}$ be given. Then for all $t \in \mathbb{N}$ the prior and posterior for $\boldsymbol{\theta_t}$ are given by

$$(\boldsymbol{\theta_t} \,|\, D_{t-1}) \sim \mathcal{N}_p(\boldsymbol{a_t}, R_t)$$
$$(\boldsymbol{\theta_t} \,|\, D_t) \sim \mathcal{N}_p(\boldsymbol{m_t}, C_t),$$

where

$$\begin{aligned}
\boldsymbol{a_t} &= G_t \boldsymbol{m_{t-1}}, \\
R_t &= G_t C_{t-1} G_t^T + W_t, \\
\boldsymbol{m_t} &= \boldsymbol{a_t} + \frac{\phi(y_t - f_t)}{\phi q_t + 1} R_t \boldsymbol{x_t}, \\
C_t &= R_t - \frac{\phi}{\phi q_t + 1} R_t \boldsymbol{x_t} \boldsymbol{x_t^T} R_t
\end{aligned}$$

for $q_t = \boldsymbol{x_t^T} R_t \boldsymbol{x_t},\ f_t = \boldsymbol{x_t^T} \boldsymbol{a_t}$.

Proof. First, notice that $R_t^T = R_t$, since R_t is a covariance matrix and thus symmetric. By induction, assuming that $(\boldsymbol{\theta_{t-1}} \mid D_{t-1}) \sim \mathcal{N}(\boldsymbol{m_{t-1}}, C_{t-1})$, by the properties of multivariate normal distribution the state evolution equation (11) leads to the following prior distribution

$$(\boldsymbol{\theta_t} \mid D_{t-1}) = (G_t \boldsymbol{\theta_{t-1}} + \boldsymbol{w_t} \mid D_{t-1}) \sim \mathcal{N}_p(G_t \boldsymbol{m_{t-1}}, G_t C_{t-1} G_t^T + W_t),$$

where the covariance matrix and the mean are equal to $\boldsymbol{a_t}$ and R_t respectively. To obtain the posterior covariance matrix and mean, we proceed as in example 2.1 and obtain

$$\begin{aligned}
C_t &\overset{(9)}{=} (R_t^{-1} + \phi \boldsymbol{x_t} \boldsymbol{x_t^T})^{-1} = (R_t^{-1} + \phi \boldsymbol{x_t} \boldsymbol{x_t^T})^{-1} R_t^{-1} R_t \\
&= (R_t(R_t^{-1} + \phi \boldsymbol{x_t} \boldsymbol{x_t^T}))^{-1} R_t = (U_p + \phi R_t \boldsymbol{x_t} \boldsymbol{x_t^T})^{-1} R_t \\
&\overset{(12)}{=} (U_p - \phi R_t \boldsymbol{x_t} (1 + \phi \boldsymbol{x_t^T} R_t \boldsymbol{x_t})^{-1} \boldsymbol{x_t^T}) R_t \\
&= R_t - \frac{\phi}{(1 + \boldsymbol{x_t^T} R_t \boldsymbol{x_t} \phi)} R_t \boldsymbol{x_t} \boldsymbol{x_t^T} R_t \\
&= R_t - \frac{\phi}{(\phi q_t + 1)} R_t \boldsymbol{x_t} \boldsymbol{x_t^T} R_t
\end{aligned}$$

as well as

$$\begin{aligned}
\boldsymbol{m_t} &\overset{(8)}{=} C_t (R_t^{-1} \boldsymbol{a_t} + \phi y_t \boldsymbol{x_t}) \\
&= (R_t - \frac{\phi}{(\phi q_t + 1)} R_t \boldsymbol{x_t} \boldsymbol{x_t^T} R_t)(R_t^{-1} \boldsymbol{a_t} + \phi y_t \boldsymbol{x_t}) \\
&= \boldsymbol{a_t} + \phi y_t R_t \boldsymbol{x_t} - \frac{\phi}{(\phi q_t + 1)} R_t \boldsymbol{x_t} \boldsymbol{x_t^T} R_t (R_t^{-1} \boldsymbol{a_t} + \phi y_t \boldsymbol{x_t}) \\
&= \boldsymbol{a_t} + \frac{\phi}{(\phi q_t + 1)} R_t \boldsymbol{x_t} (y_t \cdot (\phi \frac{(\phi q_t + 1)}{\phi} - \phi \boldsymbol{x_t^T} R_t \boldsymbol{x_t}) - \boldsymbol{x_t^T} \boldsymbol{a_t}) \\
&= \boldsymbol{a_t} + \frac{\phi}{(\phi q_t + 1)} (y_t \cdot (\phi \cdot \boldsymbol{x_t^T} R_t \boldsymbol{x_t} + 1 - \phi \boldsymbol{x_t^T} R_t \boldsymbol{x_t}) - \boldsymbol{x_t^T} \boldsymbol{a_t}) R_t \boldsymbol{x_t} \\
&= \boldsymbol{a_t} + \frac{\phi}{(\phi q_t + 1)} (y_t - \boldsymbol{x_t^T} \boldsymbol{a_t}) R_t \boldsymbol{x_t}.
\end{aligned}$$

□

2.5 Remark. In (11) the random variable $\boldsymbol{w_t}$ is used to model the increasing uncertainty over time. An alternative, more simple approach was introduced in [Ame85], where instead the same effect is controlled through multiplication with a diagonal matrix

$B_t \in \mathbb{R}^{p\times p}$ with positive diagonal elements $\frac{1}{\beta_{it}} \geq 1, i = 1, ..., p$. In this case, the term $G_t\boldsymbol{\theta_t} + \boldsymbol{w_t}$ in the state evolution equations (11) is replaced by $B_tG_t\boldsymbol{\theta_{t-1}}$, so that covariance matrix for the prior distribution of $\boldsymbol{\theta_t}$ becomes $R_t = B_tG_tC_{t-1}G_t^TB_t$.

Finally, to make the transition to the dynamic generalized linear models, we present an alternative proof to Proposition 2.4. The posterior distribution of the natural parameter η_t of the exponential family is derived using the conjugate prior given Lemma 1.6. Then, the relation between η_t and $\boldsymbol{\theta_t}$, given by the linear model, is used to derive the posterior mean and variance for $\boldsymbol{\theta_t}$. Beforehand, we review the definitions of conditional expectation and conditional variance.

2.6 Definition. ([Bli14, p. 384]) Let X and Y be two random variables. If Y is continuous and the conditional probability density function $f_{Y\,|\,X}$ exists, then we define the *conditional expected value of Y given $X = x$* as

$$\mathbb{E}[Y\,|\,X = x] = \int_{-\infty}^{\infty} y \cdot f_{Y\,|\,X}(y\,|\,x)\,dy.$$

If Y is discrete, the integral is replaced by a sum and the density $f_{Y\,|\,X}$ by the probability mass function of Y given $X = x$. For a random vector Y, the definition can be expanded by calculating the mean of the random vector Y with respect to the density $f_{Y\,|\,X}$.

2.7 Definition. ([Bli14, p. 392, 400])
For random variables X and Y, let $g : x \mapsto \mathbb{E}[Y\,|\,X = x]$. Then the *conditional expectation of Y given X* is defined as

$$E[Y\,|\,X] := g(X).$$

The *conditional variance of Y given X* is defined as

$$V(Y\,|\,X) := E[(Y - E[Y\,|\,X])^2\,|\,X] = E[Y^2\,|\,X] - (E[Y|X])^2.$$

Notice that the conditional expectation and the conditional variance are random variables and the change in notation from $\mathbb{E}(\cdot)$ to $E(\cdot)$ (analogously from $\mathbb{V}(\cdot)$ to $V(\cdot)$).

Proof. (for Proposition 2.4, [Wes85a]) The natural parameter η_t of a normal distribution is equal to the mean as described in example 1.7. Thus, according to (10) the parameters η_t and $\boldsymbol{\theta_t}$ are related through the linear model $\eta_t = \boldsymbol{x_t^T\theta_t}$. If $(\boldsymbol{\theta_t}|D_{t-1}) \sim \mathcal{N}(\boldsymbol{a_t}, R_t)$, then by properties of the normal distribution $(\eta_t\,|\,D_{t-1}) \sim \mathcal{N}(f_t, q_t)$, where $f_t = \boldsymbol{x_t^T a_t}$ and $q_t = \boldsymbol{x_t^T}R_t\boldsymbol{x_t}$. Notice that the density for Y_t given $\boldsymbol{\theta_t}$ depends on $\boldsymbol{\theta_t}$ only through $\eta_t = \boldsymbol{x_t^T\theta_t}$ i.e.

$$p(y_t\,|\,\eta_t, \boldsymbol{\theta_t}) = p(y_t\,|\,\eta_t). \tag{13}$$

Denote by $p(y_t\,|\,D_{t-1})$ the prior predictive distribution of Y_t given the prior $p(\eta_t\,|\,D_{t-1})$. The joint posterior density of θ_t and η_t can be expressed as follows

$$p(\boldsymbol{\theta_t}, \eta_t\,|\,D_t) = p(\eta_t\,|\,D_t)p(\boldsymbol{\theta_t}\,|\,\eta_t, D_{t-1}). \tag{14}$$

This can be shown with the help of Bayes' rule

$$\begin{aligned} p(\boldsymbol{\theta_t}, \eta_t \,|\, D_t) &= \frac{p(y_t \,|\, \eta_t, \boldsymbol{\theta_t}) p(\boldsymbol{\theta_t}, \eta_t \,|\, D_{t-1})}{p(y_t|D_{t-1})} \overset{(13)}{=} \frac{p(y_t \,|\, \eta_t) p(\boldsymbol{\theta_t}, \eta_t \,|\, D_{t-1})}{p(y_t|D_{t-1})} \\ &= \frac{p(y_t \,|\, \eta_t) p(\eta_t \,|\, D_{t-1}) p(\boldsymbol{\theta_t} \,|\, \eta_t, D_{t-1})}{p(y_t|D_{t-1})} \\ &= \frac{p(y_t \,|\, D_{t-1}) p(\eta_t|D_t) p(\boldsymbol{\theta_t} \,|\, \eta_t, D_{t-1})}{p(y_t|D_{t-1})} \\ &= p(\eta_t \,|\, D_t) p(\boldsymbol{\theta_t} \,|\, \eta_t, D_{t-1}). \end{aligned}$$

The term $p(\eta_t \,|\, D_t)$ is the posterior density for the natural parameter η_t for the conjugate prior $\mathcal{N}(f_t, q_t)$. Analogously to example 1.2 we have $(\eta_t \,|\, D_t) \sim \mathcal{N}(g_t, p_t)$ where

$$g_t = \frac{\frac{f_t}{q_t} + y_t \phi}{\frac{1}{q_t} + \phi} = f_t + \frac{\phi q_t (y_t - f_t)}{(\phi q_t + 1)} \tag{15}$$

$$p_t = \frac{1}{\frac{1}{q_t} + \phi} = q_t - \frac{q_t^2 \phi}{(\phi q_t + 1)}. \tag{16}$$

The term $p(\boldsymbol{\theta_t} \,|\, \eta_t, D_{t-1})$ can be calculated from the joint prior distribution of $\boldsymbol{\theta_t}$ and η_t, which is a

$$\mathcal{N}_{p+1}\left(\begin{pmatrix} f_t \\ \boldsymbol{a_t} \end{pmatrix}, \begin{pmatrix} q_t & (R_t \boldsymbol{x_t})^T \\ R_t \boldsymbol{x_t} & R_t \end{pmatrix} \right) \tag{17}$$

distribution, since $\mathrm{Cov}(\boldsymbol{\theta_t}, \eta_t) = \mathrm{Cov}(\boldsymbol{\theta_t}, \boldsymbol{x_t^T} \boldsymbol{\theta_t}) = \mathrm{Cov}(\boldsymbol{\theta_t}, \boldsymbol{\theta_t}) \boldsymbol{x_t} = R_t \boldsymbol{x_t}$. Then, by properties of the multivariate normal distribution $(\boldsymbol{\theta_t} \,|\, \eta_t, D_{t-1}) \sim \mathcal{N}(\boldsymbol{\hat{a}_t}, \hat{R}_t)$, where

$$\boldsymbol{\hat{a}_t} = \boldsymbol{a_t} + \frac{(\eta_t - f_t)}{q_t} R_t \boldsymbol{x_t} \tag{18}$$

$$\hat{R}_t = R_t - \frac{1}{q_t} R_t \boldsymbol{x_t} \boldsymbol{x_t^T} R_t. \tag{19}$$

Finally, the mean and covariance of $(\boldsymbol{\theta}_t \,|\, D_t)$ can be derived by first noting that for each component of θ_{ti} of θ_t, the marginal distribution also fulfills $p(\theta_{ti} \,|\, \eta_t, D_{t-1}) p(\eta_t \,|\, D_t) = p(\theta_{ti}, \eta_t \,|\, D_t)$ thus we have

$$\begin{aligned} \mathbb{E}\left[E[\theta_{ti} \,|\, \eta_t, D_{t-1}] \,|\, D_t\right] &= \int_{\mathbb{R}} \mathbb{E}[\theta_{ti} \,|\, \eta_t = \hat{\eta}] \cdot p(\hat{\eta} \,|\, D_t) \, d\hat{\eta} \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} \hat{\theta} \cdot p(\hat{\theta} \,|\, \eta_t = \hat{\eta}, D_{t-1}) p(\hat{\eta}|D_t) \, d\hat{\theta} \, d\hat{\eta} \\ &\overset{(14)}{=} \int_{\mathbb{R}} \int_{\mathbb{R}} \hat{\theta} \cdot p(\hat{\theta}, \hat{\eta} \,|\, D_t) \, d\hat{\theta} \, d\hat{\eta} \\ &= \int_{\mathbb{R}} \hat{\theta} \cdot p(\hat{\theta} \,|\, D_t) \, d\hat{\theta} \\ &= \mathbb{E}[\theta_{ti} \,|\, D_t]. \end{aligned}$$

Then also $\mathbb{E}\left[E\left[\boldsymbol{\theta_t} \,|\, \eta_t, D_{t-1}\right] \,|\, D_t\right] = \mathbb{E}[\boldsymbol{\theta_t} \,|\, D_t]$ since the entries of these conditional expectations are precisely the respective conditional expectations of the entries of $\boldsymbol{\theta_t}$. Then, using the previously obtained results about the distributions of $(\boldsymbol{\theta_t} \mid \eta_t, D_t)$ and $(\eta_t \mid D_t)$ we obtain

$$\begin{aligned}\mathbb{E}[\boldsymbol{\theta_t} \mid D_t] &= \mathbb{E}\left[E[\boldsymbol{\theta_t} \mid \eta_t, D_{t-1}] \mid D_t\right] = \mathbb{E}[\boldsymbol{\hat{a}_t} \mid D_t] \\ &\overset{(18)}{=} \boldsymbol{a_t} + \frac{(g_t - f_t)}{q_t} R_t \boldsymbol{x_t} \\ &\overset{(15)}{=} \boldsymbol{a_t} + \frac{\phi(y_t - f_t)}{(\phi q_t + 1)} R_t \boldsymbol{x_t}.\end{aligned}$$

Similarly, there holds

$$\begin{aligned}\mathbb{V}[\boldsymbol{\theta_t} \mid D_t] &= \mathbb{V}[E[\boldsymbol{\theta_t} \mid \eta_t, D_{t-1}] \mid D_t] + \mathbb{E}[V[\boldsymbol{\theta_t} \mid \eta_t, D_{t-1}] \mid D_t] \\ &= \mathbb{V}[\boldsymbol{\hat{a}_t} \mid D_t] + \mathbb{E}[\hat{R}_t \mid D_t] \\ &\overset{(17),(19)}{=} R_t - \frac{(1 - \frac{p_t}{q_t})}{q_t} R_t \boldsymbol{x_t} \boldsymbol{x_t^T} R_t \\ &\overset{(16)}{=} R_t - \frac{\phi}{(\phi q_t + 1)} R_t \boldsymbol{x_t} \boldsymbol{x_t^T} R_t.\end{aligned}$$

□

3 Dynamic Generalized Linear Model

The generalized linear model extends the linear model by dropping the normality assumption and relating a function of the mean linearly to a vector of regression parameters.

3.1 Definition. ([McN89, p. 27]) Let $Y_1, ..., Y_n$ be independent random variables with $Y_i \sim \text{EXPF}(\eta_i, \phi, a, b)$ for some functions a, b, and $\mu_i := E(Y_i)$. Denote by $M \subseteq \mathbb{R}$ the set of possible values for μ_i. Let $\boldsymbol{x_1}, ..., \boldsymbol{x_n} \in \mathbb{R}^p$ be known vectors and $g : M \to \mathbb{R}$ an invertible function. Then, for an unknown parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$, a generalized linear model (GLM) is defined through the equations

$$g(\mu_i) = \boldsymbol{x_i^T \theta} \ , i = 1, ..., n, \tag{20}$$

which are often referred to as the *systematic component* of the GLM. The function g is called the *link function*.

3.2 Remark. In [Wes85a], the equation $g(\eta_i) = \boldsymbol{x_i^T \theta_i}$ is used instead. Since $\mu_i = a'(\eta_i)$ (compare with (3)), then for an invertible a', we can write

$$\hat{g}(\mu_i) := g(a'^{-1}(\mu_i)) = g(\eta_i) = \boldsymbol{x_i^T \theta_i}.$$

Thus these two models are interchangeable. We continue by using the relation $g(\eta_i) = \boldsymbol{x_i^T \theta_i}$.

To extend Definition 2.2 to GLMs, we need to consider how prior information about $\boldsymbol{\theta_t}$ should be represented. As in Chapter 2, for ease of use, the prior distribution for η_t should be of conjugate form given in Lemma 1.6. Before, the conjugate prior distribution for η_t was a normal distribution, therefore, assuming the prior distribution of $\boldsymbol{\theta_t}$ to be multivariate-normal, the linear model $\eta_t = \mu_t = \boldsymbol{x_t^T \theta_t}$ leads to a normal prior for η_t. But choosing an appropriate distribution for $\boldsymbol{\theta_t}$, in general, if η_t is supposed to be of conjugate form given in Lemma 1.6 and instead of the linear model we have a more complex relation as in equation (20) is not a straightforward task. Thus an alternative approach is needed. Instead of a full prior distribution for $\boldsymbol{\theta_t}$, only the first two moments of $\boldsymbol{\theta_t}$ are considered in [Wes85a]. This suffices to determine the parameters for the conjugate prior of η_t.

3.3 Definition. ([Wes85a])
Let $\phi > 0$ and for $\Theta, \mathcal{Y} \subset \mathbb{R}$, $a : \Theta \to \mathbb{R}$, $b : \mathbb{R}_{>0} \times \mathcal{Y} \to \mathbb{R} \setminus \{0\}$. For $t \in \mathbb{N}$, let

- $Y_t \sim \mathrm{EXPF}(\eta_t, \phi, a, b)$ be a random variable taking values in $\mathcal{Y}$,
- $\eta_t \sim \mathrm{CP}[\alpha, \beta]$ be a random variable taking values in Θ, where $\mathrm{CP}[\alpha, \beta]$ is the conjugate prior for $\mathrm{EXPF}(\eta_t, \phi, a, b)$,
- $\boldsymbol{x_t} \in \mathbb{R}^p$ be a known vector,
- $G_t \in \mathbb{R}^{p \times p}$ be the transition matrix,
- $B_t \in \mathbb{R}^{p \times p}$ be a diagonal matrix of positive discount factors (compare Remark 2.5).

Additionally, let

- $C_0 \in \mathbb{R}^{p \times p}$ a symmetric, positive definite matrix and $\boldsymbol{m_0} \in \mathbb{R}^p$,
- $g : \mathbb{R} \to \mathbb{R}$ an invertible function.

Denote $\boldsymbol{a_t} := \mathbb{E}[\boldsymbol{\theta_t} \,|\, D_{t-1}], R_t := \mathbb{V}[\boldsymbol{\theta_t} \,|\, D_{t-1}]$ (the prior moments at time t) for $t \in \mathbb{N}$ and $\boldsymbol{m_t} = \mathbb{E}[\boldsymbol{\theta_t} \,|\, D_t], C_t = \mathbb{V}[\boldsymbol{\theta_t} \,|\, D_t]$ (the posterior moments at time t) for $t \in \mathbb{N}_0$. The notation D_t is used as in Definition 2.2.

The *dynamic generalized linear model* (DGLM) $((G_t)_{t\in\mathbb{N}}, (B_t)_{t\in\mathbb{N}}, (m_0), C_0, g)$ for the process $(Y_t \,|\, \eta_t, \phi)$, at time t is specified by the *guide relation*

$$g(\eta_t) \simeq \boldsymbol{x_t^T \theta_t} \tag{21}$$

and the state evolution equations

$$\boldsymbol{a_t} = G_t \boldsymbol{m_{t-1}} \tag{22}$$

$$R_t = B_t G_t C_{t-1} G_t^T B_t. \tag{23}$$

3.4 Remark. ([Wes85a]) The guide relation (21) is to be interpreted as one possible way to choose the parameters α_t, β_t for the prior distribution of η_t. That means if $g(\eta_t) = \boldsymbol{x_t^T}\boldsymbol{\theta}_t$, the mean and variance of $\boldsymbol{x_t^T \theta_t}$ determine α_t, β_t. Alternatively, the prior parameters might be fixed in some other way. In the following analysis, the guide relation is used to determine the prior for η_t.

Notice that the crucial step to provide analysis for the model given in Definition 3.3 is the derivation of posterior moments for $\boldsymbol{\theta_t}$. As no distributional assumptions were made about $\boldsymbol{\theta_t}$, such moments can't be obtained directly. Instead, the linear Bayesian approach is used in [Wes85a]. In the linear Bayesian approach, the expectation is replaced by a linear predictor. To motivate this approach, let's review the fact that the mean minimises the quadratic risk function.

3.5 Definition. ([Rob94, p. 49, 70]) Let θ be a, possibly vector valued, random variable taking values in some set Θ. The *quadratic risk function* for θ is defined as

$$r(\theta, \hat{\theta}) := \mathbb{E}[(\theta - \hat{\theta})(\theta - \hat{\theta})^T],\ \hat{\theta} \in \Theta.$$

We might also want to calculate the expectation based on a conditional density, e.g. $p(\boldsymbol{\theta_t} \,|\, \eta_t, D_{t-1})$ for $\boldsymbol{\theta_t}$. In this case we write $r(\boldsymbol{\theta_t}, \hat{\boldsymbol{\theta}}_{\boldsymbol{t}} \,|\, \eta_t, D_{t-1})$ for the conditional expectation $\mathbb{E}[(\boldsymbol{\theta_t} - \hat{\boldsymbol{\theta}}_{\boldsymbol{t}})(\boldsymbol{\theta_t} - \hat{\boldsymbol{\theta}}_{\boldsymbol{t}})^T \,|\, \eta_t, D_{t-1}]$.

3.6 Remark. The expectation $r(\boldsymbol{\theta_t}, \hat{\boldsymbol{\theta}}_{\boldsymbol{t}} \,|\, \eta_t, D_{t-1})$ can either denote the conditional expected value for some fixed value of η_t or the conditional expectation (which is a function of η_t) if no value is yet fixed for η_t. The same holds for similar expressions later, which involve conditioning on η_t.

3.7 Lemma. ([Ross98, p. 350]) Let θ, η be scalar valued random variables such that $E[\theta \,|\, \eta]$ is defined. Denoted by $H \subset \mathbb{R}$ the set of possible values for η. Then, the conditional expectation $E[\theta \,|\, \eta]$ is the best predictor in the sense that for any function $f : H \to \mathbb{R}$ that is measurable with respect to the appropriate probability spaces, there holds

$$r(\theta, f(\eta) \,|\, \eta) \geq r(\theta, E[\theta \,|\, \eta] \,|\, \eta).$$

Additionally,

$$r(\theta, E[\theta \,|\, \eta] \,|\, \eta) = \mathbb{V}(\theta \,|\, \eta).$$

An appropriate extension of Lemma 3.7 for vector valued θ is obtained by taking the trace of the quadratic risk.

3.8 Corollary. For $p \in \mathbb{N}$ let $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)^T$ be a random vector and η a scalar valued random variable such that $E[\boldsymbol{\theta} \,|\, \eta]$ is defined. Denoted by $H \subset \mathbb{R}$ the set of possible values for η. For a matrix A, denote by $\mathrm{tr}(A)$ the trace of A. Then the conditional expectation $E[\boldsymbol{\theta} \,|\, \eta]$ is the best predictor in the sense that for any measurable function $f : H \to \mathbb{R}^p$, there holds

$$\mathrm{tr}(r(\boldsymbol{\theta}, f(\eta) \,|\, \eta)) \geq \mathrm{tr}(r(\boldsymbol{\theta}, E[\boldsymbol{\theta} \,|\, \eta] \,|\, \eta)).$$

Proof. Minimizing the quadratic risk of the individual entries of $\boldsymbol{\theta}$ minimizes the sum of them. Since the trace of $r(\boldsymbol{\theta}, f(\eta) \,|\, \eta)$ is the sum of its diagonal entries, which are $r(\theta_i, f(\eta)_i \,|\, \eta)$, the trace is minimized by applying Lemma 3.7 to all the diagonal entries. The minimal value is obtained at $E[\boldsymbol{\theta} \,|\, \eta]$ since per definition its entries are $E[\theta_i \,|\, \eta]$. □

As stated beforehand, the conditional expectation $E[\theta \,|\, \eta]$ will be replaced by a linear predictor. The linear predictor minimizing the quadratic risk is chosen for this.

3.9 Lemma. ([Gol07, p. 56]) Let θ, η be scalar valued random variables such that $E[\theta \,|\, \eta]$ is defined. Denoted by $H \subset \mathbb{R}$ the set of possible values for η. Let $g : H \to \mathbb{R}$ be some measurable function. Then

$$\operatorname*{argmin}_{(d_0, d_1)^T \in \mathbb{R}^2} r(\theta, d_0 + d_1 \cdot g(\eta) \,|\, \eta) = \big(\mathbb{E}[\theta] - \mathrm{Cov}(\theta, g(\eta)))\, \mathbb{V}(g(\eta))^{-1}\, \mathbb{E}[g(\eta)], \mathrm{Cov}(\theta, g(\eta)))\, \mathbb{V}(g(\eta))^{-1}\big)$$
$$=: (d_0^*, d_1^*).$$

3.10 Corollary. Let $\boldsymbol{\theta}$ be a random vector taking values in a k-dimensional space and η a scalar valued random variable such that $E[\boldsymbol{\theta} \mid \eta]$ is defined. Denoted by $H \subset \mathbb{R}$ the set of possible values for η. Let $g : H \to \mathbb{R}$ be some measurable function. Then

$$\begin{aligned}\underset{(\boldsymbol{d_0},\boldsymbol{d_1})^{\boldsymbol{T}}\in\mathbb{R}^{2k}}{\operatorname{argmin}} \ & \operatorname{tr}\left(r(\boldsymbol{\theta}, \boldsymbol{d_0} + \boldsymbol{d_1} \cdot g(\eta) \mid \eta)\right) \\ &= \left(\mathbb{E}[\boldsymbol{\theta}] - \operatorname{Cov}(\boldsymbol{\theta}, g(\eta)))\, \mathbb{V}(g(\eta))^{-1}\, \mathbb{E}[g(\eta)], \operatorname{Cov}(\boldsymbol{\theta}, g(\eta))\, \mathbb{V}\left(g(\eta)\right)^{-1}\right) \\ &=: \boldsymbol{(d_0^*, d_1^*)}.\end{aligned}$$

Proof. The proof follows analogously to the proof of Corollary 3.8 using Lemma 3.9. □

We now apply these results to our model in order to obtain linear predictors for the posterior moments for $\boldsymbol{\theta_t}$. The posterior moments for $\boldsymbol{\theta_t}$ can also be obtained by considering information about η_t at time t. The next result has already been proven in the second proof of Proposition 2.4.

3.11 Lemma. In the setting of Definition 3.3, there holds

$$\begin{aligned}&\mathbb{E}[\boldsymbol{\theta_t} \mid D_t] = \mathbb{E}[E[\boldsymbol{\theta_t} \mid \eta_t, D_{t-1}] \mid D_t] \\ &\mathbb{V}[\boldsymbol{\theta_t} \mid D_t] = \mathbb{V}[E[\boldsymbol{\theta_t} \mid \eta_t, D_{t-1}] \mid D_t] + \mathbb{E}[V[\theta_t \mid \eta_t, D_{t-1}] \mid D_t].\end{aligned}$$

Motivated by the guide relation, the predictor for $E[\boldsymbol{\theta_t} \mid \eta_t, D_{t-1}]$ is sought in the class of predictors of the form $\boldsymbol{d_0} + \boldsymbol{d_1} \cdot g(\eta_t)$, $\boldsymbol{d_0}, \boldsymbol{d_1} \in \mathbb{R}^p$. At time t, all the necessary prior information about $g(\eta_t)$ to apply Corollary 3.10 is available.

3.12 Lemma. ([Wes85a]) In the setting of Definition 3.3 and assuming that the guide relation $g(\eta_t) = \boldsymbol{x_t}^T \boldsymbol{\theta_t}$ holds for all $t \in \mathbb{N}$, there holds

$$\begin{aligned}\mathbb{E}[g(\eta_t) \mid D_{t-1}] &= \boldsymbol{x_t^T a_t} =: f_t \\ \mathbb{V}[g(\eta_t) \mid D_{t-1}] &= \boldsymbol{x_t^T} R_t \boldsymbol{x_t} =: q_t \\ \operatorname{Cov}(\boldsymbol{\theta_t}, g(\eta_t)) \mid D_{t-1}) &= \boldsymbol{x_t^T} R_t\end{aligned}$$

Proof. These equations follow from the guide relation and state evolution equations by the properties of expectation and covariance. □

3.13 Corollary. ([Wes85a]) In the setting of Definition 3.3 and assuming $g(\eta_t) = \boldsymbol{x_t}^T \boldsymbol{\theta_t}$, the linear predictor $\boldsymbol{d^*} = \boldsymbol{d_0^*} + \boldsymbol{d_1^*} \cdot g(\eta)$ minimizing $tr\left(r(\boldsymbol{\theta}, \boldsymbol{d_0} + \boldsymbol{d_1} \cdot g(\eta_t) \mid \eta_t, D_{t-1})\right)$ is given by

$$\boldsymbol{d}^* = \boldsymbol{a_t} + \frac{(g(\eta_t) - \boldsymbol{x_t^T a_t})}{\boldsymbol{x_t^T} R_t \boldsymbol{x_t}} R_t \boldsymbol{x_t}. \tag{24}$$

The value of the quadratic risk function at $\boldsymbol{d^*}$ is

$$r(\boldsymbol{\theta}, \boldsymbol{d}^* \mid \eta, D_{t-1}) = R_t - \frac{1}{\boldsymbol{x_t^T} R_t \boldsymbol{x_t}} R_t \boldsymbol{x_t} \boldsymbol{x_t^T} R_t.$$

3.14 Proposition. Let a DGLM $\{(G_t)_t, (B_t)_t, \boldsymbol{m_0}, C_0, g\}$ for a process $\{Y_t \mid \eta_t, \phi\}$ be given. Denote by $g_t := \mathbb{E}[g(\eta_t) \mid D_t]$ and $p_t := \mathbb{V}[g(\eta_t) \mid D_t]$ the posterior moments of $g(\eta_t)$. Replacing in Lemma 3.11 $E[\boldsymbol{\theta_t} \mid \eta_t, D_{t-1}]$ by the linear predictor $\boldsymbol{d^*}$ obtained in Corollary 3.13 and $V[\boldsymbol{\theta_t} \mid \eta_t, D_{t-1}]$ by $r(\boldsymbol{\theta_t}, \boldsymbol{d} \mid \eta_t, D_{t-1})$ we obtain

$$\mathbb{E}[\boldsymbol{\theta_t} \,|\, D_t] \simeq \mathbb{E}[\boldsymbol{d}^* \,|\, D_t] = \boldsymbol{a_t} + \frac{(g_t - \boldsymbol{x_t^T a_t})}{q_t} R_t \boldsymbol{x_t} \tag{25}$$

$$\mathbb{V}[\boldsymbol{\theta_t} \,|\, D_t] \simeq \mathbb{V}[\boldsymbol{d}^* \,|\, D_t] + \mathbb{E}[r(\boldsymbol{\theta_t}, \boldsymbol{d}^* \,|\, \eta_t, D_{t-1}) \,|\, D_t] = R_t - \frac{(1 - \frac{p_t}{q_t})}{q_t} R_t \boldsymbol{x_t x_t^T} R_t, \tag{26}$$

where $q_t = \boldsymbol{x_t^T} R_t \boldsymbol{x_t}$.

Proof. In equation (24), the only random term is $g(\eta_t)$, thus we obtain (25)

$$\mathbb{E}[\boldsymbol{d}^* \,|\, D_t] = \mathbb{E}[\boldsymbol{a_t} + \frac{(g(\eta_t) - \boldsymbol{x_t^T a_t})}{\boldsymbol{x_t^T} R_t \boldsymbol{x_t}} \boldsymbol{x_t^T} R_t \,|\, D_t] = \boldsymbol{a_t} + \frac{(g_t - \boldsymbol{x_t^T a_t})}{q_t} R_t \boldsymbol{x_t}.$$

The expression $r(\boldsymbol{\theta}, \boldsymbol{d}^* \,|\, \eta, D_{t-1})$ is a constant, so we get

$$\begin{aligned}
\mathbb{V}[\boldsymbol{d}^* \,|\, D_t] + \mathbb{E}[r(\boldsymbol{\theta_t}, \boldsymbol{d}^* \,|\, \eta_t, D_{t-1}) \,|\, D_t] &\overset{(26)}{=} \mathbb{V}\left[\boldsymbol{a_t} + \frac{(g(\eta_t) - \boldsymbol{x_t^T a_t})}{\boldsymbol{x_t^T} R_t \boldsymbol{x_t}} R_t \boldsymbol{x_t} \,|\, D_t\right] \\
&\quad + \mathbb{E}\left[R_t - \frac{1}{\boldsymbol{x_t^T} R_t \boldsymbol{x_t}} R_t \boldsymbol{x_t x_t^T} R_t \,|\, D_t\right] \\
&= \frac{R_t \boldsymbol{x_t x_t^T} R_t}{(\boldsymbol{x_t^T} R_t \boldsymbol{x_t})^2} \mathbb{V}[(g(\eta_t) \,|\, D_t] + R_t - \frac{1}{\boldsymbol{x_t^T} R_t \boldsymbol{x_t}} R_t \boldsymbol{x_t x_t^T} R_t \\
&= \frac{R_t \boldsymbol{x_t x_t^T} R_t}{(\boldsymbol{x_t^T} R_t \boldsymbol{x_t})^2} p_t + R_t - \frac{1}{\boldsymbol{x_t^T} R_t \boldsymbol{x_t}} R_t \boldsymbol{x_t x_t^T} R_t \\
&= R_t - \frac{(1 - \frac{p_t}{q_t})}{q_t} R_t \boldsymbol{x_t x_t^T} R_t.
\end{aligned}$$

□

3.15 Remark. Notice that the DLM is a special case of the DGLM. The results in Proposition 2.4 coincide with those in Proposition 3.14 for the normal distribution.

3.16 Remark. We now summarize the analysis based on the above results. At time $t \in \mathbb{N}$

1. Calculate the prior moments $\boldsymbol{a_t}$, R_t for $\boldsymbol{\theta_t}$ according to the state evolution equations (22) and (23).
2. Set $(\eta_t \,|\, D_{t-1}) \sim CP[\alpha_t, \beta_t]$ with α_t, β_t chosen so that equations in Lemma 3.12 hold.
3. Calculate the posterior for η_t based on observed data y_t. This has the conjugate form given in Lemma 1.6.
4. Based on the posterior distribution for η, determine the posterior moments g_t and p_t for $g(\eta_t)$.
5. Set the posterior moments of $\boldsymbol{\theta_t}$ to (25) and (26).

As the last step, the means to make predictions about future data are provided. For a random variable $\boldsymbol{\theta}$ with expectation $\boldsymbol{a}$ and covariance matrix R, we write $\boldsymbol{\theta} \sim [\boldsymbol{a}, R]$.

3.17 Remark. ([Wes85a]) Let a DGLM $\{(G_t)_t, (B_t)_t, \boldsymbol{m_0}, C_0, g\}$ for a process $(Y_t \mid \eta_t, \phi)_{t\in\mathbb{N}}$ be given. For $t \in \mathbb{N}, k \in \mathbb{N}$ define

$$D_t(k) := D_{t+k} \setminus \{Y_{t+1}, ..., Y_{t+k}\}.$$

The prior moments for $\boldsymbol{\theta_{t+k}}$ given $D_{t-1}(k)$ are defined by extending the state evolution equations over multiple time steps recursively for $j = 1, ..., k$ as follows

$$\begin{aligned}&\boldsymbol{a_t(0)} = \boldsymbol{a_t},\ \boldsymbol{a_t(j)} = G_{t+j}\boldsymbol{a_t(j-1)},\\ &R_t(0) = R_t,\ R_t(j-1) = B_{t,j}G_{t+j}R_t(j-1)G_{t+j}^T B_{t,j},\end{aligned}$$

So we set

$$(\boldsymbol{\theta_{t+k}} \mid D_{t-1}(k)) \sim [a_t(k), R_t(k)].$$

The discount matrix $B_{t,j}$ can, for example, be chosen as $B_{t,j} = B_{t+j}$. As before, the conjugate prior $C[\alpha_t(k), \beta_t(k)]$ for the natural parameter η_{t+k} is chosen in accordance to the guide relation that means

$$\begin{aligned}\mathbb{E}[g(\eta_{t+k}) \mid D_{t-1}(k)] &= \boldsymbol{x_{t+k}^T a_t(k)},\\ \mathbb{V}[g(\eta_{t+k}) \mid D_{t-1}(k)] &= \boldsymbol{x_{t+k}^T} R_t(k)\boldsymbol{x_{t+k}}.\end{aligned}$$

Then, the prior predictive distribution $p(Y_{t+k} \mid D_{t-1}(k))$ has the conjugate form given in Lemma 1.6.

In the next example, an application of the previous results is presented.

3.18 Example. Let $Y_t \sim \mathcal{B}(k, \mu_t)$ for $k, t \in \mathbb{N}$. The density of a $\mathcal{B}(k, \mu_t)$-distribution can be written in exponential family form (see Definition 1.4)

$$p(y_t \mid \mu_t) = \binom{k}{y_t}\mu_t^{y_t}(1-\mu_t)^{k-y_t} = \exp\left[\phi\left\{y_t\eta_t - a(\eta_t)\right\}\right] b(y_t, \phi),\ y_t \in \{0, ..., k\}$$

for $\phi := 1,\ \eta_t := \ln\left(\frac{\mu_t}{1-\mu_t}\right),\ a(\eta_t) := k \cdot \ln\left(1 + \exp(\eta_t)\right),\ b(y_t, \phi) := \binom{k}{y_t}$. To find the moments of η_t it is easier to first consider a conjugate prior for the mean μ_t instead.

The conjugate prior for the parameter μ_t is the beta distribution $\mathrm{Beta}(\alpha_t, \beta_t)$ for $\alpha_t, \beta_t > 0$. For an observed value y_t of Y_t, the posterior is a $\mathrm{Beta}(\alpha_t + y_t, \beta_t + k - y_t)$ distribution (see [Hof09, p. 51]). If $\mu_t \sim \mathrm{Beta}(\alpha_t, \beta_t)$, then the natural parameter $\eta_t = \ln\left(\frac{\mu_t}{1-\mu_t}\right)$ has a distribution obtained from transforming the $\mathrm{Beta}(\alpha_t, \beta_t)$ distribution, which is exactly of the conjugate form $CP(\hat{\alpha}_t, \hat{\beta}_t)$ given in Lemma 1.6 with $\hat{\alpha}_t = \alpha_t$, $\hat{\beta}_t = \frac{\alpha+\beta}{k}$ and $c(\hat{\alpha}_t, \hat{\beta}_t) = \frac{1}{\mathrm{B}(\hat{\alpha}_t, k\hat{\beta}_t - \hat{\alpha}_t)}$. Also, the posterior distribution of η_t is a $CP(\hat{\alpha}_t + y_t, \hat{\beta}_t + 1)$-distribution which coincides with the result from Lemma 1.6.

If we choose g to be the identity function, then the following moments are of interest

$$\mathbb{E}\left[\eta_t \mid D_{t-1}\right] = \mathbb{E}\left[\ln\left(\frac{\mu_t}{1-\mu_t}\right) \mid D_{t-1}\right] = \gamma(\alpha_t) - \gamma(\beta_t) \tag{27}$$

$$\mathbb{V}\left[\eta_t \mid D_{t-1}\right] = \mathbb{V}\left[\ln\left(\frac{\mu_t}{1-\mu_t}\right) \mid D_{t-1}\right] = \dot{\gamma}(\alpha_t) + \dot{\gamma}(\beta_t), \tag{28}$$

where $\gamma(\cdot)$ is the digamma function and $\dot{\gamma}(\cdot)$ is the trigamma function (see [Wes85a, Ch 5.]). In Step 2 of Remark 3.16 we need to solve the equations

$$\mathbb{E}[\eta_t \mid D_{t-1}] = f_t = \boldsymbol{x_t^T a_t},\ \mathbb{V}[\eta_t \mid D_{t-1}] = q_t = \boldsymbol{x_t^T} R_t \boldsymbol{x_t}$$

in α_t, β_t. At best, the digamma and trigamma functions can be approximated by non-linear functions (see [Abr72, p. 258 - 259]) so solving the above equations becomes a non-trivial task. Alternatively, in [Wes85a, Ch. 5], the parameters α_t and β_t are obtained by replacing f_t through the mode of $(\eta_t \mid D_{t-1})$ and q_t^{-1} through the curvature $-\frac{\partial^2}{\partial \eta_t} \ln(p(\eta_t \mid D_{t-1}))$ at the mode. For the mode f_t, there holds

$$\frac{\partial}{\partial \eta_t} p(\eta_t \mid D_{t-1})|_{\eta_t = f_t} = 0 \Leftrightarrow \hat{\alpha}_t = \hat{\beta}_t \frac{\partial}{\partial \eta_t} a(\eta_t)|_{\eta_t = f_t} \Leftrightarrow \frac{\alpha_t}{\alpha_t + \beta_t} = \frac{\exp(f_t)}{1 + \exp(f_t)}.$$

And for the curvature

$$q_t^{-1} = -\frac{\partial^2}{\partial \eta_t^2} \ln(p(\eta_t \mid D_{t-1})|_{\eta_t = f_t} = \hat{\beta}_t \frac{\partial^2}{\partial \eta_t^2} a(\eta_t)|_{\eta_t = f_t} = (\alpha_t + \beta_t) \frac{\exp(f_t)}{(1 + \exp(f_t))^2}.$$

Which can be easily solved for α_t and β_t, resulting in

$$\alpha_t = q_t^{-1} \left(1 + \exp(f_t)\right),\ \beta_t = q_t^{-1} \left(1 + \exp(-f_t)\right) \tag{29}$$

Then, after calculating the posterior parameters for μ_t, the posterior moments g_t, p_t for η_t can be approximated from analogous equations to 27 and 28.

Finally, from Lemma 1.6, we obtain the density of the prior predictive distribution for Y_t

$$\begin{aligned} p(y_t \mid D_{t-1}) &= \frac{c(\hat{\alpha}_t, \hat{\beta}_t)}{c(\hat{\alpha}_t + y_t, \hat{\beta}_t + 1)} b(y_t, \phi) \\ &= \frac{\text{Beta}\left(\hat{\alpha}_t + y_t, k(\hat{\beta}_t + 1) - (\hat{\alpha}_t + y_t)\right)}{\text{Beta}\left(\hat{\alpha}_t, k\hat{\beta}_t - \hat{\alpha}_t\right)} \binom{k}{y_t} \\ &= \frac{\text{Beta}(\alpha_t + y_t, \beta_t + k - y_t)}{\text{Beta}(\alpha_t, \beta_t)} \binom{k}{y_t},\ y_t \in \{0, ..., k\}, \end{aligned}$$

which is the density of a beta-binomial distribution $\text{BetaBin}(k, \alpha_t, \beta_t)$ with mean $k\frac{\alpha_t}{\alpha_t + \beta_t}$ (see [Gel04, p. 576 - 577]). Applications of this model on real data sets are demonstrated in Chapter 5.

4 Random Scale Parameters and Outliers

In [Wes85a], some additional extensions concerning the scale parameters and outliers to the models discussed in Chapters 2 and 3 were proposed. These extensions are discussed in greater detail in [Wes85b]. As first step in this chapter, we discuss the extension of the dynamic linear model where the scale parameter ϕ is also random. For this, a conjugate prior for the vector $(\boldsymbol{\theta}, \phi)$, the normal-gamma distribution, is introduced.

4.1 Definition. ([Ber94, p. 140]) Let $\boldsymbol{a} \in \mathbb{R}^p, \vartheta > 0, \delta > 0$ and $R \in \mathbb{R}^{p \times p}$ symmetric and positive-definite. Let $\boldsymbol{\theta}$ be a continuous, $p-$dimensional random vector and ϕ a continuous random variable. The random vector $(\boldsymbol{\theta}, \phi)$ has a p-dimensional *normal-gamma distribution* $\mathcal{NG}_p(\boldsymbol{a}, R, \vartheta, \delta)$ if $(\boldsymbol{\theta} \,|\, \phi) \sim \mathcal{N}_p(\boldsymbol{a}, \phi^{-1}R)$ and $\phi \sim \mathcal{G}(\vartheta, \delta)$, where $\mathcal{G}(\vartheta, \delta)$ denotes the gamma distribution with parameters ϑ and δ.

4.2 Remark. The density of $\mathcal{NG}_p(\boldsymbol{a}, R, \vartheta, \delta)$ is the product of densities of the $\mathcal{N}_p(\boldsymbol{a}, \phi^{-1}R)$ and $\mathcal{G}(\vartheta, \delta)$ distributions.

In the context of linear models, the normal-gamma priors can be used both for modelling the prior information about $(\boldsymbol{\theta}, \phi)$ as well as about (η, ϕ). The key to understand the relationship between these is the next property.

4.3 Lemma. Let $(\boldsymbol{\theta}, \phi) \sim \mathcal{NG}_p(\boldsymbol{a}, R, \vartheta, \delta)$. Then for $\boldsymbol{x} \in \mathbb{R}^p \setminus \{0\}$, there holds

$$(\boldsymbol{x^T\theta}, \phi) \sim \mathcal{NG}(\boldsymbol{x^T a}, \boldsymbol{x}^T R\boldsymbol{x}, \vartheta, \delta).$$

Proof. By the properties of the normal distribution, $(\boldsymbol{x^T\theta} \,|\, \phi) \sim \mathcal{N}(\boldsymbol{x^T a}, \phi^{-1}\boldsymbol{x}^T R\boldsymbol{x})$, thus by the definition of the one-dimensional normal-gamma distribution, the distribution of $(\boldsymbol{x^T\theta}, \phi)$ is a $\mathcal{NG}(\boldsymbol{x^T a}, \boldsymbol{x}^T R\boldsymbol{x}, \vartheta, \delta)$-distribution. □

Notice that the transformation $\boldsymbol{x^T\theta}$ does not change the distribution of ϕ. Therefore, information about ϕ can also be obtained by considering its joint distribution with $\boldsymbol{x^T}\theta$ instead of $\boldsymbol{\theta}$. For example, a posterior distribution for ϕ can be derived this way.

4.4 Lemma. Assuming that $(\eta, \phi) \sim \mathcal{NG}(\eta_0, q, \vartheta, \delta)$ and $(Y \,|\, \eta, \phi) \sim \mathcal{N}(\eta, \phi^{-1})$, the posterior distribution $(\eta, \phi \,|\, Y = y)$ is

$$\mathcal{NG}(\eta', q', \vartheta + \frac{1}{2}, \delta'),$$

where $\eta' = \frac{y+q^{-1}\eta_0}{1+q^{-1}}, q' = (1+q^{-1})^{-1}$ and $\delta' = \delta + \frac{1}{2(1+q)}(y-\eta_0)^2$,
i.e. the normal-gamma distribution is a conjugate prior for the normal distribution.

Proof. Using the Bayes rule and the assumptions about the distributions at hand, there holds

$$\begin{aligned} p(\eta, \phi \,|\, y) &\propto p(y \,|\, \eta, \phi) \cdot p(\eta, \phi) \\ &\propto \phi^{1/2} \exp\left[-\frac{\phi}{2}(y-\eta)^2\right] \cdot \phi^{1/2} \exp\left[\frac{-\phi q^{-1}}{2}(\eta-\eta_0)^2\right] \cdot \phi^{\vartheta-1}\exp[-\phi\delta]. \end{aligned}$$

We can write

$$\begin{aligned} \exp\left[-\frac{\phi}{2}(y-\eta)^2\right] \cdot \exp\left[\frac{-\phi q^{-1}}{2}(\eta-\eta_0)^2\right] &= \exp\left[-\frac{\phi}{2}\left((1+q^{-1})\eta^2 - 2\eta(y+q^{-1}\eta_0) + y^2 + q^{-1}\eta_0^2\right)\right] \\ &= \exp\left[-\frac{\phi(1+q^{-1})}{2}\left(\eta^2 - 2\eta\frac{(y+q^{-1}\eta_0)}{(1+q^{-1})} + \frac{y^2+q^{-1}\eta_0^2}{(1+q^{-1})}\right)\right] \\ &= \exp\left[-\frac{\phi(1+q^{-1})}{2}\left(\eta - \frac{(y+q^{-1}\eta_0)}{(1+q^{-1})}\right)^2\right] \\ &\cdot \exp\left[-\frac{\phi(1+q^{-1})}{2}\left(\frac{y^2+q^{-1}\eta_0^2}{(1+q^{-1})} - \left(\frac{(y+q^{-1}\eta_0)}{(1+q^{-1})}\right)^2\right)\right] \\ &= \exp\left[-\frac{\phi(1+q^{-1})}{2}\left(\eta - \frac{(y+q^{-1}\eta_0)}{(1+q^{-1})}\right)^2\right] \cdot \exp\left[-\phi\left(\frac{1}{2(1+q)}(y-\eta_0)^2\right)\right]. \end{aligned}$$

Thus, we have

$$p(\eta,\phi\,|\,y) \propto \phi^{\frac{1}{2}} \exp\left[-\frac{\phi(1+q^{-1})}{2}\left(\eta-\frac{(y+q^{-1}\eta_0)}{(1+q^{-1})}\right)^2\right]\cdot\phi^{\vartheta+\frac{1}{2}-1}\exp\left[-\phi\left(\delta+\frac{1}{2(1+q)}(y-\eta_0)^2\right)\right]$$

which is proportional to the density of $\mathcal{NG}(\eta', q', \vartheta+\frac{1}{2}, \delta')$. □

4.5 Remark. To extend Definition 2.2, ϕ is assumed to be a random variable with $(\phi\,|\,D_t) \sim \mathcal{G}(\vartheta_t, \delta_t)$, $t \in \mathbb{N}_0$. The state evolution equations are now replaced by

$$(\boldsymbol{\theta_t}\,|\,\phi, D_{t-1}) \sim \mathcal{N}_p(\boldsymbol{a_t}, \phi^{-1}R_t),\ t \in \mathbb{N},$$

with $\boldsymbol{a_t}, R_t$ as in the results from Proposition 2.4. Then, $(\boldsymbol{\theta_t}, \phi\,|\,D_{t-1}) \sim \mathcal{NG}_p(\boldsymbol{a_t}, R_t, \vartheta_{t-1}, \delta_{t-1})$. We denote a DLM with a random scale parameter ϕ by $((G_t)_t, (W_t)_t, \boldsymbol{m_0}, C_0, \vartheta_0, \delta_0)$.

4.6 Proposition. Let a DLM $\{(G_t)_t, (W_t)_t, \boldsymbol{m_0}, C_0, \vartheta_0, \delta_0\}$ for a normal process $(Y_t\,|\,\theta_t, \phi)_{t\in\mathbb{N}}$ be given. Then $(\boldsymbol{\theta_t}, \phi\,|\,D_t) \sim \mathcal{NG}_p(m_t, C_t, \vartheta_t, \delta_t)$ for all $t \in \mathbb{N}$, where

$$\begin{aligned}
\boldsymbol{m_t} &= \boldsymbol{a_t} + \frac{(y_t - f_t)}{\phi q_t + 1} R_t \boldsymbol{x_t},\\
C_t &= R_t - \frac{1}{\phi q_t + 1} R_t \boldsymbol{x_t} \boldsymbol{x_t^T} R_t,\\
\vartheta_t &= \vartheta_{t-1} + \frac{1}{2},\\
\delta_t &= \delta_{t-1} + \frac{1}{2(1+q_t)}(y_t - f_t)^2,
\end{aligned}$$

for $q_t = \boldsymbol{x_t^T} R_t \boldsymbol{x_t}$ and $f_t = \boldsymbol{x_t^T}\boldsymbol{a_t}$.

Proof. The parameters $\boldsymbol{m_t}$ and C_t are the posterior moments for the normal prior $(\boldsymbol{\theta_t}\,|\,\phi, D_{t-1})$ and normal sample distribution of Y_t and can be calculated as in Example 2.1 (analogously to the proof of Proposition 2.4). By Lemma 4.3 we also have $(\eta_t, \phi\,|\,D_{t-1}) = (\boldsymbol{x_t}\boldsymbol{\theta_t}, \phi\,|\,D_{t-1}) \sim \mathcal{NG}(f_t, q_t, \vartheta_{t-1}, \delta_{t-1})$. Using Lemma 4.4, the values for the parameters ϑ_t and δ_t are obtained. □

Modelling ϕ can also be used as a tool to determine how good the systematic component is at describing the data. Recall from Remark 1.5 that the variance in the exponential family equals $\phi^{-1}a''(\eta)$. If the posterior for ϕ indicates a high probability for values smaller than one, the systematic component of the model is better at describing data. If the opposite is the case, there is some extra variation in the data that can't be explained by the systematic components alone (compare with [Wes85b, Ch. 2.1]).

For continuous distributions in the exponential family, conjugate priors for ϕ exist, so in these cases the analysis is similar to DLMs (see [Wes85b, Ch. 2.1]). For discrete distributions, the scale parameter ϕ is always fixed at one. In [Wes85b], a method for discrete distributions is developed to assess whether the systematic component is adequate at describing the variance in the data. For this, the model is embedded into a general class of models where ϕ not restricted to one. The general models are constructed by using the concept of deviance which is introduced next.

4.7 Definition. ([McN89, p. 33]) Let Y be a random variable with possible values in $\mathcal{Y}$ following a distribution P_μ, where μ is some parameter with possible values in some set M. Denote by $p(y \,|\, \mu)$ the density of P_μ. For a value $y \in \mathcal{Y}$, let $\mu_s(y)$ be the maximum likelihood estimate for μ. The *deviance* is defined as

$$d(y \,|\, \hat{\mu}) := -2 \ln \left[\frac{p(y \,|\, \hat{\mu})}{p(y \,|\, \mu_s(y))} \right], \ \hat{\mu} \in M.$$

The deviance compares a model to the *saturated model*, that is the model which best fits the data. In the context of GLMs the values for the means μ_i obtained from the equations (20) can be thought of as "predictions" for the already observed data y_i. For the saturated model, these values are exactly y_i. Deviance is used to measure how much the predicted values differ from the data, i.e. how good is the model at describing the data that it was fitted from (compare with [McN89, p. 33]).

In the Bayesian setting, if the parameters have prior (or posterior) distributions, the possible values for the parameter can be compared in similar fashion to the parameter which maximizes the prior (or the posterior) density. In [Wes85a, Ch 4.2] the following definitions are proposed.

4.8 Definition. ([Wes85a, Ch. 4.2]) Let $Y \sim \mathrm{EXPF}(\eta, \phi, a, b)$ for suitable functions a, b and fixed $\phi > 0$ and let $\eta \sim \mathrm{CP}[\alpha, \beta]$ for some parameters α, β, where $CP[\alpha, \beta]$ is the conjugate prior for $\mathrm{EXPF}(\eta, \phi, a, b)$.

Let η^* be the prior mode for η, that is the value that maximizes the density of $CP[\alpha, \beta]$. The *prior deviance* for η is defined as

$$d(\eta) := -2 \ln \left[\frac{p(\eta)}{p(\eta^*)} \right].$$

In the context of DGLMs, we denote by $d(\eta_t \,|\, D_{t-1}, \phi = 1)$ the prior deviance based on the density of $(\eta_t \,|\, D_{t-1}, \phi = 1)$.

Let η^{**} be the posterior mode for η, that is the value that maximizes the density of the posterior distribution $CP[\alpha + \phi y, \beta + \phi]$ for an observed value y of Y (compare to Lemma 1.6). The *posterior deviance* for η is defined as

$$d(\eta \,|\, y) := -2 \ln \left[\frac{p(\eta \,|\, y)}{p(\eta^{**} \,|\, y)} \right].$$

Similarly, in the context of DGLMs, we denote by $d(\eta_t \,|\, D_t, \phi = 1)$ the prior deviance based on the density of $(\eta_t \,|\, D_t, \phi = 1)$. Finally, the *residual deviance* is defined as

$$d(y) := [d(y \,|\, \eta) + d(\eta)] \,|_{\eta = \eta^{**}}.$$

4.9 Lemma. ([Wes85a]) Let $Y \sim \mathrm{EXPF}(\eta, \phi, a, b)$ for suitable functions a, b and fixed $\phi > 0$ and let $\eta \sim \mathrm{CP}[\alpha, \beta]$ for some parameters α, β, where $CP[\alpha, \beta]$ is the conjugate prior for $\mathrm{EXPF}(\eta, \phi, a, b)$. There holds

$$d(y \,|\, \eta) + d(\eta) = d(y) + d(\eta \,|\, y) \tag{30}$$

4.10 Definition. Let $Y \sim \mathrm{EXPF}(\eta, \phi, a, b)$ for suitable functions a, b and fixed $\phi > 0$ and let $\eta \sim \mathrm{CP}[\alpha, \beta]$ for some parameters α, β, where $CP[\alpha, \beta]$ is the conjugate prior for $\mathrm{EXPF}(\eta, \phi, a, b)$. We define

$$m(y \,|\, \eta, \phi) := \phi^{1/2} \exp\left[-\frac{\phi}{2} d(y \,|\, \eta, \phi = 1)\right] \tag{31}$$

4.11 Remark. The general class of models introduced in [Wes85b], assumes a sampling distribution with density proportional to (31). This function can be written as

$$m(y \,|\, \eta, \phi) = \phi^{1/2} \exp\left[\ln \frac{p(y \,|\, \eta, \phi = 1)}{p(y \,|\, \eta_s(y), \phi = 1)}\right]^{\phi} = \frac{\phi^{1/2}}{p(y \,|\, \eta_{s(y)}, \phi = 1)^{\phi}} p(y \,|\, \eta, \phi = 1)^{\phi}$$

Therefore, it fulfils the following properties (compare with [Wes85b, Ch. 2]):

- As a function of η it is proportional to $p(y \,|\, \eta, \phi = 1)^{\phi}$.
- For $\phi = 1$ it is proportional to the exponential family density $p(y \,|\, \eta, \phi = 1)$.
- Since $p(y \,|\, \eta, \phi = 1)^{\phi} \propto \exp\left[\phi\{y\eta - a(\eta)\}\right]$ as a function of η, it fits the definition of exponential family distribution (compare with Definition 1.4).

Based on these properties, this model is suitable for evaluating whether the standard model with $\phi = 1$ is appropriate.

4.12 Remark. In [Wes85b] a gamma prior $\mathcal{G}(\vartheta_{t-1}, \delta_{t-1})$ was used for ϕ in the general model. Using the conjugate prior for η from Lemma 1.6 does in general not lead to a gamma posterior for ϕ. For example in case of a Poisson model, the normalization term of the conjugate prior density involves gamma functions in ϕ. Then of course, the posterior density $p(\eta_t, \phi) \propto p(\eta_t \,|\, \phi, D_{t-1})$ is of no convenient form (compare with [Wes85b, Ch 2.3]). To obtain a posterior distribution for ϕ that is again a gamma distribution, the conjugate prior for η_t is approximated by $p(\eta_t \,|\, \phi, D_{t-1}) \propto \phi^{1/2} \exp\left[-\frac{\phi}{2} d(\eta_t \,|\, \phi = 1, D_{t-1})\right]$. In [Wes85b, Ch. 2.3] more details on how and when this approximation is appropriate are given.

4.13 Lemma. Let $\phi \sim \mathcal{G}(\vartheta, \delta)$ and $(\eta \,|\, \phi)$ have a distribution with density proportional to $\phi^{1/2} \exp\left[-\frac{\phi}{2} d(\eta \,|\, \phi = 1)\right]$. For an observed y from a sampling distribution with density proportional to (31), the posterior distribution of ϕ is a $\mathcal{G}\left(\vartheta + \frac{1}{2}, \delta + \frac{1}{2} d(y \,|\, \phi = 1)\right)$-distribution and the posterior distribution of η has a density proportional to the function $\phi^{1/2} \exp\left[-\frac{\phi}{2} d(\eta \,|\, y, \phi = 1)\right]$.

Proof. There holds

$$\begin{aligned}
p(\eta, \phi \,|\, y) &\propto m(y \,|\, \eta, \phi) p(\eta \,|\, \phi) p(\phi) \\
&= \phi^{1/2} \exp\left[-\frac{\phi}{2} d(y_t \,|\, \eta, \phi = 1)\right] \cdot \phi^{1/2} \exp\left[-\frac{\phi}{2} d(\eta \,|, \phi = 1)\right] \cdot \phi^{\vartheta - 1} \exp\left[-\phi\delta\right] \\
&= \phi^{1/2} \exp\left[-\frac{\phi}{2} \left(d(y \,|\, \eta, \phi = 1) + d(\eta \,|, \phi = 1)\right)\right] \phi^{\vartheta + \frac{1}{2} - 1} \exp\left[-\phi\delta\right] \\
&\overset{(30)}{=} \phi^{1/2} \exp\left[-\frac{\phi}{2} \left(d(y \,|\, \phi = 1) + d(\eta \,|\, y, \phi = 1)\right)\right] \\
&= \phi^{1/2} \exp\left[-\frac{\phi}{2} d(\eta \,|\, y, \phi = 1)\right] \cdot \phi^{\vartheta + \frac{1}{2} - 1} \exp\left[-\phi\left(\delta + \frac{1}{2} d(y \,|\, \phi = 1)\right)\right].
\end{aligned}$$

□

4.14 Remark. To evaluate whether the standard model with $\phi = 1$ is a good fit, the DGLM is now extended in the following way. The analysis at time $t \in \mathbb{N}$ is carried out as described in Chapter 3 with ϕ set to one (see [Wes85a, Ch 4.2]). Additionally, the scale parameter is treated as a random variable with $(\phi \mid D_0) \sim \mathcal{G}(\vartheta_0, \delta_0)$ for some prior parameters ϑ_0, δ_0 and $(\phi \mid D_t) \sim \mathcal{G}(\vartheta_t, \delta_t)$. The parameters ϑ_t, δ_t are calculated at each step as in Lemma 4.13, where the prior distribution for ϕ is $(\phi \mid D_{t-1})$ and the deviance functions are based on the distributions $(\eta_t \mid D_{t-1})$ and $(\eta_t \mid D_t)$ calculated beforehand. At each step, the distribution $(\phi \mid D_t)$ is evaluated to inspect whether the model with $\phi = 1$ is suitable.

Finally, we summarize some of the ideas from [Wes85a, Ch. 4.3] and [Wes85b, Ch. 3] concerning outlier modelling. Let's consider the DGLM at time $t \in \mathbb{N}$. For the conjugate prior $\mathrm{CP}(\alpha_t, \beta_t)$ and an observed value y_t of Y_t, the parameters of the posterior distribution $(\eta_t \mid D_t)$ are (compare with Lemma 1.6) $\alpha_t + \phi y_t$ and $\beta_t + \phi$. So if y_t is an outlier, the first parameter of the posterior distribution is strongly affected by it. This leads to a noticeable change in the posterior mode and thus the posterior distribution favoring more extreme values (compare with [Wes85b, Ch. 3]). To evaluate the sensitivity of the chosen model to outliers, the *influence function* $g(y_t \mid \eta_t) = \frac{\partial}{\partial \eta_t} \ln p(y_t \mid \eta_t)$ is introduced. The effect of observations of Y_t on the posterior can be seen in the following equation derived from the Bayes' rule (see [Wes85b, Ch. 3.2])

$$\frac{\partial}{\partial \eta_t} p(\eta_t \mid y_t) = \frac{\partial}{\partial \eta_t} p(\eta_t) + g(y_t \mid \eta_t).$$

So if the influence of outliers on the posterior should be limited, the function $g(y_t \mid \eta_t)$ needs to possess certain properties such as boundedness. In [Wes85b, Ch. 3] such properties are guaranteed by using the function (31).

5 Applications

5.1 Example. In [Wes85a] the binomial model described in Example 3.18 is applied to advertising data. We have extracted the data presented in [Wes85a, Figure 3 (a)] with the help of WebPlotDigitizer ([Roh21]), which might have lead to some slight inaccuracies in the data used here. The data comes from a weekly survey of $k = 66$ people. Weekly counts of the number of people, who responded positively to a question concerning the advertisement of a chocolate bar, and a measure of advertising called *adstock* are provided. The adstock in week t combines the effect of advertising from week t and advertising from previous weeks, where the diminishing effect of previous advertising over time is taken into account (for more information, see [Mig85]).

In our model, y_t are the weekly counts of positive responses and $\boldsymbol{x_t} = (1, \hat{x}_t)$, where $\hat{x}_t$ are the weekly adstock values for $t \in \{1, ..., 170\}$. Following the example from [Wes85a], the parameters are set to $G_t = U_2$, $B_t = \mathrm{diag}(0.95^{-1/2}, 0.9^{-1/2})$ for $t \in \{1, ..., 170\} \setminus \{96, 156\}$ and $B_{96} = B_{156} = 0.1^{-1/2} \cdot U_2$. The discount parameters are changed at $t \in \{96, 156\}$ to reflect increased uncertainty. This is done due to a change in the advertising strategy at those times. We tested the following values for the moments of $(\theta_0 \mid D_0)$

$$m_{0(1)} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}, m_{0(2)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
$$C_{0(1)} = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix},\ C_{0(2)} = \begin{bmatrix} 100^2 & 0 \\ 0 & 100^2 \end{bmatrix}.$$

By implementing the steps in Remark 3.16 in R with α_t and β_t chosen as in Example 3.18, the posterior mean $\boldsymbol{m_t} = (m_{t1}, m_{t2})^T$ and posterior covariance matrix $C_t \in \mathbb{R}^2$ for $\boldsymbol{\theta_t} = (\theta_{t1}, \theta_{t2})^T$ as well as the means of the prior predictive distribution for Y_t for all $t \in \{1, ..., 170\}$ were obtained. In Figure 1, posterior means for m_{t2} for θ_{t2}, which is the coefficient for the adstock, and the standard deviation values (obtained by taking the square root of C_{t22}) are displayed.

Comparing the results of the two models, we can see that the different prior values produce slightly different results in the beginning. In the later steps, the this difference is not noticeable anymore. This is an expected result, as with time the observed data contributes more to the prior values than the original priors.

As a final step, we test whether the DGLM is more suited for the time dependent data at hand than the GLM. For this, for each $t \in \{2, ..., 170\}$ we compare the predictions made by the DGLM and GLM to the actual values. As predictions for y_t, the mean of the prior predictive distribution at time $t \in \mathbb{N}$ obtained from the DGLM analysis (compare with Example 3.18) and the prediction obtained by fitting a GLM in R based on the data up until to time $t-1$ are chosen and denoted by $\hat{y}_t$ and $\tilde{y}_t$. To compare the precision of the predictions made by the different models, for each $t \in \{2, ..., 170\}$ the squared errors $(y_t - \hat{y}_t)^2$ and $(y_t - \tilde{y}_t)^2$ are calculated. The first and the third quartile as well as the mean and median of the squared errors over all $t \in \{2, ..., 170\}$ for the two models are displayed in Table 2. The errors for the DGLM are noticeably smaller from which we can conclude that the DGLM is better at predicting the values for this particular data set than GLM.

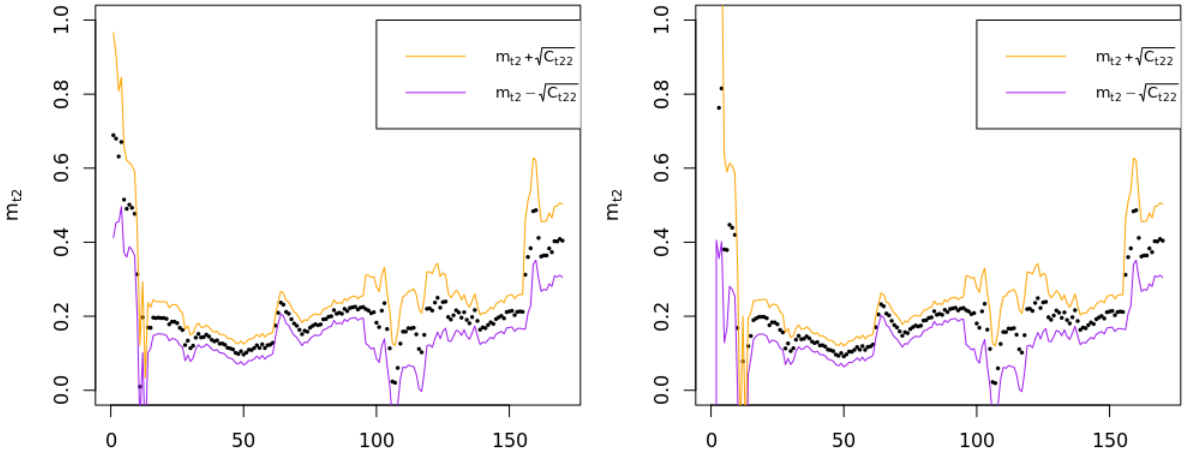

Figure 1: Second components of the posterior means m_t and the differences to the standard deviations $\sqrt{C_{t22}}$ for the starting values $m_{0(1)}, C_{0(1)}$ (on the left) and $m_{0(2)}, C_{0(2)}$ (on the right).

Model	first quartile	median	mean	third quartile
GLM	3.11	14.41	77.74	60.76
DGLM	1.9460	10.0411	49.7881	37.8955

Table 1: Summary of the squared errors of the predictions made by the GLM and DGLM.

5.2 Example. In [Wes85a, Example 4], the binomial DGLM from Example 3.18 is applied to a data set where the observations are not time dependent and these are compared to the results obtained by fitting a static GLM. The authors were able to produce similar results with the two models.The data, consisting of $n = 39$ observations, originally stem from [Fin47] and a GLM was also applied to this data set in [Pre81]. The data was obtained from a study of the effect of the rate and volume of air in skin on the occurrence of vasoconstriction (narrowing of the blood vessels). The data were obtained by multiple measurements from three different subjects. In this case, the response, which describes the occurence or non-occurence of vasoconstriction, is binary, so the sampling distribution of Y_t is a $\mathcal{B}(k, \mu_t)$-distribution for $k = 1$. For the vector $\boldsymbol{x_t} = (1, \hat{x}_{t1}, \hat{x}_{t2})^T$, the values $\hat{x}_{t1}$ and $\hat{x}_{t2}$ are obtained by transforming the rate and the volume variables from the data set by the natural logarithm.

The prior values were set to

$$m_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \; C_0 = 100^2 \cdot U_3$$

and $G_t = B_t = U_3$ for all $t \in \{1, ..., 39\}$. The choice of G_t reflects the time independence of $\boldsymbol{\theta}$ and the prior values lead to a relatively uninformative prior i.e. no particular values are favoured. This particular example leads to overflow in R when calculating exponentials in equations 29, so the results from [Wes85a] can't be reproduced here. Instead, we consider a simulated data set in the next example to test whether the DGLM is appropriate for time dependent data.

5.3 Example. To create similar conditions to Example 5.2, for each $t \in \{1, ..., 300\}$ we have sampled an observation y_t from $\mathcal{B}(k, \mu_t)$, where $\mu_t = \frac{\exp(x_t^T \theta_t)}{1+\exp(x_t^T \theta_t)}$ for

$$\theta_t = \begin{cases} (-1, 0.1)^T & , \; 1 \leq t \leq 100 \\ (-1.5, 0.3)^T & , \; 101 \leq t \leq 200 \\ (0.3, -0.05)^T & , \; 201 \leq t \leq 300 \end{cases} .$$

For $\boldsymbol{x_t} = (1, \hat{x}_t)$, the $\hat{x}_t$ are chosen as equidistant points in the interval $[1, 20]$ for $1 \leq t \leq 100$, $[3, 21]$ for $101 \leq t \leq 200$ and $[0, 17]$ for $201 \leq t \leq 300$. To avoid the numerical problems that occurred in Example 5.2, the parameter k is set to equal 30. The prior values are set to

$$m_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \; C_0 = 100^2 \cdot U_2$$

and $G_t = B_t = U_2$ for all $t \in \{1, ..., 300\}$.

To test whether the DGLM is also suitable for time independent data, we compare the results from the DGLM analysis to those produced by fitting a static GLM in R. The results obtained by fitting a GLM are independent of the ordering of data. To see whether the DGLM also possess the same property, the DGLM analysis is also conducted on the data once in randomly permuted order and once in the reversed order. In Table 5.3 the resulting coefficient estimates and the corresponding standard errors for the different models are displayed. As coefficient estimates for $\boldsymbol{\theta} = (\theta_1, \theta_2)$, the entries of the posterior means $\boldsymbol{m_{300}}$ resulting from the DGLM analyses are chosen. The corresponding standard errors are taken to be the square roots of the diagonal entries of the posterior covariance matrix C_{300}. Here, we can observe that the results produced by the DGLM are very similar to those produced by the GLM and permuting the data has no major effect on the results.

Model	**estimate for θ_1 (st. error)**	**estimate for θ_2 (st. error)**
GLM	-0.563 (0.047)	0.093 (0.004)
DGLM	-0.516 (0.048)	0.087 (0.004)
DGLM (reverse order)	-0.554 (0.047)	0.093 (0.004)
DGLM (random order)	-0.586 (0.048)	0.097 (0.004)

Table 2: Estimates (rounded to three decimal places) produced by the GLM and DGLM analyses.

Conclusion

The dynamic generalized linear model provides a flexible tool to model time dependent data. The model is better at capturing structural changes in the data and the Bayesian approach allows for incorporation of subjective information into the model. In Example 5.1 knowledge about changes in the data could be incorporated by changing the discount factors. The dynamic model outperformed the static GLM at predicting the values of the data set. The dynamic model is also adaptable to cases where the data is not time dependent as illustrated in Example 5.3.

On the flip side, the flexibility also leads to more ambiguity. The theory of dynamic models is not yet able to capture particularities of the different sampling distributions. In Example 3.18 the original approach was partly abandoned for some more ad hoc methods to avoid computational difficulties. In Example 5.2 numerical problems occurred and the original results from [Wes85a] couldn't be reproduced.

References

[Abr72] M. Abramowitz, I. A. Stegun (eds.): *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Applied Mathematics Series. Vol. 55, Tenth Printing, Dover Publications, 1972.

[Ame85] J. R. M. Ameen, P. J. Harrison: *Normal Discount Bayesian Models.* Bayesian Statistics 2, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smirth, Elsevier Science Publishers B.V., North-Holland, pp. 271-298, 1985.

[Ata08] V. Atanasiu: *Mathematical models in regression credibility theory.* BASM, No 3(58), pp. 18-33, 2008.

[Ber94] J. M. Bernardo, A. F. Smith: *Bayesian Theory.* John Wiley and Sons, 1994.

[Bli14] J. K. Blitzstein, J, Hwang: *Introduction to Probability.* Chapman and Hall, 2014.

[Fin47] D. J. Finney: *The estimation from individual records of the relationship between dose and quantal response.* Biometrika Vol. 34, No. 3/4, pp. 320-334, 1947.

[Gel04] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin: *Bayesian Data Analysis.* Second Edition, Chapman and Hall, 2004.

[Gol07] M. Goldstein, D. Wooff: *Bayes Linear Statistics Theory and Methods.* Wiley, 2007.

[Hof09] P. D. Hoff: *A First Course in Bayesian Statistical Methods.* Springer-Verlag New York, 2009.

[Lin97] J. K. Lindsey: *Applying Generalized Linear Models.* Springer-Verlag New York, Inc., 1997.

[McN89] P. McCullagh, J.A. Nelder : *Generalized Linear Models.* Second edition, Chapman and Hall London New York, 1989.

[Mig85] H. S. Migon, P. J. Harrison: *An Application of Non-linear Bayesian Forecasting to Television Advertising.* Bayesian Statistics 2, J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith (eds.), Elsevier Science Publishers B.V., North-Holland, pp. 681-696, 1985.

[Nel72] J. A. Nelder, R. W. M. Wedderburn: *Generalized Linear Models.* Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3, pp. 370-384, 1972.

[Pre81] D. Pregibon: *Logistic Regression Diagnostics.* The Annals of Statistics, Vol 9, No. 4, pp. 705-724, 1981.

[Rob94] C. P. Robert: *The Bayesian Choice.* Springer-Verlag New York, Inc, 1994.

[Roh21] A. Rohatgi: *Webplotdigitizer: Version 4.5.* `https://automeris.io/WebPlotDigitizer`, 2021. Accessed 02.09.2022.

[Ross98] S. Ross: *A First Course in Probability.* Fifth Edition, Prentice-Hall, Inc. 1998.

[Wes85a] M. West, P. Jeff Harrison, H. S. Migon: *Dynamic Generalized Linear Models and Bayesian Forecasting*. Journal of the American Statistical Association, Vol. 80, No. 389, pp. 77-83, 1985.

[Wes85b] M. West: *Generalized Linear Models: Scale Parameters, Outlier Accommodation and Prior Distributions*. Bayesian Statistics 2, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith, Elsevier Science Publishers B.V., North-Holland, pp. 531-558, 1985.