FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

ASSIGNMENT OF MASTER’S THESIS

Title:	Sequential Bayesian Poisson regression
Student:	Bc. Radomír Žemlička
Supervisor:	Ing. Kamil Dedecius, Ph.D.
Study Programme:	Informatics
Study Branch:	Knowledge Engineering
Department:	Department of Applied Mathematics
Validity:	Until the end of summer semester 2020/21

Instructions

Regression modelling of counts generally relies on the generalized linear models (GLMs), particularly on the Poisson regression model with the logarithmic link function. In the Bayesian realm, such models are estimated via convenient prior distributions. However, regardless of the functional form of the prior, the posterior distributions are neither standard nor analytically tractable. In his 1973 paper, G.M. El-Sayyad suggests circumventing the intractability issue by means of normal approximations of the Poisson likelihood.
The main points of the thesis are:
1. Overview of the GLMs and the Poisson regression model, focus on El-Sayyad's approach to its Bayesian estimation. Propose a sequential variant.
2. Propose methods for stabilization of the estimation procedure and study their behaviour on convenient examples.
3. If possible, suggest a use case of the proposed modelling approach in the signal processing domain.

References

Will be provided by the supervisor.

Ing. Karel Klouda, Ph.D.
Head of Department

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
Dean

Prague January 9, 2020

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Master's thesis

Sequential Bayesian Poisson Regression

Bc. Radomír Žemlička

Department of Applied Mathematics
Supervisor: Ing. Kamil Dedecius, Ph.D.

May 28, 2020

Acknowledgements

With many thanks to my supervisor and mentor Ing. Kamil Dedecius, Ph.D., for his guidance, valuable time, and endless patience. To my dear wife for her tremendous support.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as a school work under the provisions of Article 60 (1) of the Act.

In Prague on May 28, 2020 …………………

Citation of this thesis

Žemlička, Radomír. *Sequential Bayesian Poisson Regression.* Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2020.

Abstrakt

Poissonovská regrese je populární zobecněný lineární model používaný k modelování diskrétních náhodných veličin, typicky počtů. Tato práce je zaměřena na problematiku jejího sekvenčního odhadování s regresními koeficienty potenciálně pomalu proměnnými v čase. Je použita vhodná aproximace normálním rozdělením, aby tak bylo možné učinit v Bayesovském kontextu. Rovněž je diskutována kalibrační technika pro zvýšení kvality odhadů. Na závěr je navržen případ použití představeného přístupu v doméně zpracování signálu, zejména jeho použití v difuzních sítích (diffusion networks) pro realizaci distribuovaného kolaborativního odhadování.

Klíčová slova Poissonovská regrese, Bayesovská inference, distribuované odhadování, kolaborativní odhadování

Abstract

The Poisson regression is a popular generalized linear model used to model discrete count variables. This thesis is focused on the problem of its sequential estimation under potentially slowly time-varying regression coefficients. A convenient approximation by normal distribution is used to do so in the Bayesian setting. Also, a calibration technique is discussed to enhance the estimation quality. Finally, a use case of the proposed approach in the signal processing domain is suggested, in particular, its application in diffusion networks to perform distributed collaborative estimation.

Keywords Poisson regression, Bayesian inference, distributed estimation, collaborative estimation

Contents

List of Figures

List of Tables

Introduction

This thesis is focused on models of discrete counts used, e.g., to describe epidemiological data, the number of stock market transactions in finance, traffic intensities in networks and transportation, the number of particle arrivals in physics, or phenomena in social networks [1, 2]. High counts can be generally approximated by continuous data models, but those can fail if the counts are small and include many zeros [3]. The thesis is specifically focused on the Poisson regression model. At first, its low-cost real-time sequential estimation is proposed to deal with streaming data. Then a method for its distributed inference in networks of collaborating agents (sensors) is devised as an application in the signal processing domain. The author is not aware of any existing sequential distributed or non-distributed alternative. The known non-distributed Poisson models rely on computationally intensive optimization techniques [4, 5], making their usage in online tasks, such as dealing with streaming data or effective processing of big data, difficult.

Distributed inference of unknown variables in networks of collaborating agents has become an established discipline in the signal processing domain. Its applications may be found in sensor networks, smart grids and microgrids, IoT (Internet of Things), big data, social networks, and other types of networked systems [6, 7, 8, 9].

Generally, three communication and information processing strategies can be distinguished: the incremental strategy, consensus, and diffusion [10, 11]. In this thesis, the point of interest is the diffusion strategy, where the information exchange runs on a single time scale and within one network hop distance [12]. Many popular sequential inference algorithms have found their more or less modified diffusion counterparts. To name only a few: the LMS (least mean squares) [13, 14, 15, 16, 7], RLS (recursive least squares) [12], Kalman filters [17, 18], Bernoulli filters [19], particle filters [20, 21, 22], or the quasi-Bayesian mixture estimation algorithm [23]. A unifying Bayesian framework for diffusion inference of a wide class of models was designed in [24] and [25].

There are several major difficulties. First, the sequential estimation of the

Poisson model is generally impossible due to its functional form and the static estimation requires a numerical optimization method to be used (as described in Section 1.1.1). It is shown that a way towards the solution provides the Bayesian paradigm along with a couple of approximations providing stable and analytically tractable results. There are several novel points in the proposed algorithm. Firstly, the static Bayesian Poisson regression [26] is recasted into an algorithm for *online* estimation of potentially slowly time-varying regression coefficients. Secondly, a rule for the combination of these estimates in diffusion networks is shown.

Some of the key aspects of this thesis, along with the results of a simulated example, have already been presented in [27]. However, in this thesis, the theoretical background is further explored, along with the detailed discussion of alternative methods. The effects of different hyperparameter values and different network configurations on the accuracy of estimation depending on the nature of the data are also examined. Finally, some other topics are mentioned that can be researched in the future as part of the author's postgraduate studies.

The thesis is structured as follows: The problem is described in Chapter 1, where an algorithm for the sequential estimation is also shown. The following Chapter 2 sheds light on the usage of the proposed modeling approach in the signal processing domain, specifically the distributed estimation. In Chapter 3, several sets of simulated examples are presented to demonstrate the efficiency of the proposed method. Finally, Chapter 4 summarizes topics that can be further explored in the future.

Chapter **1**

Sequential Inference of the Poisson Model

We consider a discrete-time modeling of a stochastic process $\{Y_t; t = 0, 1, \ldots\}$ with mutually independent observations $y_t \in \mathbb{N}$. Let Y_t be a random variable that is determined by a known regressor $x_t \in \mathbb{R}^n$, and an unknown vector of possibly slowly time-varying regression coefficients $\beta_t \in \mathbb{R}^n$. The relationship characterizes the GLM (generalized linear model) [4]

$$\mathbb{E}[Y_t|x_t, \beta_t] = g^{-1}(\beta_t^\intercal x_t), \tag{1.1}$$

where $g(\cdot)$ is a known link function. The product $\beta_t^\intercal x_t$ is called the linear predictor. It is useful to note that the identity function $g(\cdot)$ provides the ordinary linear regression model [28]. In the case of $y_t \in \mathbb{N}$, the natural logarithm plays the role of the link function,

$$g(\mathbb{E}[Y_t|x_t, \beta_t]) = \log(\mathbb{E}[Y_t|x_t, \beta_t]) = \beta_t^\intercal x_t, \tag{1.2}$$

resulting in the Poisson regression model

$$Y_t \sim Po(\lambda_t) = Po\left(g^{-1}(\beta_t^\intercal x_t)\right) = Po\left(\exp\left(\beta_t^\intercal x_t\right)\right). \tag{1.3}$$

The expected value and the variance are

$$\mathbb{E}[Y_t|x_t, \beta_t] = \operatorname{var}(Y_t|x_t, \beta_t) = \lambda_t = \exp(\beta_t^\intercal x_t), \tag{1.4}$$

and the pdf (probability density function) of the model reads

$$f(y_t|x_t, \beta_t) = \frac{\lambda_t^{y_t} e^{-\lambda_t}}{y_t!} = \frac{e^{\beta_t^\intercal x_t y_t} e^{-\exp(\beta_t^\intercal x_t)}}{y_t!}. \tag{1.5}$$

Generally, direct Bayesian inference of GLMs is analytically intractable (except for the linear regression model) due to the lack of convenient conjugate prior distributions. Therefore, the inference mostly relies on MCMC

(Markov chain Monte Carlo) methods [5], which are not suitable for real-time sequential analyses. Some GLM-specific lower-complexity workarounds were proposed, e.g., the normal Laplacian approximation of the posterior pdf in logistic regression [29], or the three-stage approximation Poisson → log-gamma → normal pdf in the static Poisson model [26]. The later will be adopted below to propose the sequential Bayesian estimator.

1.1 Statical approach to estimation

As stated earlier, direct sequential inference of the Poisson model is analytically intractable, hence a different approach must be chosen. There are several ways to work around this problem. This section focuses mainly on MLE (maximum likelihood estimation) [30] and Bayesian estimation using approximation of the posterior distribution [26].

1.1.1 Maximum likelihood estimation

From the statistical point of view, a given set of data $y = [y_0, y_1, \ldots, y_{m-1}]$ is a random sample from an unknown population. The aim of maximum likelihood estimation is to find a specific population from which the sample has most likely been drawn. More precisely, every population is mathematically described by a corresponding probability distribution. Every probability distribution is associated with a specific model and its parameters. Our goal is to find a unique vector $\beta = [\beta_0, \beta_1, \ldots, \beta_{n-1}]^\intercal$ containing values of the parameters [30].

Let $f(y_t|\beta)$ be the pdf that specifies the probability of an observation y_t depending on the parameters β. Given a specific vector of parameters, the corresponding pdf will show that some observations are more probable than others. If all individual observations y_t are statistically independent, then the pdf of the joint distribution is just a simple multiplication of individual pdfs, meaning

$$f(y|\beta) = f(y_0, y_1, \ldots, y_{m-1}|\beta) = \prod_{t=0}^{m-1} f(y_t|\beta). \tag{1.6}$$

However, we usually do not know the actual values of the parameters, hence we have to find them. To solve this, we need to define the likelihood function by reversing the roles of the observations y and the parameter vector β, resulting in

$$L(\beta|y) = f(y|\beta). \tag{1.7}$$

The value of the function represents a likelihood of the parameter vector β given the observations y. It is also important to note that there is a significant difference between the two functions $f(y|\beta)$ and $L(\beta|y)$, as they are both defined on different variables [30].

Now, using the previously defined function, our goal is to find a vector $\hat{\beta}$ for which the likelihood function value is the highest,

$$\hat{\beta} = \underset{\forall\beta}{\arg\max}\ L(\beta|y). \tag{1.8}$$

To do this, we can use a variety of numerical optimization methods. In some cases, it is possible to find the maximum analytically using the derivation of the likelihood function (if it exists). Working with this type of formula is generally difficult. However, it is possible to work with the logarithm of the likelihood function,

$$\ell(\beta|y) = \log L(\beta|y). \tag{1.9}$$

Finally, to find the maximum of the function, we need to solve an equation

$$\frac{\partial \ell(\beta|y)}{\partial \beta} = 0, \tag{1.10}$$

and identify its critical points. Since logarithm is a monotonic function, the maximum of $\ell(\beta|y)$ occurs at the same point as does the maximum of $L(\beta|y)$ [30].

Now, let us again get back to the Poisson regression. Suppose that we have a set of m vectors

$$x_t \in \mathbb{R}^n, \quad t = 0, \dots, m-1, \tag{1.11}$$

along with a set of m values

$$y_t \sim Po\left(e^{\beta^\intercal x_t}\right). \tag{1.12}$$

First, we need to define the pdf of the joint distribution of all observations. The observations are independent, therefore using the Equation (1.6), the pdf for all k values y_t can be written as

$$f(y_0, \dots, y_{m-1}|x_0, \dots, x_{m-1}, \beta) = \prod_{t=0}^{m-1} \frac{e^{\beta^\intercal x_t y_t} e^{-\exp(\beta^\intercal x_t)}}{y_t!}. \tag{1.13}$$

Then, to perform MLE and find the values of the vector β, we need to define a likelihood function as shown in the Equation (1.7), resulting in

$$L(\beta|X, Y) = \prod_{t=0}^{m-1} \frac{e^{\beta^\intercal x_t y_t} e^{-\exp(\beta^\intercal x_t)}}{y_t!}. \tag{1.14}$$

As described earlier, it is easier to work with the log-likelihood rather than with the original form (1.14). We can transform the Equation (1.14) using the principle presented in the Equation (1.9), leading us to

$$\ell(\beta|X,Y) = \sum_{t=0}^{m-1} \left(\beta^\intercal x_t y_t - e^{\beta^\intercal x_t} - \log(y_t!)\right). \tag{1.15}$$

Finally, to find a maximum of (1.15), we need to solve the equation

$$\frac{\partial \ell(\beta|X,Y)}{\partial \beta} = 0. \tag{1.16}$$

The main problem is that the equation has no closed-form solution. However, the negative log-likelihood, $-\ell(\beta|X,Y)$, is a convex function, therefore a convex optimization approach, such as gradient descent, can be used to find the optimal value of $\hat{\beta}$.

1.1.2 Bayesian estimation

There is also a different approach that does not rely on iterative optimization algorithms. In [26], G. M. El-Sayyad describes a method to analytically estimate parameters β in a Bayesian setting using an approximation based on the work of M. S. Bartlett and D. G. Kendall [31].

Let y and β be random variables with pdfs $f(y|\beta)$ and $\pi(\beta)$, respectively. Then, according to Bayes' theorem

$$\pi(\beta|y) = \frac{f(y|\beta)\pi(\beta)}{f(y)}, \tag{1.17}$$

where $\pi(\beta|y)$ is a posterior density of β, $\pi(\beta)$ is a prior density of β, $f(y|\beta)$ is a likelihood of observations, and $f(y)$ is a marginal density of observations. The marginal density serves as a normalizing constant and due to this fact we often write only a proportionality

$$\pi(\beta|y) \propto f(y|\beta)\pi(\beta). \tag{1.18}$$

As a result of this we get a posterior distribution of β.

We can now now apply Bayes' theorem to the Poisson regression problem. As already mentioned in the previous section, the likelihood of observations can be written as

$$f(y_0,\ldots,y_{m-1}|x_0,\ldots,x_{m-1},\beta) = \prod_{t=0}^{m-1} \frac{e^{\beta^\intercal x_t y_t} e^{-\exp(\beta^\intercal x_t)}}{y_t!}. \tag{1.19}$$

Choosing a reasonably vague prior [26], the posterior distribution reads as

$$\pi(\beta|y_0,\ldots,y_{m-1},x_0,\ldots,x_{m-1}) \propto f(y_0,\ldots,y_{m-1}|x_0,\ldots,x_{m-1},\beta). \tag{1.20}$$

If both the posterior distribution $\pi(\beta|y)$ and the prior distribution $\pi(\beta)$ are from the same probability distribution family, then they are called conjugate distributions [32]. Unfortunately, there is no conjugate prior for β due to the functional form of the likelihood function of the Poisson GLM. This, however, can be solved if we apply an approximation to the distribution of our observations.

Suppose that we have a random variable

$$Z \sim \Gamma(k, \theta), \tag{1.21}$$

where a corresponding pdf reads

$$f(z|k, \theta) = \frac{1}{\Gamma(k)\theta^k} z^{k-1} e^{-\frac{z}{\theta}}. \tag{1.22}$$

If we transform the said random variable in the sense

$$\tilde{Z} = \log Z \sim \log \Gamma(k, \theta), \tag{1.23}$$

we get a distribution whose pdf reads

$$f(\tilde{z}|k, \theta) = \frac{1}{\Gamma(k)\theta^k} e^{\tilde{z}k} e^{-\frac{\exp(\tilde{z})}{\theta}}. \tag{1.24}$$

According to [26] and [31], if the value of k is large and $\theta = 1$, then the distribution of $\tilde{Z}$ can be approximated as

$$\tilde{Z} \sim \mathcal{N}(\log k, k^{-1}). \tag{1.25}$$

If we take a closer look at Equation (1.19), we can see a familiar pattern in the formula. By using the previously described approximation, we can rewrite the pdf of the observations as

$$\begin{aligned} f(y_0, \ldots, y_{m-1}|x_0, \ldots, x_{m-1}, \beta) &= \prod_{t=0}^{m-1} \frac{e^{\beta_t^\intercal x_t y_t} e^{-\exp(\beta_t^\intercal x_t)}}{y_t!} && (1.26) \\ &= \prod_{t=0}^{m-1} \frac{1}{y_t} \frac{1}{\Gamma(y_t) 1^{y_t}} e^{\beta_t^\intercal x_t y_t} e^{-\frac{\exp(\beta_t^\intercal x_t)}{1}} && (1.27) \\ &\approx \prod_{t=0}^{m-1} \frac{1}{y_t} \mathcal{N}(\beta_t^\intercal x_t | \log y_t, y_t^{-1}) && (1.28) \\ &\propto \exp\left(-\frac{1}{2} \sum_{t=0}^{m-1} y_t (\beta^\intercal x_t - \log y_t)^2 \right). && (1.29) \end{aligned}$$

The pdf of the posterior distribution of β can be then approximated as

$$\pi(\beta|y_0, \ldots, y_{m-1}, x_0, \ldots, x_{m-1}) \propto \exp\left(-\frac{1}{2} \sum_{t=0}^{m-1} y_t (\beta^\intercal x_t - \log y_t)^2 \right). \tag{1.30}$$

Finally, we need to find the expected value of the posterior distribution. According to [26], by setting a vector

$$t = \left[\sqrt{y_0} \log y_0 \quad \cdots \quad \sqrt{y_{m-1}} \log y_{m-1}\right]^\intercal, \tag{1.31}$$

and a matrix

$$U = \begin{bmatrix} x_{0,0}\sqrt{y_0} & \cdots & x_{0,n-1}\sqrt{y_0} \\ \vdots & \ddots & \\ x_{m-1,0}\sqrt{y_{m-1}} & & x_{m-1,n-1}\sqrt{y_{m-1}} \end{bmatrix}, \tag{1.32}$$

we can obtain the posterior expected value of β as

$$\mathbb{E}(\beta) = (U^\intercal U)^{-1} U^\intercal t. \tag{1.33}$$

As we can see in Equation (1.33), the expected value is obtained using only a simple formula, which is computationally more efficient than numerical methods used with MLE. That, however, can have a negative impact on accuracy of resulting estimates.

To compare both methods, MLE and Bayesian estimation using the previously described approximation, in terms of accuracy of their estimates, we can use a simple simulation where we incrementally increase the data size. Fig. 1.1 shows estimates of parameters β for different data sizes. In both cases, the size was incrementally increased by 10. Real value of the parameters $\beta = [0.9, 0.6, 0.3, 0.1]^\intercal$. Fig. 1.2 shows evolution of the RMSE (root mean square error)

$$\begin{aligned} RMSE_t(\beta_i) &= \sqrt{\frac{1}{t+1}\sum_{\tau=0}^{t}(\hat{\beta}_{i,\tau} - \beta_i)^2} \\ i &= 0, \dots, n-1, \\ t &= 0, \dots, m-1, \end{aligned} \tag{1.34}$$

for both methods. As we can see, the second method is less accurate.

1.2 Stabilization of variance

A commonly encountered problem of the Poisson models evident from (1.4) is that the variance is equal to the expected value. Suppose that we have a Poisson variable

$$Y_t \sim Po(\lambda_t). \tag{1.35}$$

It is generally known that for large values of λ_t, it approaches the normal distribution with both the expected value and the variance equal to λ_t [33].

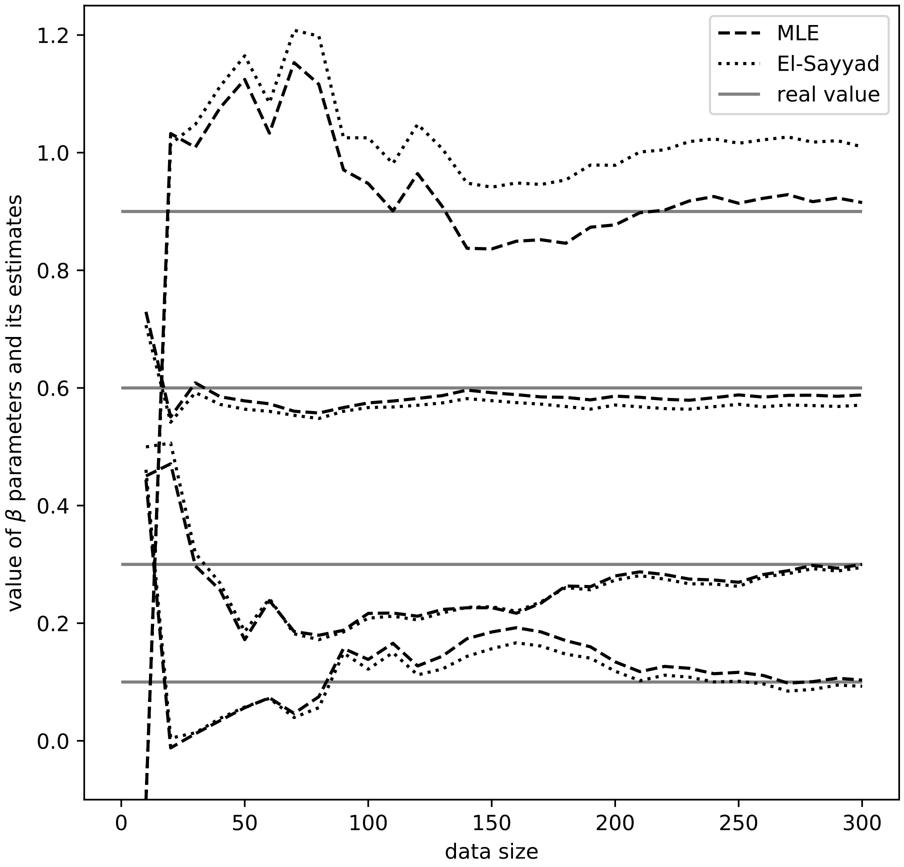

Figure 1.1: Real and estimated values of β for different sizes of data. The solid line depicts the real value of parameters, the dashed line represents estimates of MLE, and the dotted line represents estimates of the Bayesian method using the approximation presented by El-Sayyad [26].

Hence, a convenient function $h(\cdot)$, serving as a variance-stabilizing transformation, may be used to improve the estimation quality. This function can take many forms with varying accuracy and ease of implementation. Lists of these transformations are summarized, e.g., in [4] or [33].

As shown in [34] and [33], if we standardize the Poisson variable, meaning

$$g(y_t) = \frac{y_t - \lambda_t}{\sqrt{\lambda_t}}, \tag{1.36}$$

then for large λ_t, it has approximately standard normal distribution with mean $\mathbb{E}(\widetilde{Y}_t) = 0$ and variance $var(\widetilde{Y}_t) = 1$.

One of the simplest of these transformations is the square-root transformation $h(y_t) = \sqrt{y_t}$ [35, 31]. When we apply the transformation to the said

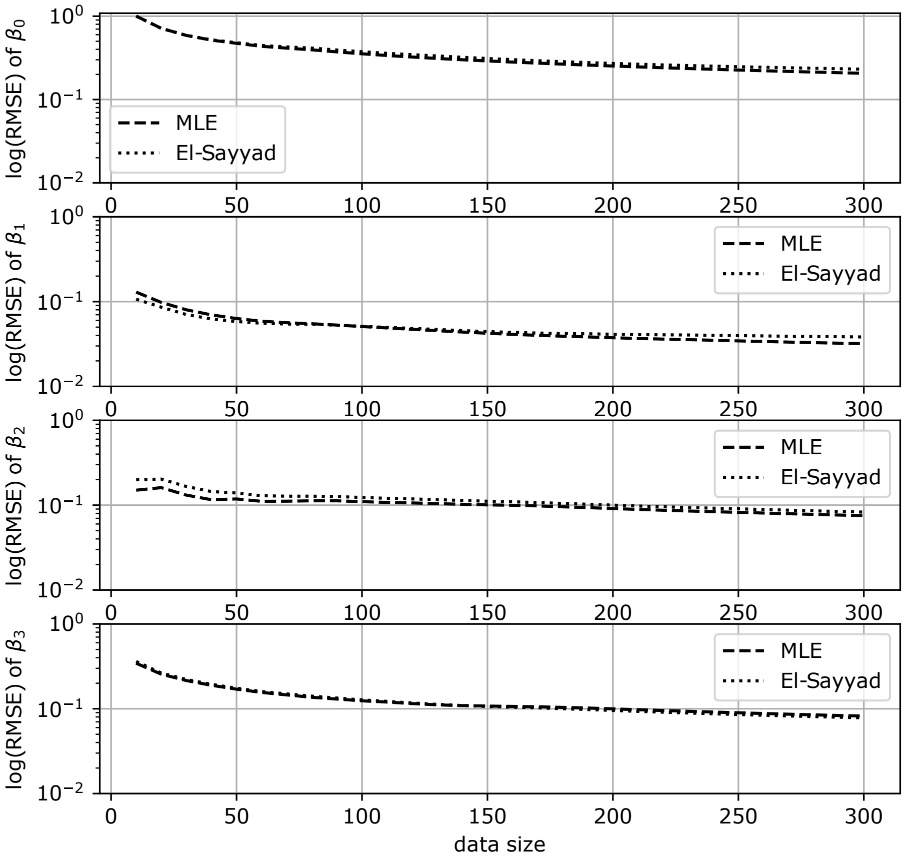

Figure 1.2: Evolution of the RMSE. The dashed line represents RMSE of MLE estimates, and the dotted line represents RMSE of estimates of the El-Sayyad's method [26].

variable, then the transformed variable is approximately normally distributed [4],

$$\widetilde{Y}_t = \sqrt{Y_t} \sim \mathcal{N}\left(\sqrt{\lambda_t}, \frac{1}{4}\right), \tag{1.37}$$

and the error term is $O(\lambda_t^{-1})$. As also stated in [4], the variance is only approximately constant. To be more precise, examining the asymptotic expansion, the subsequent terms actually read as

$$\mathbb{E}(\widetilde{Y}_t) \approx \sqrt{\lambda_t}\left(1 - \frac{1}{8\lambda_t}\right), \tag{1.38}$$

$$var(\widetilde{Y}_t) \approx \frac{1 + \frac{3}{8\lambda_t}}{4}. \tag{1.39}$$

There is another power transformation, $Y^{\frac{2}{3}}$, meaning $h(y_t) = y_t^{\frac{2}{3}}$. As stated in [4], the transformed variable is more symmetric than the previous one, with the skewness being $O(\lambda_t^{-\frac{3}{2}})$. The important feature of both this and the previous transformation is that none of them requires any direct knowledge about the actual value of λ_t.

In [4], there is another alternative transformation that provides approximate symmetry and also stability of variance. It requires knowledge about λ_t and is described as

$$h(y_t) = \begin{cases} 3y_t^{\frac{1}{2}} - 3y_t^{\frac{1}{6}}\lambda_t^{\frac{1}{3}} + \frac{1}{6}\lambda_t^{-\frac{1}{2}} & \text{if } y_t \neq 0, \\ -(2\lambda_t)^{\frac{1}{2}} + \frac{1}{6}\lambda_t^{-\frac{1}{2}} & \text{if } y_t = 0. \end{cases} \tag{1.40}$$

The resulting transformed variable is standard normal for large values of λ_t.

In [36], Bartlett also discusses a logarithmic transformation for specific cases where, even after applying the square-root transformation, the variance is still slightly correlated to the mean. The transformation reads

$$h(y_t) = \lambda_t^{-1} \sinh^{-1}\left(\lambda_t \sqrt{y_t}\right), \tag{1.41}$$

or equivalently

$$h(y_t) = \lambda_t^{-1} \log\left(\sqrt{1 + \lambda_t^2 y_t} + \lambda_t \sqrt{y_t}\right). \tag{1.42}$$

The transformation is more exact due to the fact that it requires a good estimation of λ_t. For cases, where this is not possible, it is suggested to use a transformation

$$h(y_t) = \log(y_t + 1), \tag{1.43}$$

which can deal with zeros and also provides good results [36].

Figure 1.3 shows a comparison of different transformations (namely $h(y) = \sqrt{y}$ and $h(y) = y^{\frac{2}{3}}$) and their convergence to the normal distribution. Each row shows a histogram of a random sample generated from the Poisson distribution for a specific values of λ. Shown is also a density estimated using the Gaussian KDE (kernel density estimation) [37] method along with the approximated normal distribution. Table 1.1 then shows a comparison of the real mean and variance and the sample mean and variance. As we can see from both the plots and the table, the square-root transform deviates only slightly and converges relatively fast.

A convenient transformation of the Poisson variable can improve the modeling quality. We will stick with the square-root transformations for its general simplicity and effectiveness.

	$h(y) = y$		$h(y) = \sqrt{y}$		$h(y) = y^{\frac{2}{3}}$	
	$\mathbb{E}(Y)$	$var(Y)$	$\mathbb{E}(h(Y))$	$var(h(Y))$	$\mathbb{E}(h(Y))$	$var(h(Y))$
	$\bar{y}$	s^2	$\overline{h(y)}$	s^2	$\overline{h(y)}$	s^2
$\lambda = 1$	1.0000	1.0000	1.0000	0.2500	1.0000	0.4444
	1.0006	0.9999	0.7735	0.4024	0.8374	0.5266
$\lambda = 2$	2.0000	2.0000	1.4142	0.2500	1.5874	0.5600
	1.9945	2.0038	1.2652	0.3937	1.4612	0.6675
$\lambda = 3$	3.0000	3.0000	1.7321	0.2500	2.0801	0.6410
	2.9981	3.0113	1.6301	0.3408	1.9829	0.7218
$\lambda = 4$	4.0000	4.0000	2.0000	0.2500	2.5198	0.7055
	4.0073	4.0006	1.9237	0.3065	2.4417	0.7627
$\lambda = 5$	5.0000	5.0000	2.2361	0.2500	2.9240	0.7600
	5.0030	5.0049	2.1718	0.2863	2.8531	0.8021
$\lambda = 10$	10.0000	10.0000	3.1623	0.2500	4.6416	0.9575
	10.0113	10.0212	3.1224	0.2621	4.5914	0.9788
$\lambda = 20$	20.0000	20.0000	4.4721	0.2500	7.3681	1.2064
	19.9984	19.8296	4.4436	0.2531	7.3263	1.2075
$\lambda = 30$	30.0000	30.0000	5.4772	0.2500	9.6549	1.3810
	29.9809	30.0480	5.4522	0.2542	9.6145	1.3930
$\lambda = 40$	40.0000	40.0000	6.3246	0.2500	11.6961	1.5200
	40.0256	39.8460	6.3067	0.2513	11.6684	1.5199
$\lambda = 50$	50.0000	50.0000	7.0711	0.2500	13.5721	1.6373
	50.0033	50.1447	7.0534	0.2524	13.5423	1.6467

Table 1.1: Comparison of means and variances of different transformations for specific values of λ. Shown are the real mean ($\mathbb{E}(Y)$) and the real variance ($var(Y)$) along with the sampled mean ($\bar{y}$) and the sampled variance (s^2).

Due to the change of variables theorem, the pdf of $\widetilde{Y}_t$ is

$$\begin{aligned} f(\widetilde{y}_t|x_t,\beta_t) &= f\left(h^{-1}(\widetilde{y}_t)\right)\left|\frac{\mathrm{d}h^{-1}(\widetilde{y}_t)}{\mathrm{d}\widetilde{y}_t}\right| && (1.44)\\ &= \frac{\lambda_t^{\widetilde{y}_t^2}e^{-\lambda_t}}{\widetilde{y}_t^2!}\cdot 2\widetilde{y}_t && (1.45)\\ &= \frac{e^{\beta_t^\intercal x_t\widetilde{y}_t^2}e^{-\exp(\beta_t^\intercal x_t)}}{\widetilde{y}_t^2!}\cdot 2\widetilde{y}_t && (1.46)\\ &= \frac{2}{\Gamma(\widetilde{y}_t^2)}e^{\beta_t^\intercal x_t\widetilde{y}_t^2}e^{-\exp(\beta_t^\intercal x_t)}\cdot\frac{1}{\widetilde{y}_t}, && (1.47)\end{aligned}$$

where the gamma function follows from $z! = z\Gamma(z)$.

Although the square-root transformations stabilizes the properties of the modeled random variable, the problem of nonexistent conjugate prior still persists. Therefore, instead of using the normal pdf, we will stick with the functional form (1.47) in the following steps.

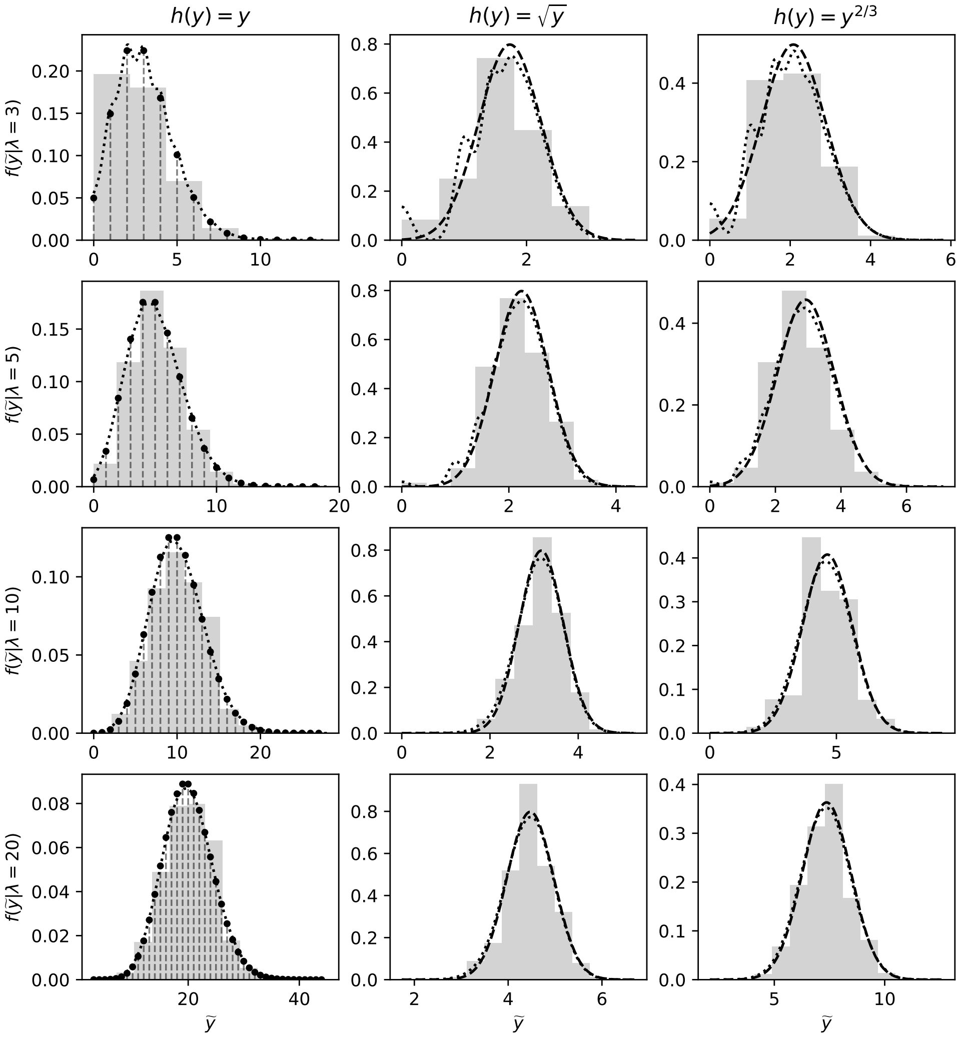

Figure 1.3: Comparison of different transformations and their convergence to the normal distribution for different values of λ. The first column shows a histogram of a sample generated from the Poisson distribution along with the real density of the Poisson distribution and KDE of the sample. The second column shows the transformation $h(y) = \sqrt{y}$ along with KDE and pdf of the corresponding normal distribution. The third column shows the same for the transformation $h(y) = y^{\frac{2}{3}}$. Grey bars represent a histogram of the random sample. Dotted lines represent a KDE density. Dashed lines in the first column represent the density of the Poisson distribution, while in the second and third columns they represent the pdf of the corresponding normal distribution.

1.3 Approximate sequential estimation of β

Let us introduce the notation:

$$x_{0:t-1} = [x_0, \dots, x_{t-1}], \tag{1.48}$$

$$\widetilde{y}_{0:t-1} = [\widetilde{y}_0, \dots, \widetilde{y}_{t-1}]. \tag{1.49}$$

The prior pdf $\pi(\beta_t|x_{0:t-1}, \widetilde{y}_{0:t-1})$ contains all available statistical information about the past observations and regressors necessary for the estimation of β_t. The initial variables $\widetilde{y}_0$ and x_0 represent the prior knowledge (pseudo-observations) available at the very beginning of the modeling, e.g., given by an expert, or obtained from historical observations.

The update of the prior distribution of β_t by recently observed y_t and x_t provides the Bayes' theorem

$$\begin{aligned}\pi(\beta_t|x_{0:t}, \widetilde{y}_{0:t}) &= \frac{f(\widetilde{y}_t|x_t, \beta_t)\pi(\beta_t|x_{0:t-1}, \widetilde{y}_{0:t-1})}{f(x_{0:t}, \widetilde{y}_{0:t})} \\ &= \frac{f(\widetilde{y}_t|x_t, \beta_t)\pi(\beta_t|x_{0:t-1}, \widetilde{y}_{0:t-1})}{\int_{\mathbb{R}^n} f(\widetilde{y}_t|x_t, \beta_t)\pi(\beta_t|x_{0:t-1}, \widetilde{y}_{0:t-1}) d\beta_t},\end{aligned} \tag{1.50}$$

where the integral in the denominator serves as the normalizing constant, ensuring that the resulting posterior function is a pdf that integrates to one.

1.3.1 Sequential estimation with conjugate prior

The Bayesian update (1.50) does not generally yield posterior distributions in closed forms. An important exception is the case of models that belong to the exponential family of distributions estimated with conjugate prior distributions [32].

Definition 1.1. (Exponential family): Let $\widetilde{Y}_t$ be a random variable with a parameter β_t. The distribution of $\widetilde{Y}_t$ belongs to the exponential family if its pdf has a form

$$f(\widetilde{y}_t|x_t, \beta_t) = k(x_t, \widetilde{y}_t) l(\beta_t) e^{\eta(\beta_t)^\intercal T(x_t, \widetilde{y}_t)}, \tag{1.51}$$

where $\eta(\beta_t)$ is the natural parameter, i.e., a function of the original parameter β_t, and $T(x_t, \widetilde{y}_t)$ is a sufficient statistic that comprises all information necessary for the estimation of β_t. The functions $k(x_t, \widetilde{y}_t)$ and $l(\beta_t)$ are the base measure and the normalizing function, respectively.

Definition 1.2. (Conjugate prior distribution): The prior distribution for the estimation of β_t conjugate to the model (1.51) is characterized by the prior hyperparameters Ξ_t of the same size as $T(x_t, \widetilde{y}_t)$, and a scalar positive ν_t that is dropped if $l(\beta_t) = 1$ for all β_t. Its pdf has the form

$$\pi(\beta_t|\Xi_{t-1}, \nu_{t-1}) = m(\Xi_{t-1}, \nu_{t-1}) l(\beta_t)^{\nu_{t-1}} e^{\eta(\beta_t)^\intercal \Xi_{t-1}}, \tag{1.52}$$

where $m(\Xi_{t-1}, \nu_{t-1})$ is a known function, and $l(\beta_t)$ is the same function as in (1.51).

Lemma 1.3. The Bayesian update (1.50) multiplying the model (1.51) with the prior pdf (1.52) results in the posterior pdf of the same functional type as the prior, characterized by the posterior hyperparameters

$$\begin{aligned}\Xi_t &= \Xi_{t-1} + T(x_t, \widetilde{y}_t),\\ \nu_t &= \nu_{t-1} + 1.\end{aligned} \tag{1.53}$$

Proof. The proof is straightforward. □

In cases where we have multiple data, the Bayesian update (1.50) has a form of

$$\pi(\beta_k|\widetilde{y}_{0:k}, x_{0:k}) \propto \pi(\beta_\tau|\widetilde{y}_{0:\tau-1}, x_{0:\tau-1}) \prod_{\widetilde{\tau}=\tau}^{k} f(\widetilde{y}_{\widetilde{\tau}}|x_{\widetilde{\tau}}, \beta_{\widetilde{\tau}}). \tag{1.54}$$

The pdf of the posterior distribution is therefore characterized by the posterior hyperparameters

$$\begin{aligned}\Xi_k &= \Xi_{\tau-1} + \sum_{\widetilde{\tau}=\tau}^{k} T(x_{\widetilde{\tau}}, \widetilde{y}_{\widetilde{\tau}}),\\ \nu_k &= \nu_{\tau-1} + k - \tau + 1.\end{aligned} \tag{1.55}$$

This result allows for efficient sequential estimation of β_t, given that the Bayesian update (1.50) is equivalent to simple summations, and the functional form of the posterior density is the same as that of the prior density. Due to this fact, the posterior pdf can serve as the prior pdf for the next time instant.

To demonstrate the described principle, we can perform Bayesian estimation of the parameter λ of the Poisson distribution. Suppose that we have a Poisson variable

$$Y \sim Po(\lambda). \tag{1.56}$$

Then the conjugate prior (see Definition 1.2) of the parameter λ is the gamma distribution [38], meaning

$$\lambda \sim \Gamma(a, b), \tag{1.57}$$

where a is a shape parameter and b is an inverse scale parameter. The likelihood of observations y_t, according to Equation (1.51) (see Definition 1.1), reads

$$
\begin{aligned}
f(y_t|\lambda_t) &= \frac{1}{y_t!}\lambda_t^{y_t} e^{-\lambda_t} && (1.58)\\
&= \frac{1}{y_t!} \exp\left\{\ln\left(\lambda_t^{y_t} e^{-\lambda_t}\right)\right\} && (1.59)\\
&= \frac{1}{y_t!} \exp\left\{y_t \ln\lambda_t - \lambda_t\right\} && (1.60)\\
&= \underbrace{\frac{1}{y_t!}}_{k(y_t)} \cdot \underbrace{1}_{l(\lambda_t)} \cdot \exp\left\{ \underbrace{\begin{bmatrix}\ln\lambda_t \\ -\lambda_t\end{bmatrix}^\intercal}_{\eta(\lambda_t)} \underbrace{\begin{bmatrix} y_t \\ 1\end{bmatrix}}_{T(y_t)} \right\}. && (1.61)
\end{aligned}
$$

The pdf of the conjugate prior distribution, according to Equation (1.52), reads

$$
\begin{aligned}
\pi(\lambda_t|a_{t-1}, b_{t-1}) &= \frac{b_{t-1}^{a_{t-1}}}{\Gamma(a_{t-1})}\lambda_t^{a_{t-1}-1} e^{-b_{t-1}\lambda_t} && (1.62)\\
&= \frac{b_{t-1}^{a_{t-1}}}{\Gamma(a_{t-1})} \exp\left\{\ln\left(\lambda_t^{a_{t-1}-1} e^{-b_{t-1}\lambda_t}\right)\right\} && (1.63)\\
&= \frac{b_{t-1}^{a_{t-1}}}{\Gamma(a_{t-1})} \exp\left\{(a_{t-1}-1)\ln\lambda_t - b_{t-1}\lambda_t\right\} && (1.64)\\
&= \underbrace{\frac{b_{t-1}^{a_{t-1}}}{\Gamma(a_{t-1})}}_{m(\Xi_{t-1},\nu_{t-1})} \cdot \underbrace{1}_{l(\lambda_t)^{\nu_{t-1}}} \cdot \exp\left\{ \underbrace{\begin{bmatrix}\ln\lambda_t \\ -\lambda_t\end{bmatrix}^\intercal}_{\eta(\lambda_t)} \underbrace{\begin{bmatrix} a_{t-1}-1 \\ b_{t-1}\end{bmatrix}}_{\Xi_{t-1}} \right\}. && (1.65)
\end{aligned}
$$

If we now perform the Bayesian update (1.50), then the posterior pdf is of the same functional type as the prior and is characterized by hyperparameters Ξ_t and ν_t (see Equation (1.53)). Furthermore, if we take the individual components of the hyperparameter Ξ_t, we can obtain the actual parameters a_t and b_t of the posterior gamma distribution, leaving us with

$$
\begin{aligned}
a_t &= \Xi_{t,0} + 1, && (1.66)\\
b_t &= \Xi_{t,1}. && (1.67)
\end{aligned}
$$

Figure 1.4 shows results of a simulated Bayesian estimation of the parameter λ. In this case $\lambda = 40$ and the prior parameters of the gamma distribution $a = 0.001$, $b = 0.001$. The estimates converge to the actual value after around 1000 time steps. Figure 1.5 shows an evolution of the RMSE of estimates. As we can see, the RMSE systematically decreases with time.

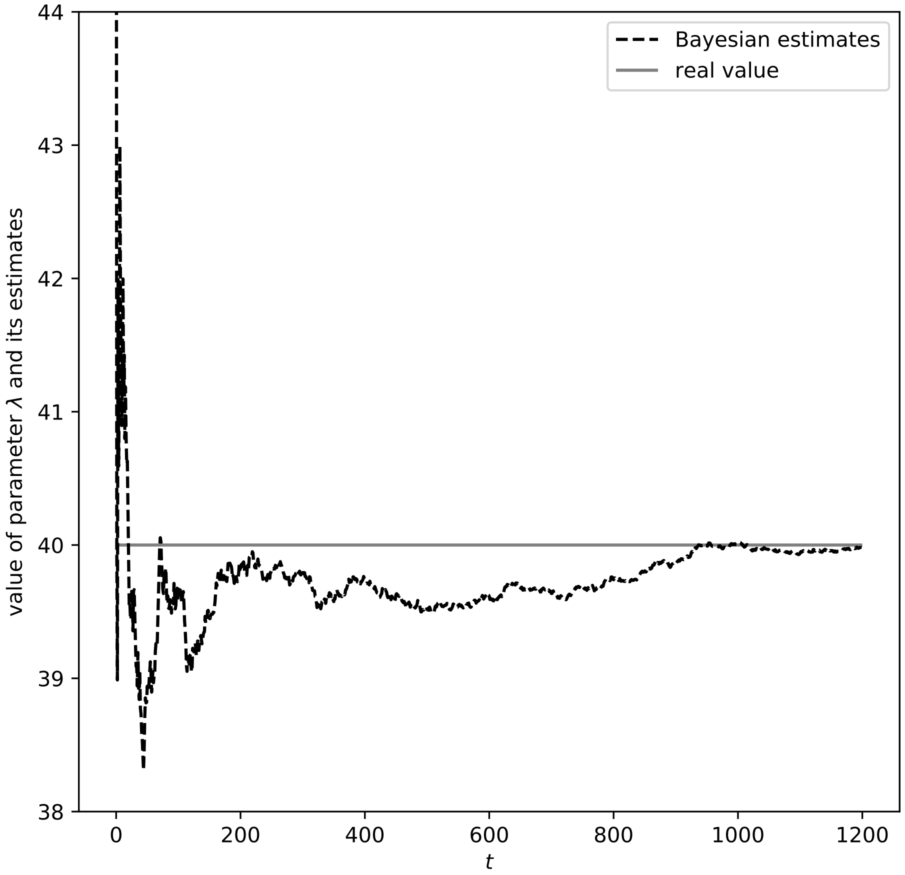

Figure 1.4: Evolution of the Bayesian estimation of parameter λ of a Poisson random variable. Solid line represents the actual value of the parameter λ. Dashed line represents the Bayesian estimates.

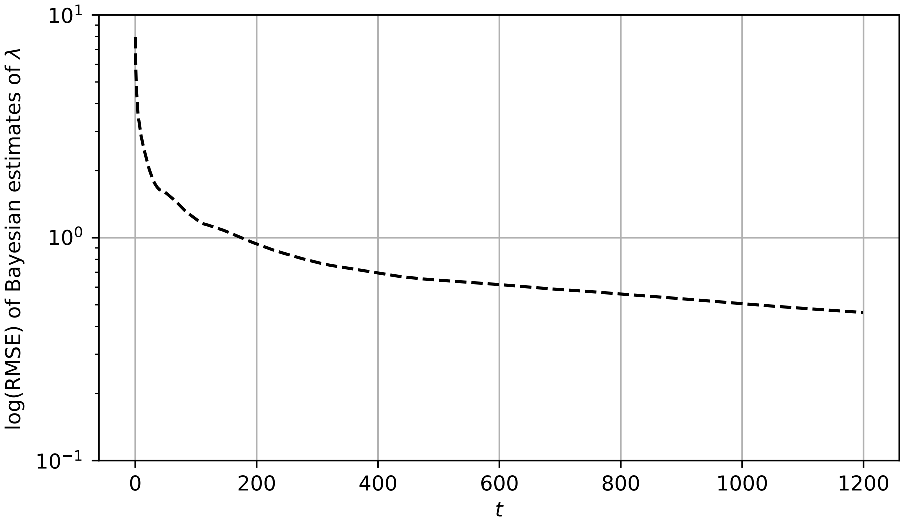

Figure 1.5: Evolution of the RMSE of the Bayesian estimates of parameter λ of a Poisson random variable.

1.3.2 Guassian approximation of the likelihood of β in the posterior distribution

When comparing the true data model (1.47) with the exponential family form (1.51), it is obvious that there cannot exist a convenient conjugate prior distribution of the form (1.52). For the one-shot static Poisson regression with no variance stabilization, a workaround is suggested in [26]. It consists of approximating the likelihood of the (time-invariant) β for all the observed data $x_{1:t}$,

$$f(y_{1:t}|\beta, x_{1:t}) = \prod_{\tau} f(y_\tau|\beta, x_\tau) \tag{1.68}$$

in the posterior distribution by a normal distribution. The approximation was originally proposed by Bartlett and Kendall [31]. Even though our aim is to update the estimate of β_t sequentially with the incoming observations (and additionally we deal with the transformed variable $\widetilde{y}_t = \sqrt{y_t}$), a similar philosophy can be adopted.

Let us once again get back to the Bayesian update (1.50). In the posterior pdf, the model (1.47) acts as a function of β_t, while x_t and $\widetilde{y}_t$ are fixed. Figure 1.6 depicts this (renormalized) function for four selected values of $\widetilde{y}_t$. Suppose that we have a real random variable u whose density function is proportional to $\exp(uz)\exp(-\exp(u))$ where z stands for a parameter. Then the density function can be approximated by a normal distribution $\mathcal{N}(\log z, z^{-1})$, provided that z is a large number [31]. In Equation (1.47) with x_t and $\widetilde{y}_t$ fixed, this leads to the approximation by $\mathcal{N}(\log \widetilde{y}_t^2, \widetilde{y}_t^{-2})$. However, the approximation is crude if $\widetilde{y}_t$ is low. The mean values differ even by 0.58 if $\widetilde{y}_t = 1$.

In order to compensate the approximation error under low values of $\widetilde{y}_t$, the following moment matching-based calibration is suggested: the bias of both the approximating mean value and the standard deviation can be predicted and suppressed (with sufficient accuracy) using the regression models with $\widetilde{y}_t$ in the role of the regressand. The calibrated approximative normal distribution with the bias removed has the mean and standard deviation

$$\begin{aligned}\mu_c &= \log \widetilde{y}_t^2 - \frac{0.5574}{\widetilde{y}_t^2},\\ \sigma_c &= \frac{1}{\widetilde{y}_t} + \frac{0.0724}{\widetilde{y}_t^2} + \frac{0.2121}{\widetilde{y}_t^4}.\end{aligned} \tag{1.69}$$

The coefficients were obtained from the OLS (ordinary least squares) over the values $\widetilde{y}_t^2 = 1, \ldots, 100$. Figures 1.7 and 1.8 depict the compensated approximation error and the related prediction error due to the model for the mean value and the standard deviation, respectively. Fig. 1.9 compares the true distribution of $\widetilde{y}_t$, the calibrated, and the noncalibrated normal approximations. It is also important to note that since $y_t \in \mathbb{N}$, it is possible to use a table of precomputed values of μ_c and σ_c^2 for low values of y_t (see Table 1.2).

μ_c σ_c^2	…+ 1	…+ 2	…+ 3	…+ 4	…+ 5	…+ 6	…+ 7	…+ 8	…+ 9	…+ 10
$y_t = 0 + \ldots$	-0.5574	0.4144	0.9128	1.2469	1.4980	1.6989	1.8663	2.0098	2.1353	2.2468
	1.6499	0.6341	0.3907	0.2823	0.2211	0.1817	0.1542	0.1339	0.1183	0.1060
$y_t = 10 + \ldots$	2.3472	2.4385	2.5221	2.5992	2.6709	2.7378	2.8004	2.8594	2.9151	2.9679
	0.0960	0.0877	0.0808	0.0748	0.0697	0.0652	0.0613	0.0578	0.0547	0.0519
$y_t = 20 + \ldots$	3.0180	3.0657	3.1113	3.1548	3.1966	3.2367	3.2752	3.3123	3.3481	3.3826
	0.0493	0.0471	0.0450	0.0431	0.0413	0.0397	0.0382	0.0368	0.0355	0.0343
$y_t = 30 + \ldots$	3.4160	3.4483	3.4796	3.5100	3.5394	3.5680	3.5959	3.6229	3.6493	3.6749
	0.0332	0.0321	0.0311	0.0302	0.0293	0.0285	0.0277	0.0270	0.0263	0.0256
$y_t = 40 + \ldots$	3.7000	3.7244	3.7482	3.7715	3.7943	3.8165	3.8383	3.8596	3.8804	3.9009
	0.0250	0.0244	0.0238	0.0233	0.0227	0.0222	0.0218	0.0213	0.0209	0.0204
$y_t = 50 + \ldots$	3.9209	3.9405	3.9598	3.9787	3.9972	4.0154	4.0333	4.0508	4.0681	4.0851
	0.0200	0.0196	0.0193	0.0189	0.0186	0.0182	0.0179	0.0176	0.0173	0.0170
$y_t = 60 + \ldots$	4.1017	4.1181	4.1343	4.1502	4.1658	4.1812	4.1964	4.2113	4.2260	4.2405
	0.0167	0.0164	0.0162	0.0159	0.0157	0.0154	0.0152	0.0150	0.0148	0.0145
$y_t = 70 + \ldots$	4.2548	4.2689	4.2828	4.2965	4.3101	4.3234	4.3366	4.3496	4.3624	4.3751
	0.0143	0.0141	0.0139	0.0138	0.0136	0.0134	0.0132	0.0130	0.0129	0.0127
$y_t = 80 + \ldots$	4.3876	4.3999	4.4121	4.4242	4.4361	4.4479	4.4595	4.4710	4.4824	4.4936
	0.0126	0.0124	0.0122	0.0121	0.0120	0.0118	0.0117	0.0115	0.0114	0.0113
$y_t = 90 + \ldots$	4.5047	4.5157	4.5266	4.5374	4.5480	4.5585	4.5690	4.5793	4.5895	4.5996
	0.0112	0.0110	0.0109	0.0108	0.0107	0.0106	0.0105	0.0104	0.0103	0.0101

Table 1.2: Table of precomputed values of μ_c and σ_c^2 for $\widetilde{y}_t^2 = 1, \ldots, 100$.

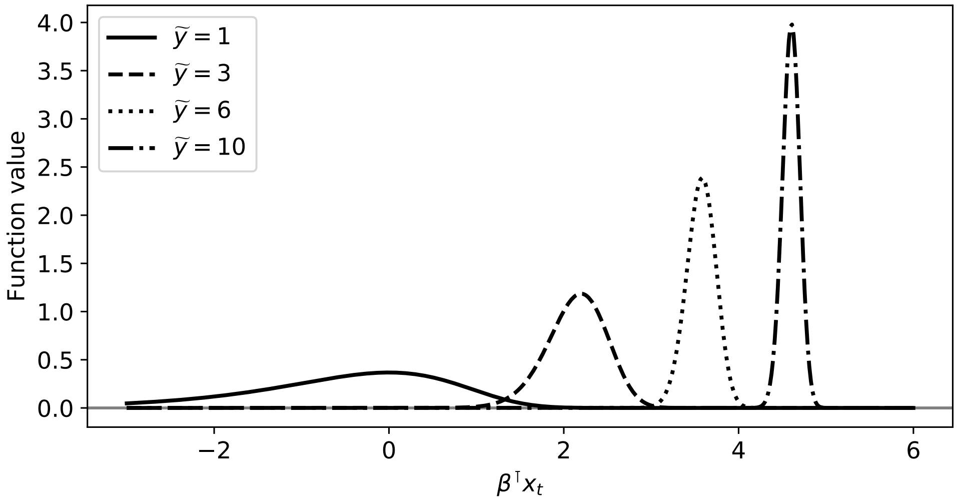

Figure 1.6: Shape of renormalized function (1.47) for different fixed observation value $\widetilde{y}_t$, regressor x_t, and variable β.

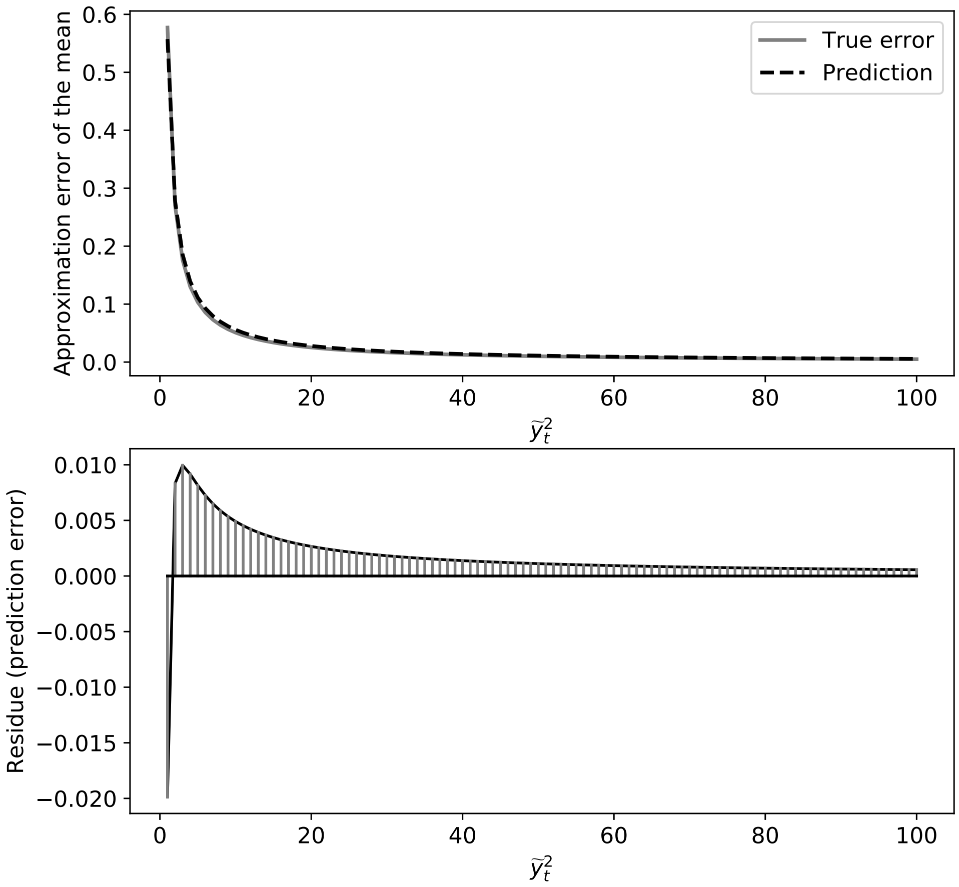

Figure 1.7: Calibration of the mean value. Top: The true value of the approximation error under $\mathcal{N}(\log \widetilde{y}_t^2, \widetilde{y}_t^{-2})$ and its regression-based prediction. Bottom: Evolution of the prediction error.

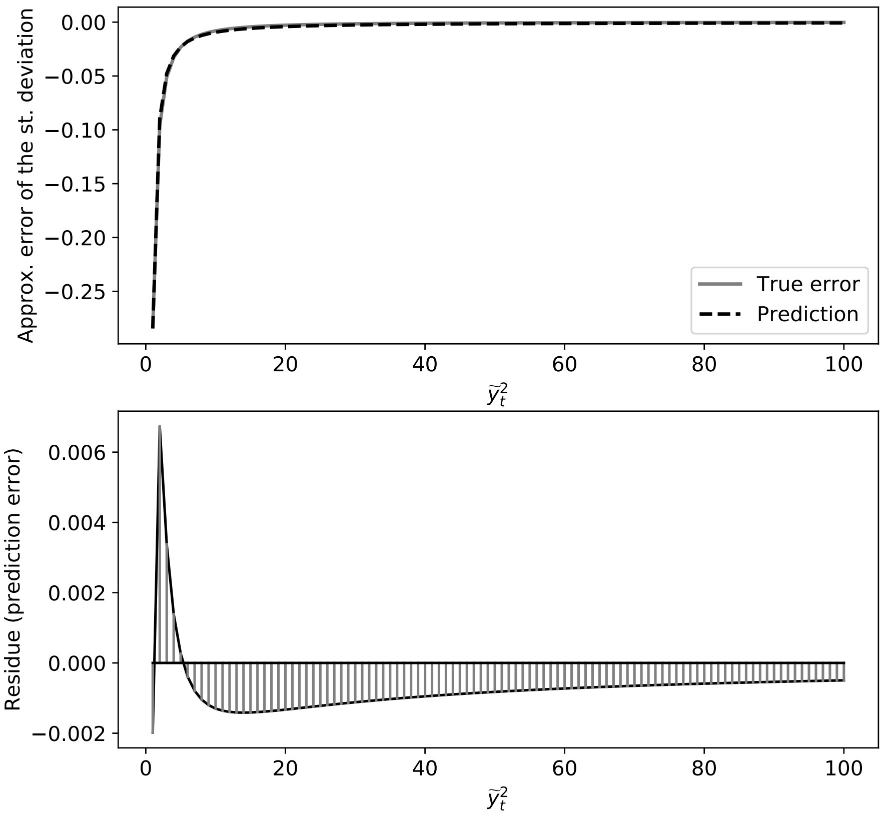

Figure 1.8: Calibration of the standard deviation. Top: The true value of the approximation error under $\mathcal{N}(\log \widetilde{y}_t^2, \widetilde{y}_t^{-2})$ and its regression-based prediction. Bottom: Evolution of the prediction error.

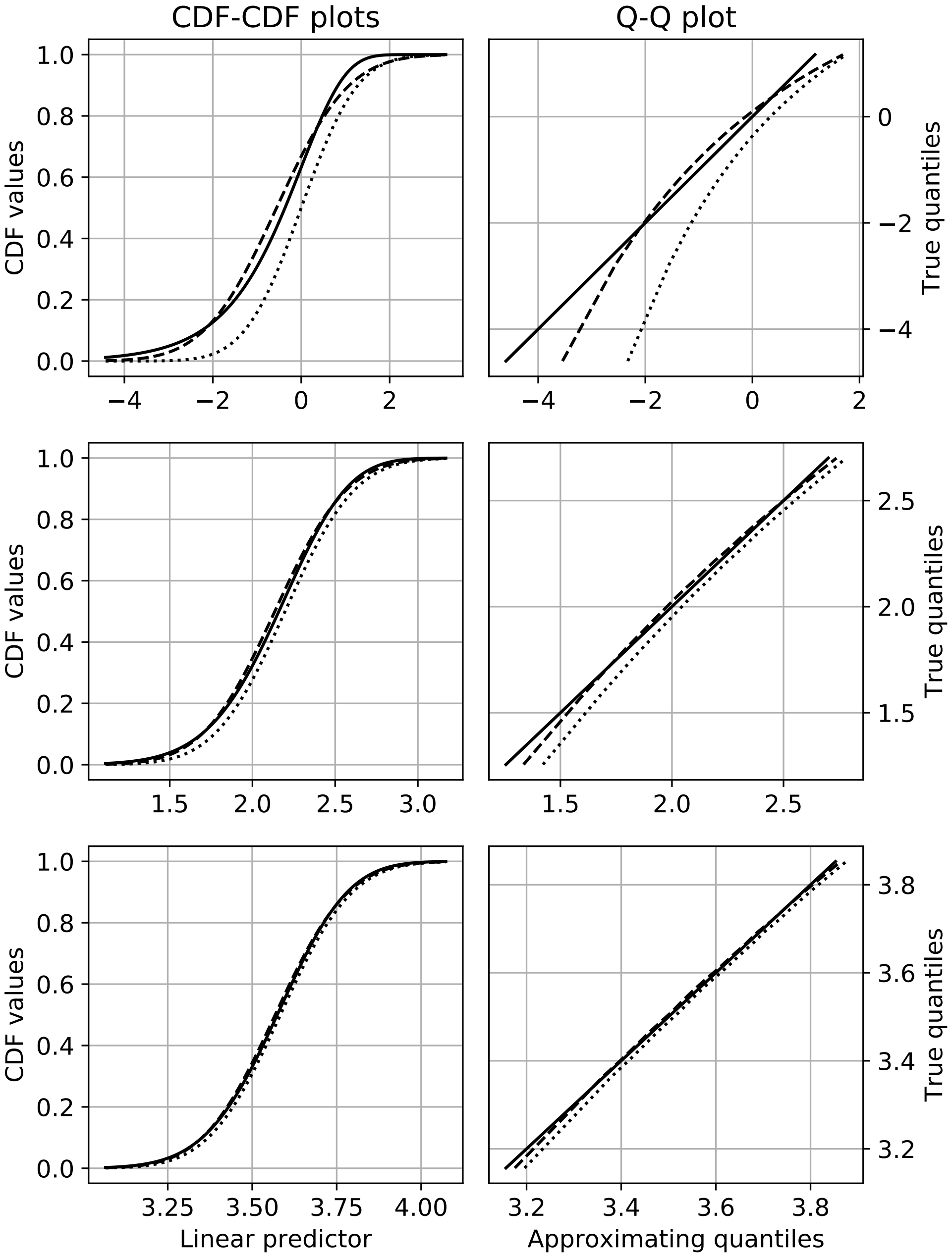

Figure 1.9: Normal approximation of the true distribution of $\widetilde{y}_t$. Shown are approximations for the values 1 (top), 3 (middle), and 6 (bottom), i.e., $y_t = 1$, 9, and 36. The solid line depicts the true distribution, the dotted line is used for the noncalibrated normal distribution, and the dashed line represents the calibrated normal distribution.

1.4 The posterior distribution

The posterior distribution $\pi(\beta_t|x_{0:t}, \widetilde{y}_{0:t})$, as specified in Equation (1.50), now consists of the normal distribution $\mathcal{N}(\mu_c, \sigma_c^2)$ defined by the moments (1.69) and a prior distribution $\pi(\beta_t|x_{0:t-1}, \widetilde{y}_{0:t-1})$. The normal distribution belongs to the exponential family (see Definition 1.1) and its pdf can be written in the form (1.51),

$$\begin{aligned} f(y_t|x_t, \beta_t) &= \frac{1}{\sqrt{2\pi\sigma_{c,t}^2}} \exp\left\{ -\frac{1}{2\sigma_{c,t}^2} ||\mu_{c,t} - \beta_t^\intercal x_t||^2 \right\} \\ &\propto \exp\Bigg\{ -\frac{1}{2} \operatorname{Tr}\Bigg(\underbrace{\begin{bmatrix} -1 \\ \beta_t \end{bmatrix} \begin{bmatrix} -1 \\ \beta_t \end{bmatrix}^\intercal}_{\eta \equiv \eta(\beta_t)} \underbrace{\begin{bmatrix} \mu_{c,t} \\ x_t \end{bmatrix} \begin{bmatrix} \mu_{c,t} \\ x_t \end{bmatrix}^\intercal \sigma_{c,t}^{-2}}_{T(x_t, \widetilde{y}_t)} \Bigg) \Bigg\}, \end{aligned} \tag{1.70}$$

where, to avoid vectorizations, a slight simplification of the notation is used. Due to this, the appropriate conjugate prior distribution is the normal distribution with the mean vector b_{t-1} and the covariance matrix P_{t-1}. The pdf of the prior distribution in the compatible form is

$$\pi(\beta|b_{t-1}, P_{t-1}) \propto \exp\Bigg\{ -\frac{1}{2} \operatorname{Tr}\Bigg(\underbrace{\begin{bmatrix} -1 \\ \beta \end{bmatrix} \begin{bmatrix} -1 \\ \beta \end{bmatrix}^\intercal}_{\eta \equiv \eta(\beta)} \underbrace{\begin{bmatrix} b_{t-1}^\intercal \\ I \end{bmatrix} P_{t-1}^{-1} \begin{bmatrix} b_{t-1}^\intercal \\ I \end{bmatrix}^\intercal}_{\Xi_{t-1}} \Bigg) \Bigg\}, \tag{1.71}$$

where I is the $n \times n$ identity matrix. Then, the posterior distribution following from the Bayes' rule (1.50) is (in terms of the update of the prior hyperparameters (1.53)) given by

$$\Xi_t = \Xi_{t-1} + T(x_t, \widetilde{y}_t). \tag{1.72}$$

If we take a closer look at the matrix Ξ_{t-1} in Equation (1.71), it can be expanded as

$$\begin{aligned} \Xi_{t-1} &= \begin{bmatrix} b_{t-1}^\intercal \\ I \end{bmatrix} P_{t-1}^{-1} \begin{bmatrix} b_{t-1}^\intercal \\ I \end{bmatrix}^\intercal && (1.73) \\ &= \begin{bmatrix} \underbrace{b_{t-1}^\intercal P_{t-1}^{-1} b_{t-1}}_{1\times 1} & \underbrace{b_{t-1}^\intercal P_{t-1}^{-1}}_{1\times n} \\ \underbrace{P_{t-1}^{-1} b_{t-1}}_{n\times 1} & \underbrace{P_{t-1}^{-1}}_{n\times n} \end{bmatrix}. && (1.74) \end{aligned}$$

Analogically

$$\Xi_t = \begin{bmatrix} \underbrace{b_t^\intercal P_t^{-1} b_t}_{1\times 1} & \underbrace{b_t^\intercal P_t^{-1}}_{1\times n} \\ \underbrace{P_t^{-1} b_t}_{n\times 1} & \underbrace{P_t^{-1}}_{n\times n} \end{bmatrix}. \tag{1.75}$$

Similarly, the matrix $T(x_t, \widetilde{y}_t)$ in Equation (1.70) can be expanded as

$$
\begin{aligned}
T(x_t, \widetilde{y}_t) &= \begin{bmatrix} \mu_{c,t} \\ x_t \end{bmatrix} \begin{bmatrix} \mu_{c,t} \\ x_t \end{bmatrix}^\intercal \sigma_{c,t}^{-2} && (1.76) \\
&= \begin{bmatrix} \underbrace{\sigma_{c,t}^{-2} \mu_{c,t}^2}_{1\times 1} & \underbrace{\sigma_{c,t}^{-2} \mu_{c,t} x_t^\intercal}_{1\times n} \\ \underbrace{\sigma_{c,t}^{-2} x_t \mu_{c,t}}_{n\times 1} & \underbrace{\sigma_{c,t}^{-2} x_t x_t^\intercal}_{n\times n} \end{bmatrix}. && (1.77)
\end{aligned}
$$

Finally, a little algebra reveals that the 'conventional' normal hyperparameters of the posterior distribution are

$$
\begin{aligned}
P_t &= \Xi_{t,1:n,1:n}^{-1} \\
&= \left(\Xi_{t-1,1:n,1:n} + T(x_t, \widetilde{y}_t)_{1:n,1:n}\right)^{-1} \\
&= \left(P_{t-1}^{-1} + \sigma_{c,t}^{-2} x_t x_t^\intercal\right)^{-1}, && (1.78)
\end{aligned}
$$

and

$$
\begin{aligned}
b_t &= P_t P_t^{-1} b_t \\
&= \Xi_{t,1:n,1:n}^{-1} \Xi_{t,1:n,0} \\
&= \left(\Xi_{t-1,1:n,1:n} + T(x_t, \widetilde{y}_t)_{1:n,1:n}\right)^{-1} \left(\Xi_{t-1,1:n,0} + T(x_t, \widetilde{y}_t)_{1:n,0}\right) \\
&= \left(P_{t-1}^{-1} + \sigma_{c,t}^{-2} x_t x_t^\intercal\right)^{-1} \left(P_{t-1}^{-1} b_{t-1} + \sigma_{c,t}^{-2} x_t \mu_{c,t}\right). && (1.79)
\end{aligned}
$$

1.5 Time-varying β_t

Having constant model parameters is rather an exception than a rule. The usual problem is that no explicit model for the evolution $\beta_{t-1} \to \beta_t$ exists. However, if the variations are slow, we may proceed by means of *forgetting*, which means heuristic discounting of possibly outdated information about β_t from the posterior distribution. The most basic yet one of the most used approaches to this problem is the exponential forgetting, flattening the prior pdf by its exponentiation [39],

$$
\pi(\beta_t | x_{0:t-1}, y_{0:t-1}) = [\pi(\beta_{t-1} | x_{0:t-1}, y_{0:t-1})]^\alpha, \quad \alpha \in [0, 1]. \qquad (1.80)
$$

In conjugate priors (see Definition 1.2) this amounts to

$$
\begin{aligned}
\nu_{t-1} &\leftarrow \alpha \nu_{t-1}, \\
\Xi_{t-1} &\leftarrow \alpha \Xi_{t-1}. && (1.81)
\end{aligned}
$$

For summary of other (more elaborate) forgetting methods, see, e.g., [40] and the references therein.

Algorithm 1 Sequential Poisson Regression

Set the prior distribution $\mathcal{N}(b_0, P_0)$. Set the forgetting factor α. For $t = 1, 2, \ldots$ do:

1. Gather observations x_t, y_t.
2. Flatten the prior distribution, Eq. (1.81).
3. Update the prior hyperparameter, Eq. (1.72)
4. Evaluate the point estimate $\bar{b}_t$ and the covariance matrix $\bar{P}_t$ from $\bar{\Xi}_t$, Equations (1.78) and (1.79).

Chapter 2

Signal Processing Domain Application

As already stated, distributed inference of unknown variables is an established discipline in the signal processing domain. Assume that we have a network consisting of a set $\mathcal{I}$ of agents that independently observe the processes

$$Y_t^{(i)} \sim Po(\exp(\beta_t^\intercal x_t^{(i)})), \tag{2.1}$$

with the observations $y_t^{(i)}$ being local, regressors $x_t^{(i)}$ being potentially local, and β_t being global (i.e., identical for all $i \in \mathcal{I}$). These types of situations may occur, e.g., in particle detection, where each agent observes different number $y_t^{(i)}$ of particles generated by a single underlying process, and employs a time-series model with $x_t^{(i)}$ consisting of past observations. Let $\mathcal{I}^{(i)}$ denote the set of adjacent neighbors of agent i, and let $i \in \mathcal{I}^{(i)}$ too. Now, suppose that at each time instant t, every agent i may perform one mutual exchange of the posterior pdfs with all its adjacent neighbors $j \in \mathcal{I}^{(i)}$ withing 1 network hop distance. Note that the pdfs are fully represented by $\Xi_t^{(i)}$.

Looking closely at the Bayesian update (1.53), we can see that $\Xi_t^{(i)}$ summarizes the information contained in the past sufficient statistics $T(x_\tau^{(i)}, \widetilde{y}_\tau^{(i)}), \tau = 0, \ldots, t$ where $T(x_0^{(i)}, \widetilde{y}_0^{(i)})$ represent the initial pseudo-observations (see Section 1.3). Therefore, the combination of the posterior pdfs in terms of the hyperparameter averaging

$$\bar{\Xi}_t^{(i)} = \frac{1}{\operatorname{card}(\mathcal{I}^{(i)})} \sum_{j \in \mathcal{I}^{(i)}} \Xi_t^{(j)}, \tag{2.2}$$

where $\operatorname{card}(\mathcal{I}^{(i)})$ denotes the cardinality of $\mathcal{I}^{(i)}$, amounts to the uniformly weighted Bayesian update by observations of adjacent neighbors. Analogically to Equation (1.75), the matrix $\Xi_t^{(i)}$ actually consists of the 'conventional' hyperparameters, meaning

$$\Xi_t^{(i)} = \begin{bmatrix} \underbrace{b_t^{\intercal,(i)} P_t^{-1,(i)} b_t^{(i)}}_{1\times 1} & \underbrace{b_t^{\intercal,(i)} P_t^{-1,(i)}}_{1\times n} \\ \underbrace{P_t^{-1,(i)} b_t^{(i)}}_{n\times 1} & \underbrace{P_t^{-1,(i)}}_{n\times n} \end{bmatrix}. \tag{2.3}$$

Thus, from a little algebra (similarly to Equations (1.78) and (1.79)) it follows that

$$\bar{P}_t^{(i)} = \left(\frac{1}{\mathrm{card}(\mathcal{I}^{(i)})} \sum_{j\in\mathcal{I}^{(i)}} P_t^{-1,(j)} \right)^{-1}, \tag{2.4}$$

$$\bar{b}_t^{(i)} = \bar{P}_t^{(i)} \left(\frac{1}{\mathrm{card}(\mathcal{I}^{(i)})} \sum_{j\in\mathcal{I}^{(i)}} P_t^{-1,(j)} b_t^{(j)} \right), \tag{2.5}$$

which is known as the *covariance intersection.* In [24], K. Dedecius shows that this result is Kullback-Leibler-optimal. A careful inspection of Equation (2.2) shows that it actually corresponds to uniformly weighted averaging of neighbors' knowledge about β_t. That is, however, not a problem, since the covariance matrices $P_t^{(j)}$ (see Equation (2.4)) effectively reflect the uncertainty about the individual estimates. More elaborate combination strategies are proposed, e.g., in [41].

Algorithm 2 DIFFUSION POISSON REGRESSION

```
For each agent $i \in \mathcal{I}$ set the prior distribution $\mathcal{N}(b_0^{(i)}, P_0^{(i)})$. Set the forgetting factor $\alpha$. For $t = 1, 2, \ldots$ and each node $i \in \mathcal{I}$ do:
Local estimation:
    1. Gather observations $x_t^{(i)}, y_t^{(i)}$.
    2. Flatten the prior distribution, Eq. (1.81).
    3. Update the prior hyperparameter, Eq. (1.72)
Combination:
    1. Get posterior pdfs $\pi(\beta_t | b_t^{(j)}, P_t^{(j)})$ of neighbors $j \in \mathcal{I}^{(i)}$.
    2. Combine the posterior hyperparameters, Eq. (2.2), or in terms of $b_t^{(j)}$ and $P_t^{(j)}$, Equations (2.4) and (2.5).
    3. Evaluate the point estimate $\bar{b}_t^{(i)}$ and the covariance matrix $\bar{P}_t^{(i)}$ from $\bar{\Xi}_t^{(i)}$.
```

Chapter **3**

Simulation Examples

3.1 Single-node estimation with static parameters

The first set of examples demonstrates the efficiency of using the calibration described in Section 1.3.2, based on the range of values of observations y_t. In all simulations, two models are compared: (i) the 'calibrated' model using the mean and variance defined in Equation (1.69), and (ii) the 'non-calibrated' model using the original mean and variance. Results of all simulations were averaged over 30 independent runs. The forgetting factor was not used, meaning $\alpha = 1$. Both models observe the same independently generated outcomes of a Poisson regression process. The initial prior distribution is the normal distribution with $b_0 = [0, 0, 0, 0]^\intercal$ and $P_0 = 100 \cdot I$ where I is the identity 4×4 matrix.

The first simulation was run with a vector of static regression coefficients $\beta = [0.9, 0.5, 0.2, 0.1]^\intercal$, and randomly generated regressors

$$x_{t,0} = 1, \tag{3.1}$$

$$x_{t,1} \sim U(0.3, 2.1), \tag{3.2}$$

$$x_{t,1} \sim U(0.5, 2.2), \tag{3.3}$$

$$x_{t,1} \sim U(0.9, 1.5). \tag{3.4}$$

Values of observations y_t ranged from 2 to 33 with a mean of 8. Fig. 3.1 shows the evolution of RMSE for both models. As we can see, the calibrated model clearly outperforms the non-calibrated model. Fig.3.2 then shows the stability of estimates for one randomly selected run of the algorithm.

The second simulation was run with a vector of regression coefficients $\beta = [0.8, 0.9, 1.1, 1.3]^\intercal$, and randomly generated regressors

$$x_{t,0} = 1, \tag{3.5}$$

$$x_{t,1} \sim U(0.2, 1.1), \tag{3.6}$$

$$x_{t,1} \sim U(0.1, 0.9), \tag{3.7}$$

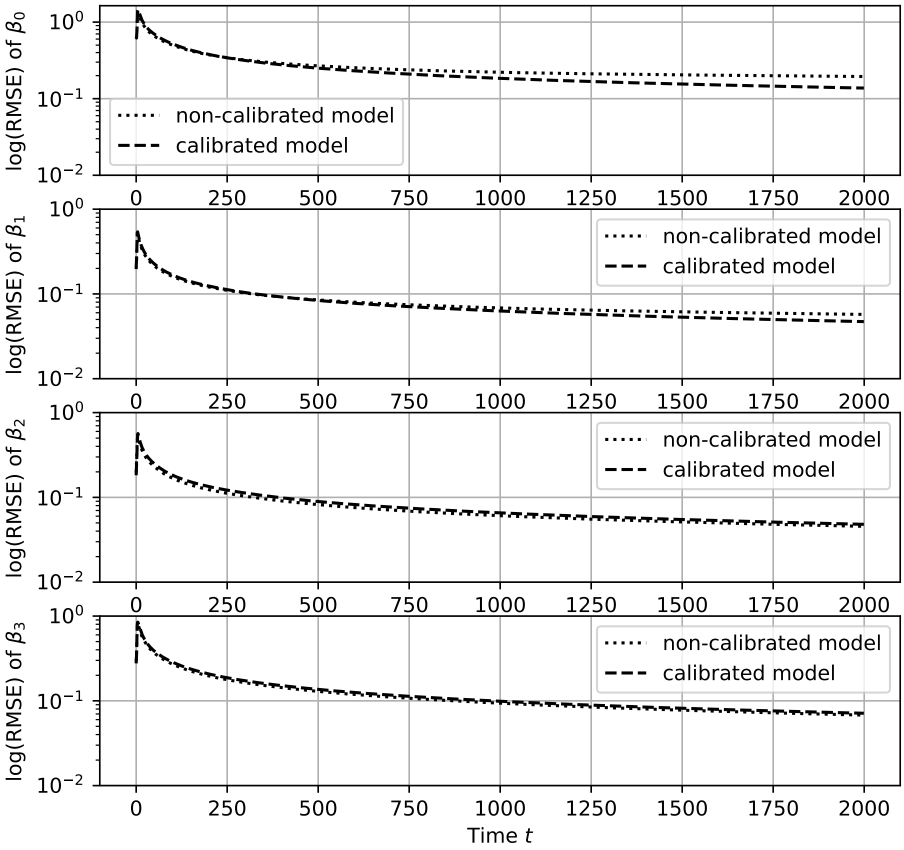

Figure 3.1: Evolution of the RMSE for low values of y_t.

$$x_{t,1} \sim U(0.3, 0.8). \tag{3.8}$$

Values of observations y_t ranged from 2 to 101 with a mean of 23. Fig. 3.3 shows the evolution of RMSE and Fig. 3.4 show the stability of estimates for one randomly selected run. As we can see, the calibrated model still dominates the non-calibrated one, although the difference is not as noticable as in the previous simulation.

The last simulation in this set was run with a vector of regression coefficients $\beta = [1.05, 2.41, 3.27, 3.87]^\intercal$, and randomly generated regressors

$$x_{t,0} = 1, \tag{3.9}$$
$$x_{t,1} \sim U(0.06, 0.1), \tag{3.10}$$
$$x_{t,1} \sim U(0.1, 0.5), \tag{3.11}$$
$$x_{t,1} \sim U(0.1, 0.7). \tag{3.12}$$

Values of observations y_t ranged from 2 to 666 with a mean of 99. Fig. 3.5 shows the evolution of RMSE and Fig. 3.6 show the stability of estimates for

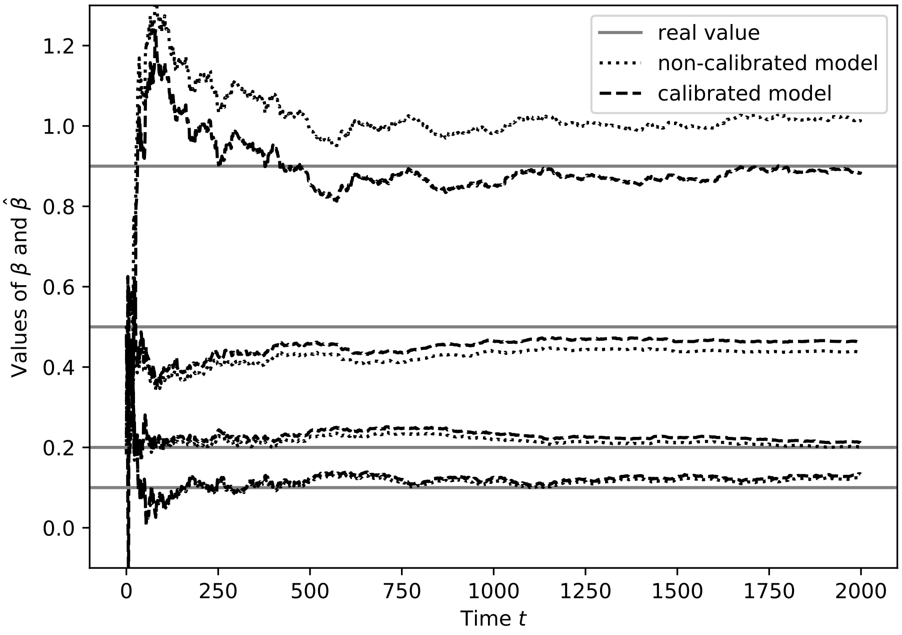

Figure 3.2: Real and estimated values of β in time for low values of y_t.

one randomly selected run. It is clear that with such large values it is no longer possible to observe a noticeable difference in the quality of estimates.

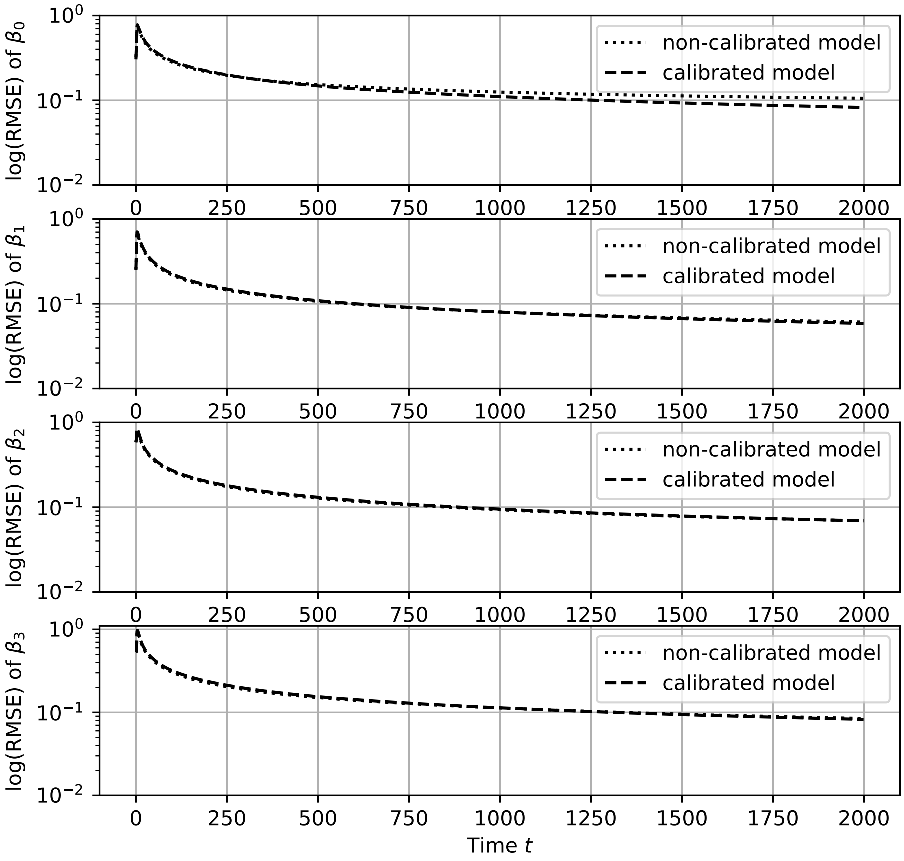

Figure 3.3: Evolution of the RMSE for slightly larger values of y_t.

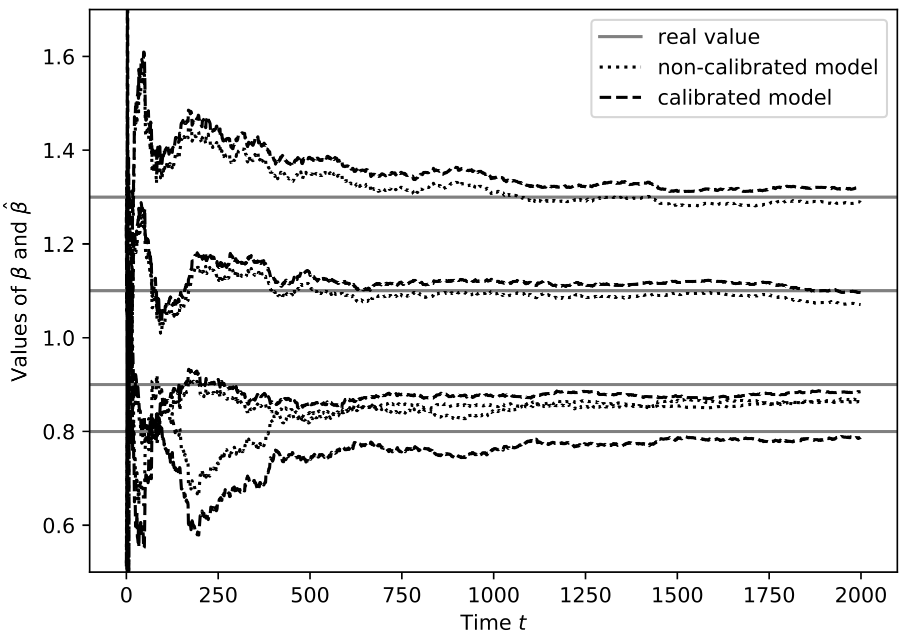

Figure 3.4: Real and estimated values of β in time for slightly larger values of y_t.

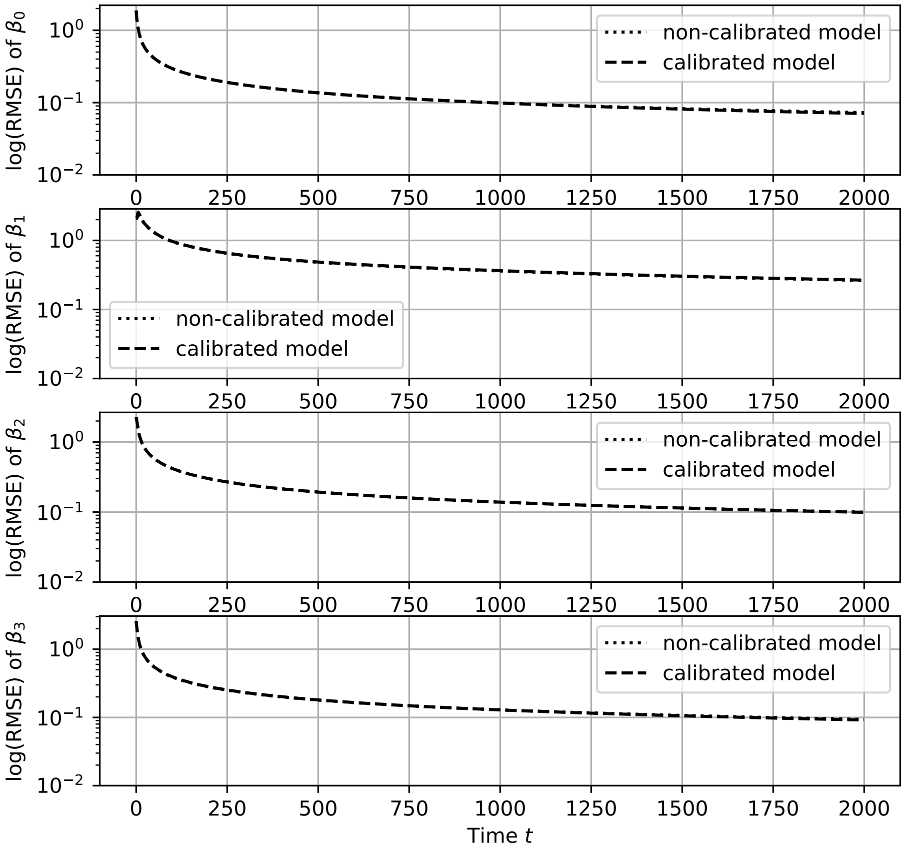

Figure 3.5: Evolution of the RMSE for large values of y_t.

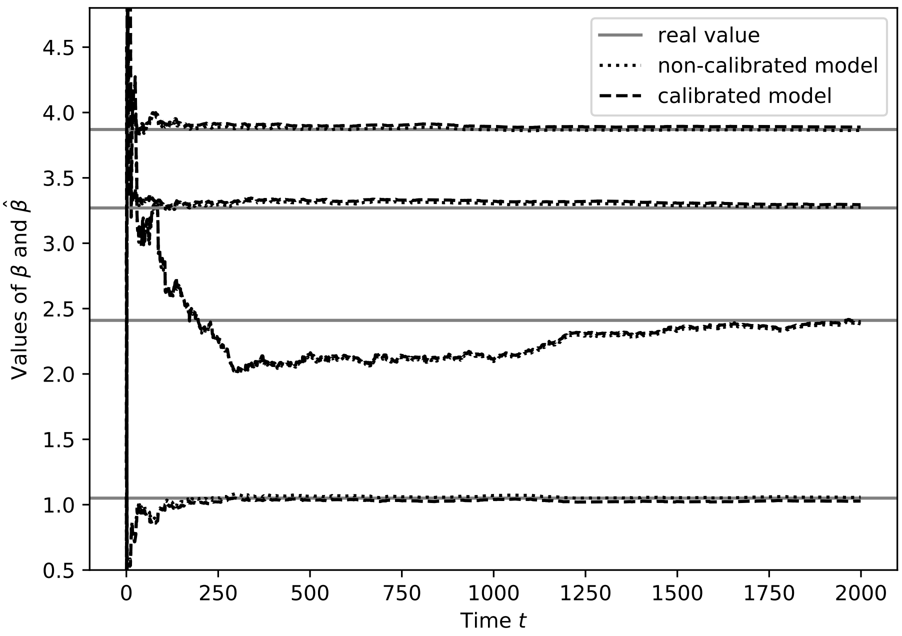

Figure 3.6: Real and estimated values of β in time for large values of y_t.

3.2 Single-node estimation with time-varying parameters

The following set of examples demonstrates the accuracy of the sequential estimation for different frequencies of parameters and different values of the forgetting factor α (within a reasonable range from 0.95 to 1), as specified in Section 1.5. Results of all simulations were averaged over 30 independent runs. All models observe the same independently generated outcomes of a Poisson regression process. The initial prior distribution is the normal distribution with $b_0 = [0, 0, 0, 0]^\intercal$ and $P_0 = 100 \cdot I$ where I is the identity 4×4 matrix.

The first simulation was run with a vector of time-varying regression coefficients

$$\beta_t = \begin{bmatrix} 0.8 + 0.08 \cdot \sin\left(4\pi \cdot \frac{t}{500}\right) \\ 0.4 + 0.07 \cdot \cos\left(2\pi \cdot \frac{t}{500}\right) \\ 0.05 \cdot \cos\left(\pi \cdot \frac{t}{500}\right) \\ -0.25 + 0.1 \cdot \sin\left(3\pi \cdot \frac{t}{500}\right) \end{bmatrix}, \quad t = 1, \dots, 500, \tag{3.13}$$

and randomly generated regressors $x_t \sim U(0, 5)^4$. Fig. 3.7 shows the evolution of RMSE of all models and Fig. 3.8 shows the stability of their estimates from one randomly selected run. As we can see, the parameters vary relatively quickly, therefore the model with the forgetting factor $\alpha = 0.95$ clearly outperforms other models in terms of accuracy.

The second simulation was run with a vector of time-varying regression coefficients

$$\beta_t = \begin{bmatrix} 0.7 + 0.02 \cdot \sin\left(\pi \cdot \frac{t}{500}\right) \\ 0.5 + 0.018 \cdot \cos\left(2\pi \cdot \frac{t}{500}\right) \\ 0.017 \cdot \cos\left(\pi \cdot \frac{t}{500}\right) \\ -0.11 + 0.007 \cdot \sin\left(\pi \cdot \frac{t}{500}\right) \end{bmatrix}, \quad t = 1, \dots, 500, \tag{3.14}$$

and randomly generated regressors $x_t \sim U(0, 5)^4$. Fig. 3.9 shows the evolution of RMSE of all models and Fig. 3.10 shows the stability of their estimates from one randomly selected run. In this case, the parameters are more stable and vary relatively slowly. It is clear that the models with the forgetting factor close to 1 perform better, especially the model with $\alpha = 0.995$, while the usage of lower values results in overall worse quality of estimation.

The last simulation was run with a vector of constant regression coefficients

$$\beta_t = \begin{bmatrix} 0.9 \\ 0.4 \\ 0.1 \\ -0.2 \end{bmatrix}, \quad t = 1, \dots, 500, \tag{3.15}$$

and randomly generated regressors $x_t \sim U(0, 5)^4$. Fig. 3.11 shows the evolution of RMSE of all models and Fig. 3.12 shows the stability of their estimates

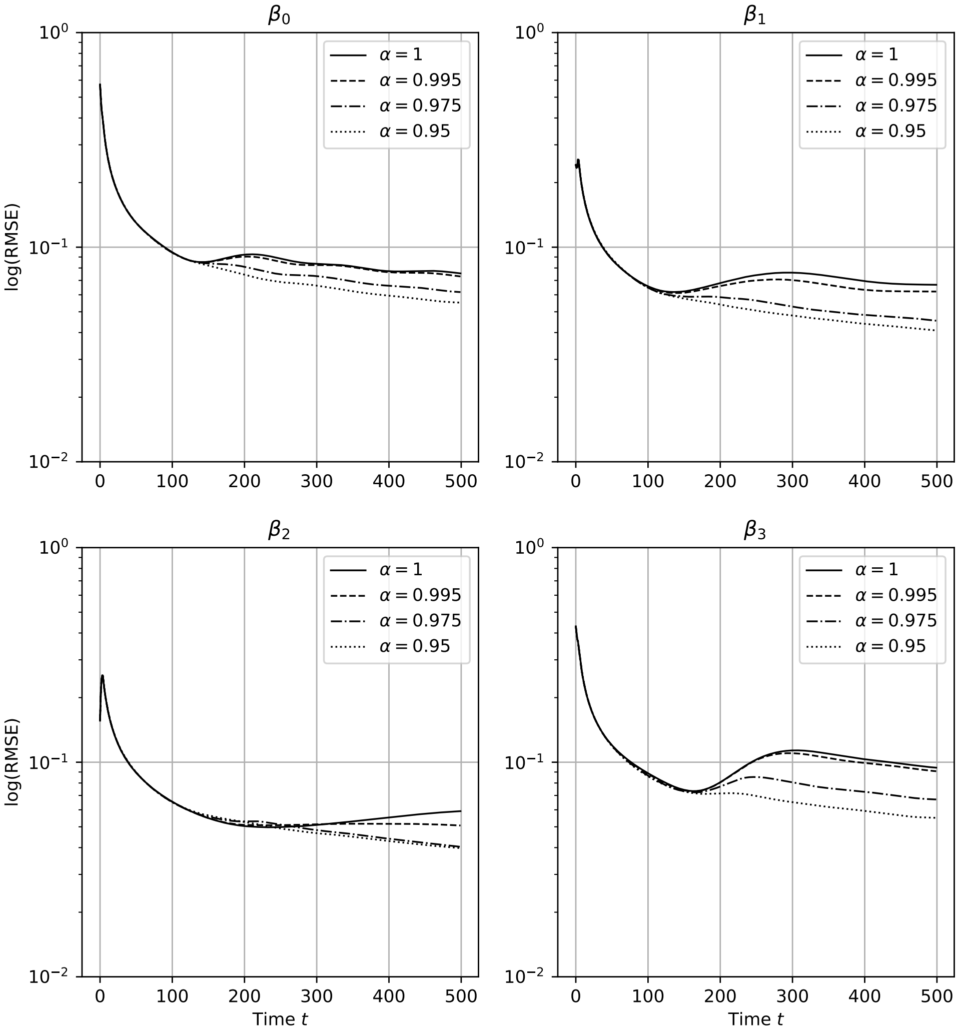

Figure 3.7: Evolution of the RMSE for high frequencies of time-varying parameters.

from one randomly selected run. Naturally, the model with $\alpha = 1$ has the best performance in this case. It is clear that lower values of the forgetting factor have considerably negative impact on both the stability of estimates and their accuracy.

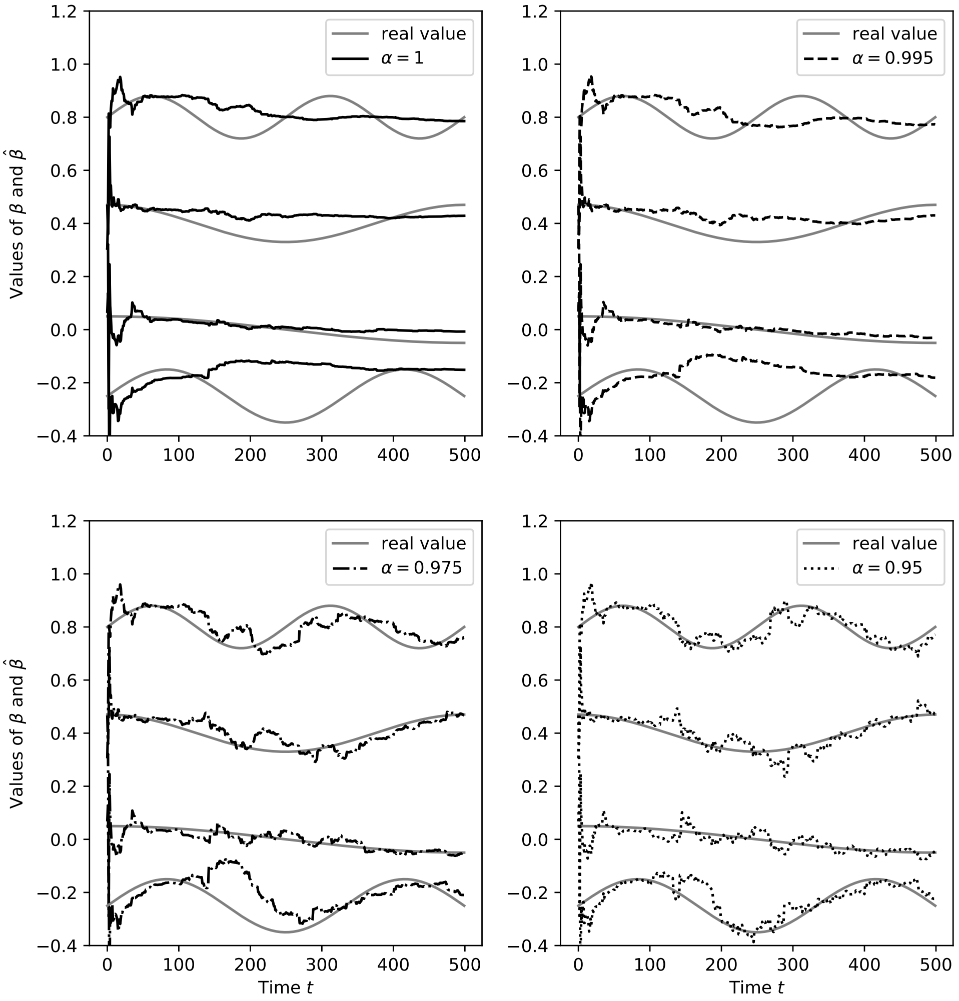

Figure 3.8: Real and estimated values of β in time for high frequencies of time-varying parameters.

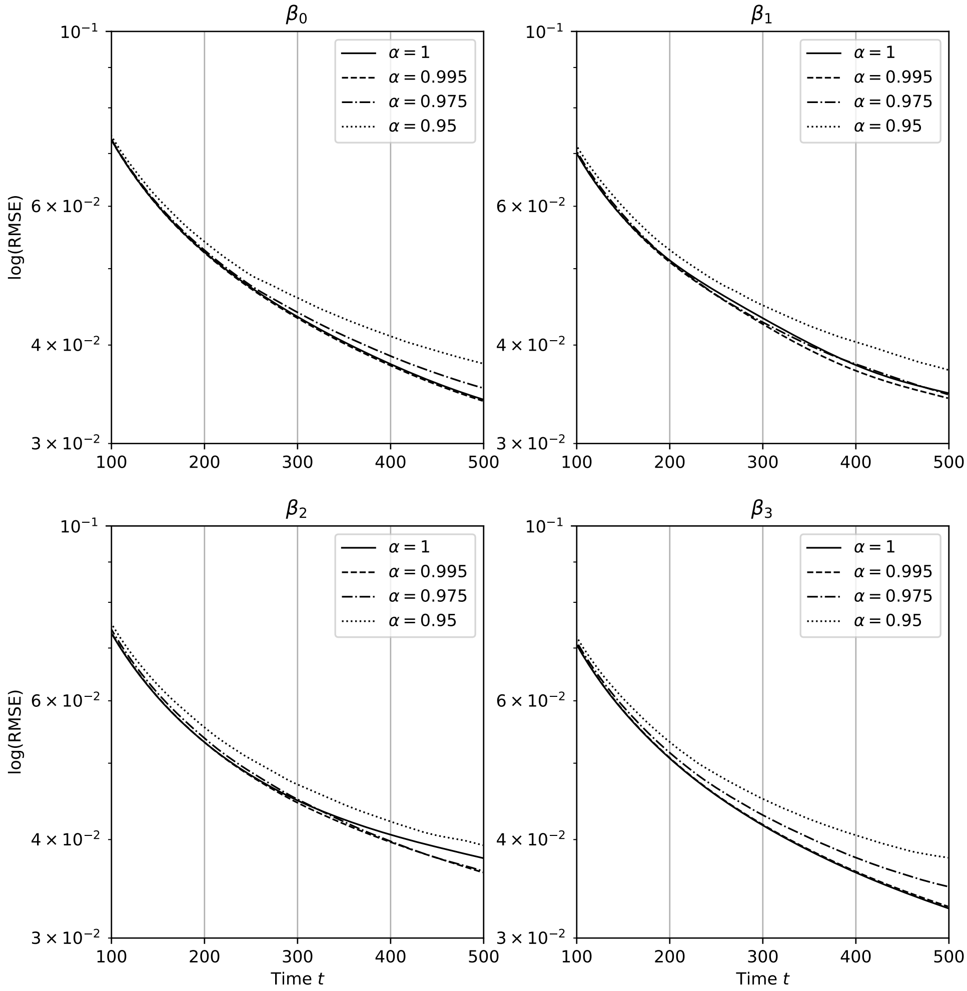

Figure 3.9: Evolution of the RMSE for low frequencies of time-varying parameters.

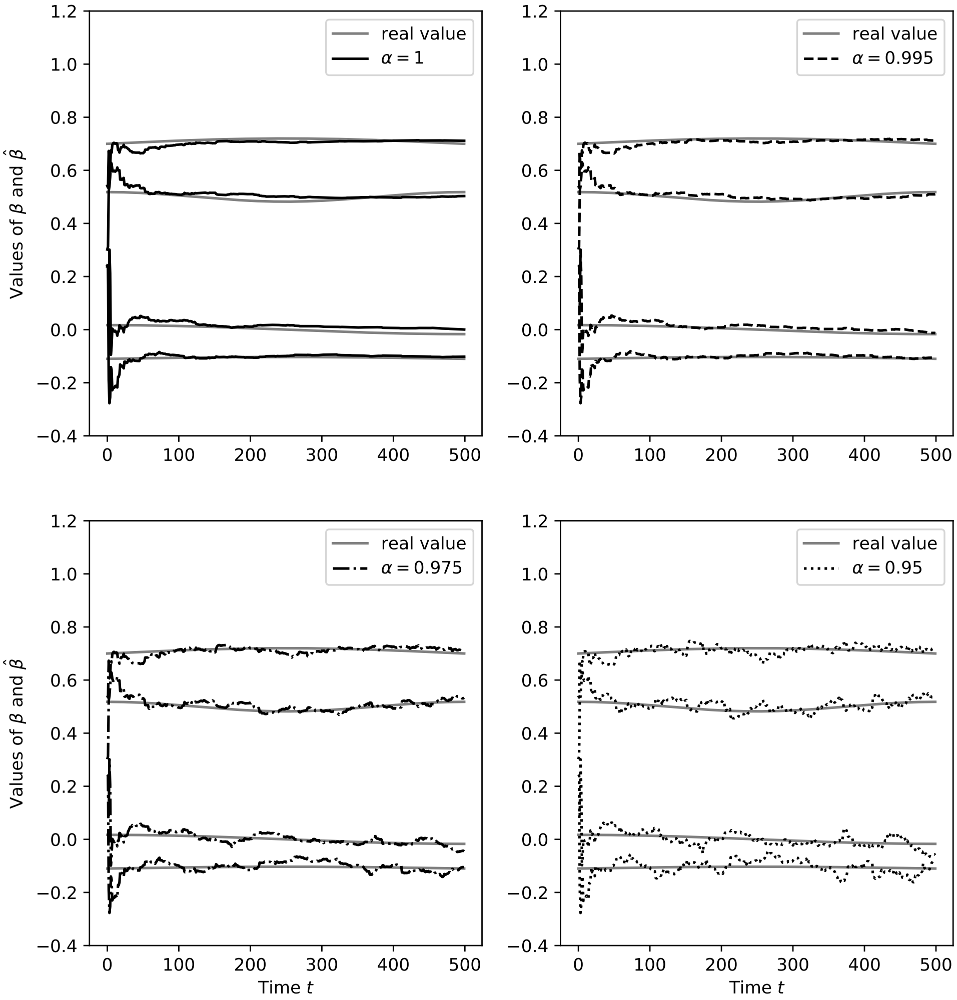

Figure 3.10: Real and estimated values of β in time for low frequencies of time-varying parameters.

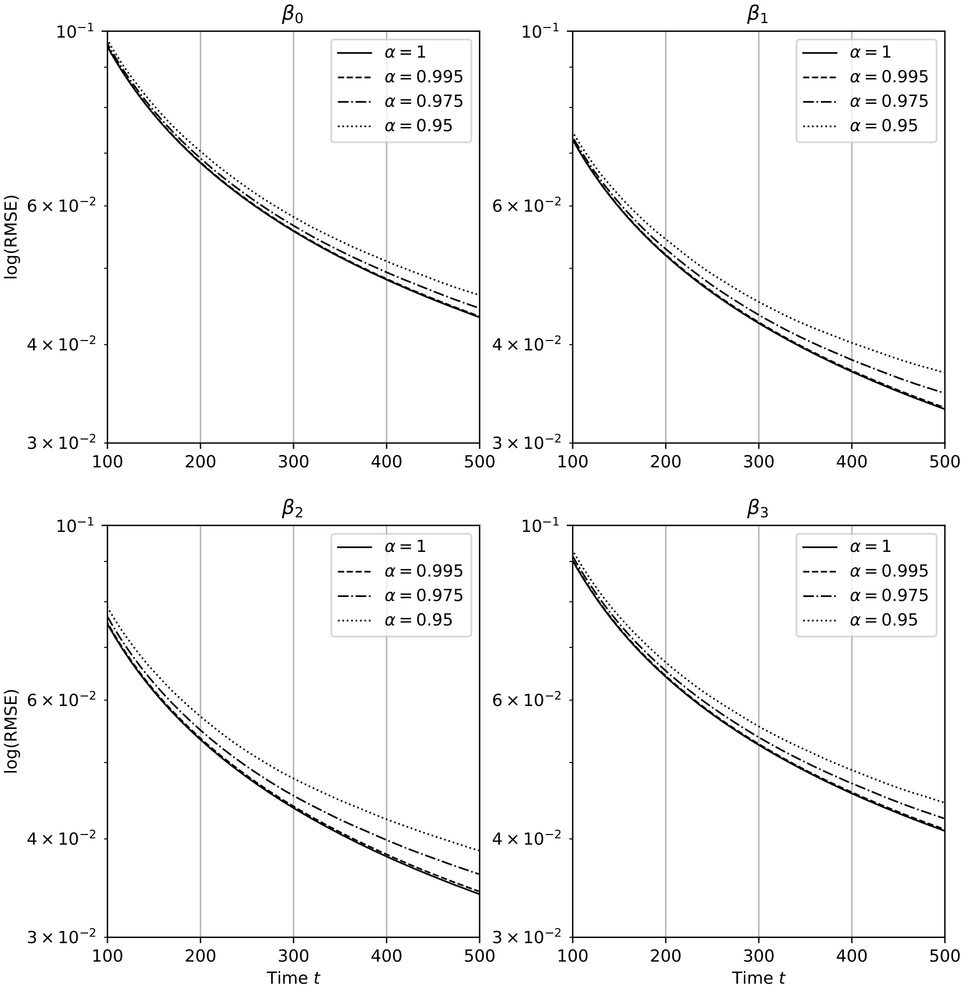

Figure 3.11: Evolution of the RMSE for constant parameters.

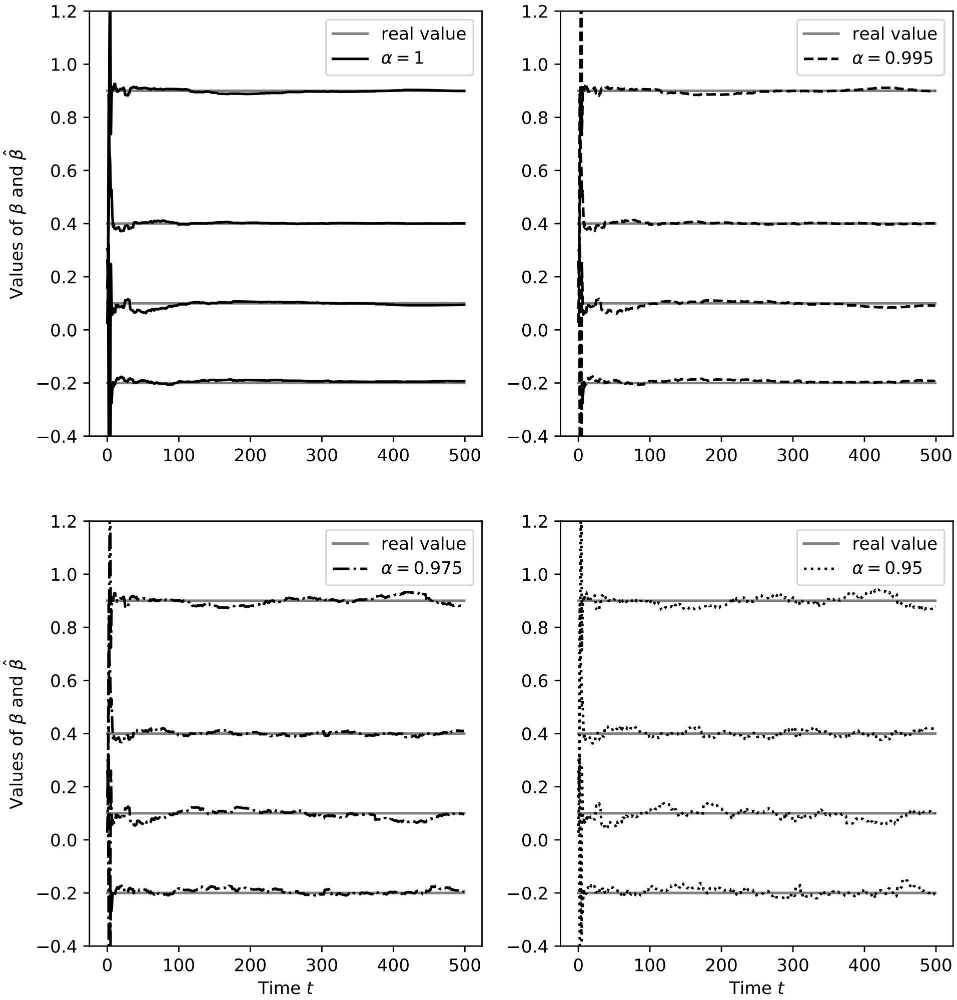

Figure 3.12: Real and estimated values of β in time for constant parameters.

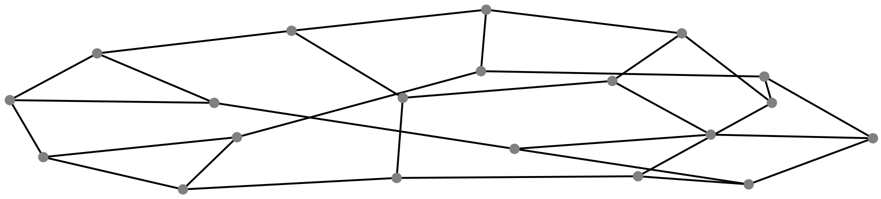

Figure 3.13: Network topology used in the first simulation.

3.3 Diffusion estimation with time-varying parameters

The final set of examples demonstrates the performance of the method proposed in Chapter 2. As the author is not aware of any alternative method for sequential modeling of counts in diffusion networks, two scenarios are compared: (i) the 'combination' scenario using Algorithm 2, and (ii) the isolated 'no combination' scenario.

Fig. 3.13 depicts the randomly generated diffusion network of 20 nodes with degree 3. They observe independently generated outcomes of a Poisson regression process simulated with a vector of time-varying regression coefficients

$$\beta_t = \begin{bmatrix} 0.7 + 0.075 \cdot \sin\left(3\pi \cdot \frac{t}{500}\right) \\ 0.5 + 0.05 \cdot \cos\left(2\pi \cdot \frac{t}{500}\right) \\ -0.2 \\ 0.05 \cdot \cos\left(\pi \cdot \frac{t}{500}\right) \end{bmatrix}, \quad t = 1, \dots, 500, \tag{3.16}$$

and randomly generated regressors $x_t^{(i)} \sim U(0,5)^4$. For all the nodes $i \in \{1, \dots, 20\}$, the initial prior distribution is the normal distribution with $b_0^{(i)} = [0,0,0,0]^\intercal$ and $P_0^{(i)} = 100 \cdot I$ where I is the identity 4×4 matrix. The forgetting factor $\alpha = 0.95$. The results are averaged over 100 independent runs.

Fig. 3.14 depicts the evolution of the RMSE averaged over all nodes. The distributed estimation clearly improves the estimation quality, especially in terms of the convergence rate. Note that when the estimates stabilize, the RMSE may slightly vary due to the time-varying nature of β_t. Fig. 3.15 shows a comparison of the stability of estimates at one randomly selected node of the network. The results show that the estimation performance of the proposed method is generally good. It can also be concluded that the estimates of the time-varying β_t are more stable in terms of smoothness, naturally at the cost of the communication overhead.

The second simulation demonstrates the effect of increasing the number of nodes in the network while maintaining the same degree of nodes. Fig. 3.16

depicts the randomly generated network of 50 nodes with degree 3. Fig. 3.17 depicts the RMSE evolution averaged over all nodes and Fig. 3.18 shows a comparison of the stability of estimates at one randomly selected node of the network. The results show no noticeable improvement over the smaller network.

The third simulation demonstrates the effect of increasing the degree of nodes. Fig. 3.19 depicts the randomly generated network of 20 nodes with degree 6. Fig. 3.20 depicts the RMSE evolution averaged over all nodes and Fig. 3.21 shows a comparison of the stability of estimates at one randomly selected node of the network. In this case, the improvement in terms of both the RMSE and the smoothness of estimates is obvious.

The last simulation shows similar results. Fig. 3.22 depicts the randomly generated network of 50 nodes with degree 6. Fig. 3.23 depicts the RMSE evolution averaged over all nodes and Fig. 3.24 shows a comparison of the stability of estimates at one randomly selected node of the network. From these observations, it can be concluded that with the increasing degree of nodes, the accuracy of estimates also increases.

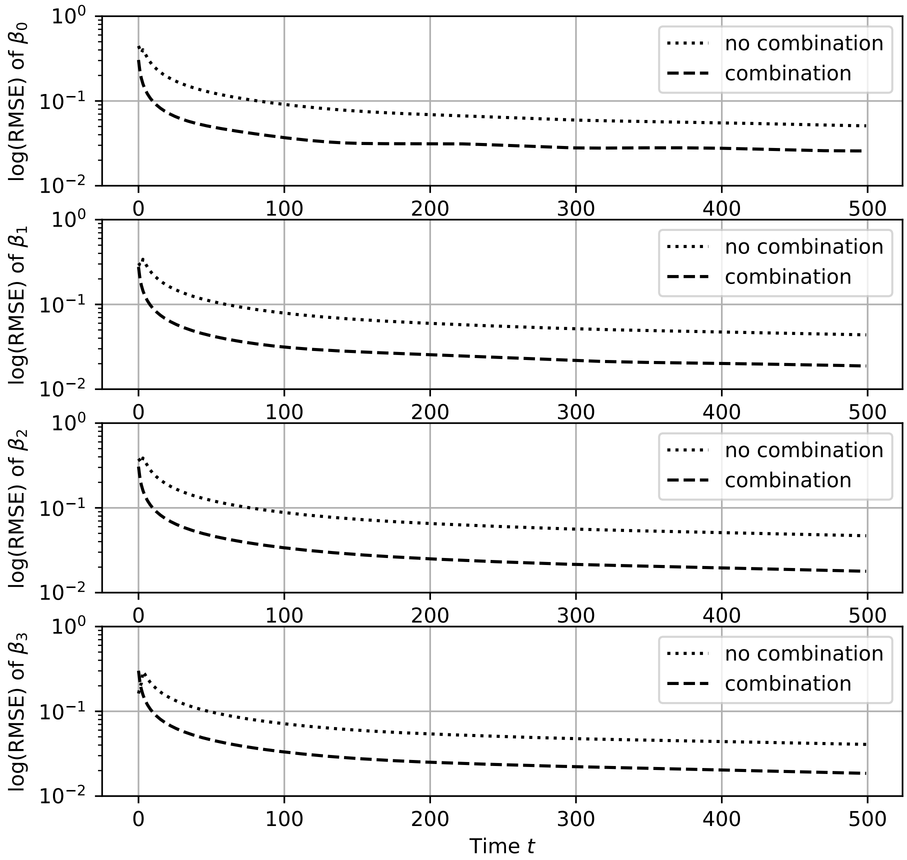

Figure 3.14: Evolution of the RMSE averaged over all 20 network nodes with degree 3.

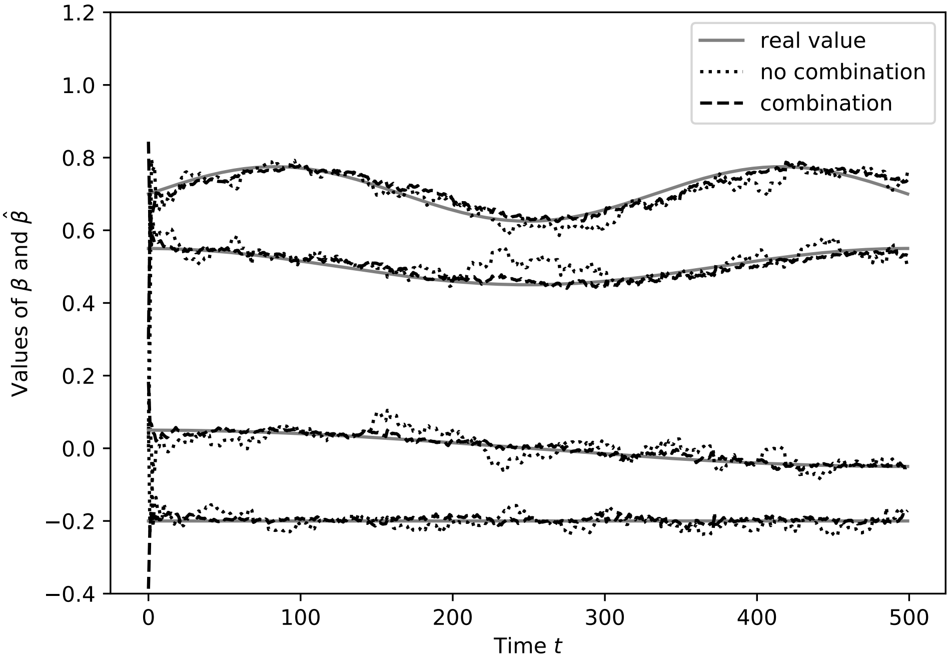

Figure 3.15: Real and estimated values of β in time in a single node of the network of 20 nodes with degree 3.

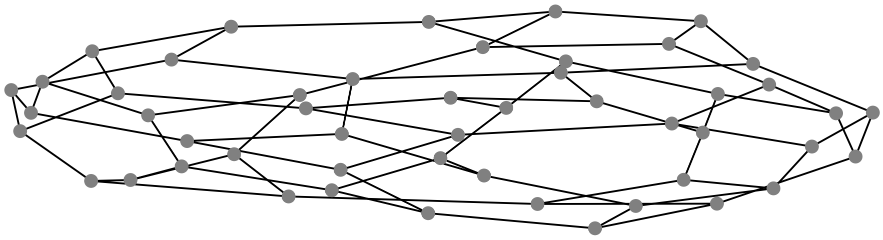

Figure 3.16: Network topology used in the second simulation.

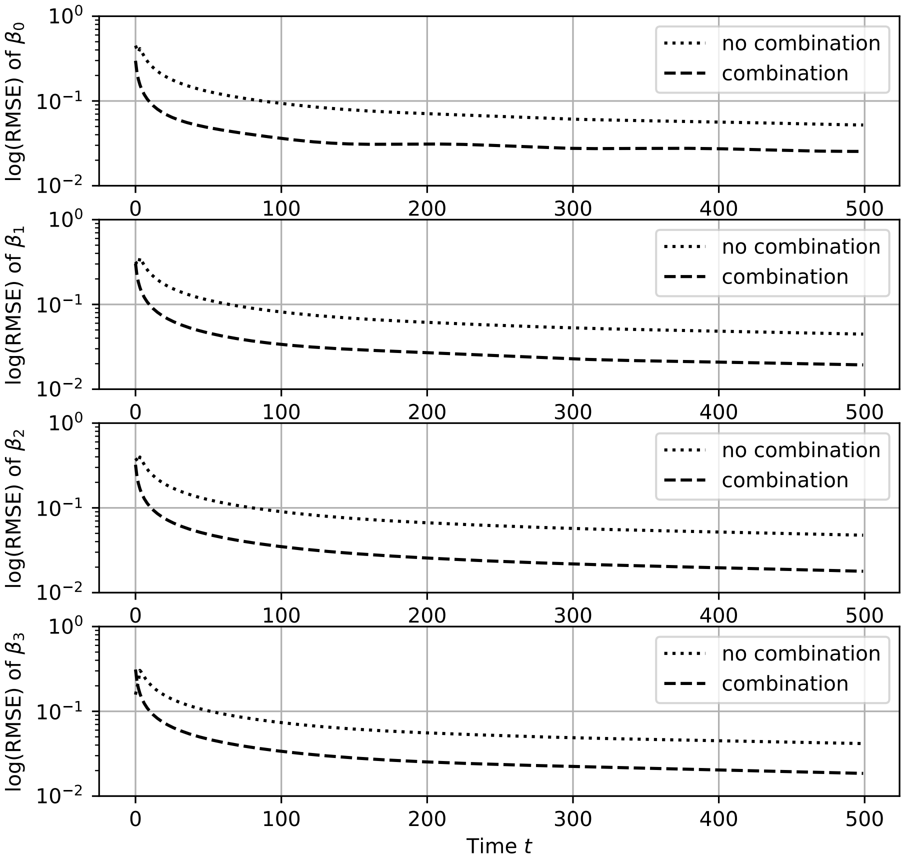

Figure 3.17: Evolution of the RMSE averaged over all 50 network nodes with degree 3.

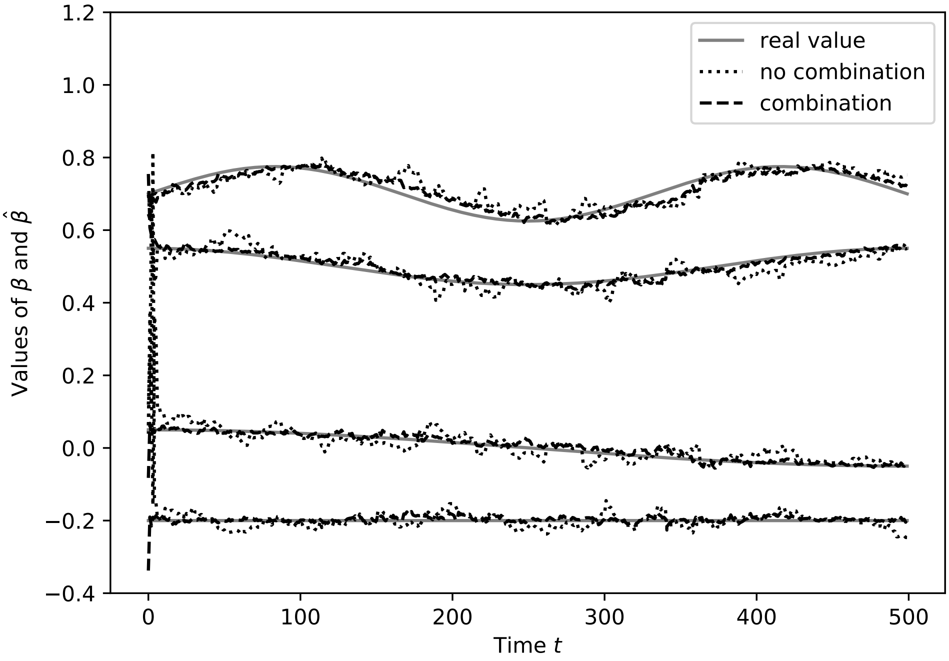

Figure 3.18: Real and estimated values of β in time in a single node of the network of 50 nodes with degree 3.

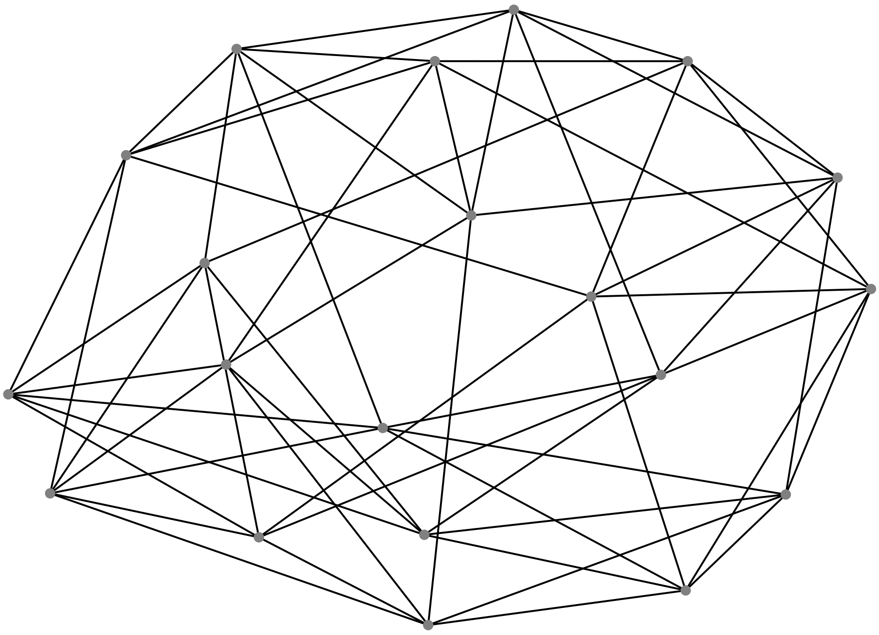

Figure 3.19: Network topology used in the third simulation.

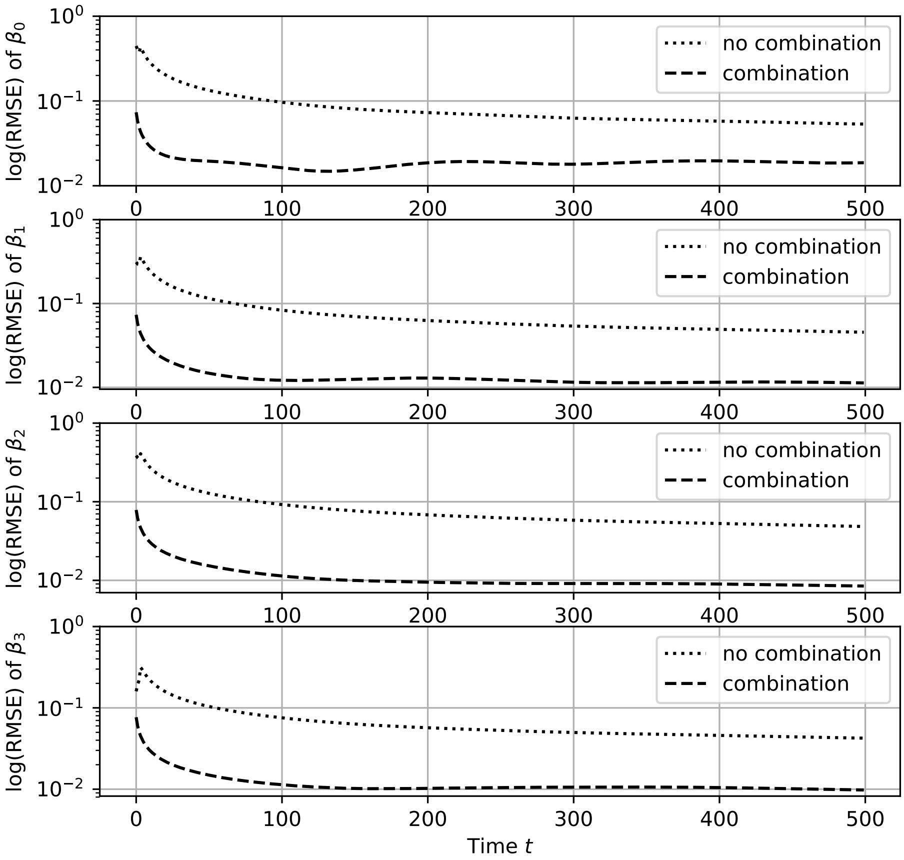

Figure 3.20: Evolution of the RMSE averaged over all 20 network nodes with degree 6.

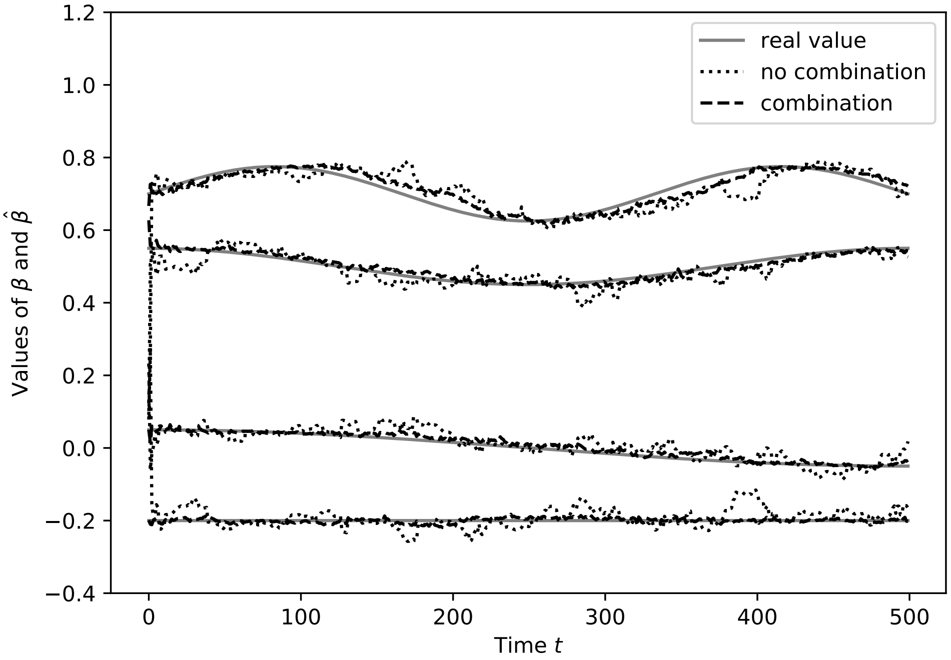

Figure 3.21: Real and estimated values of β in time in a single node of the network of 20 nodes with degree 6.

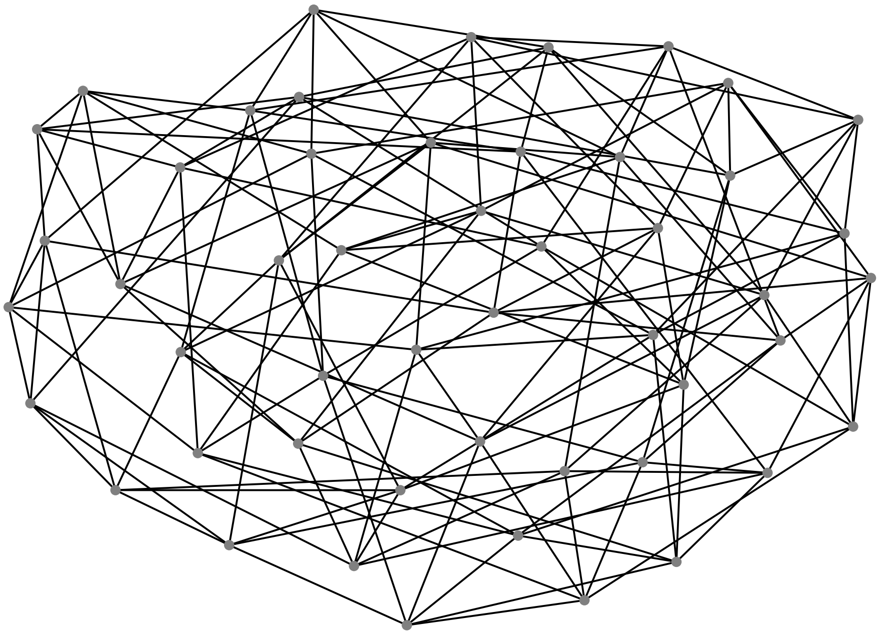

Figure 3.22: Network topology used in the fourth simulation.

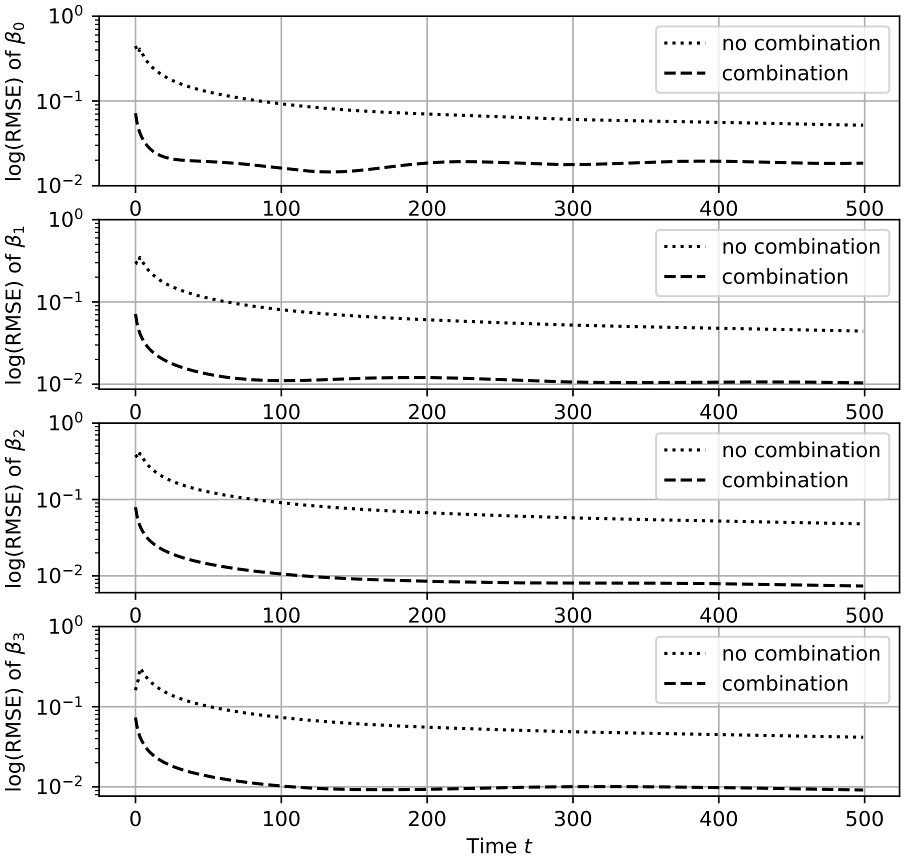

Figure 3.23: Evolution of the RMSE averaged over all 50 network nodes with degree 6.

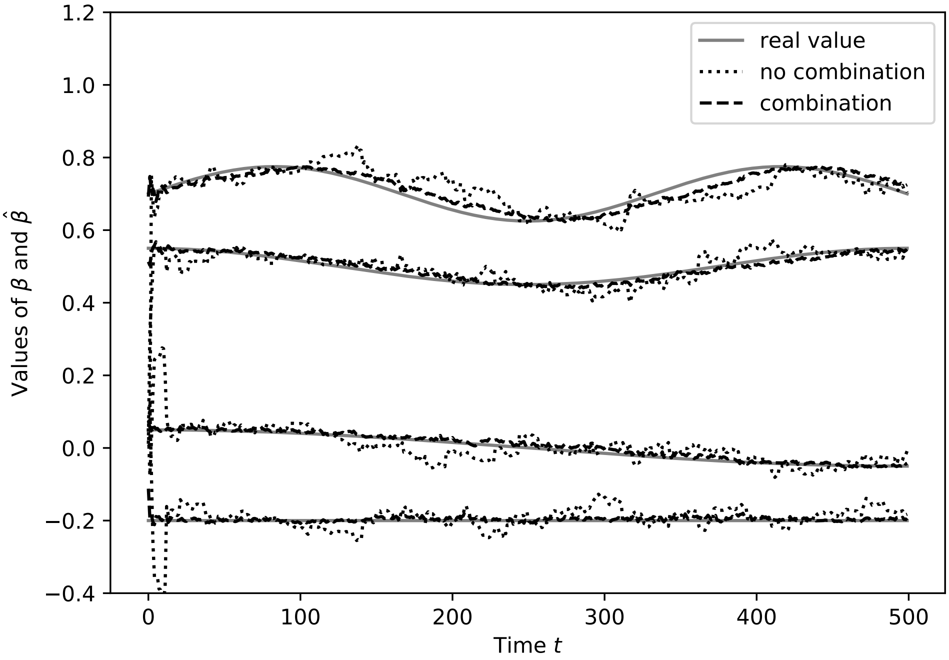

Figure 3.24: Real and estimated values of β in time in a single node of the network of 50 nodes with degree 6.

Chapter 4

Future Work

This thesis focuses on the sequential inference of the standard Poisson model and its application in a distributed environment. However, there are several topics that can be discussed in future work.

4.1 Models of counts

One of the topics is the zero-inflated Poisson regression, which is used for modeling count data with excess zeros, e.g., number of defects in cases where the manufacturing equipment is misaligned. It assumes that with probability p the only possible observation is 0, and with probability $1-p$, a standard Poisson random variable is observed [42, 43]. Let $Y = [Y_1, \dots, Y_n]^\intercal$ be a vector of independent responses. Then

$$\begin{array}{ll} Y_i \sim 0 & \text{with probability } p_i, \\ Y_i \sim Poisson(\lambda_i) & \text{with probability } 1-p_i, \end{array} \tag{4.1}$$

so that

$$Y_i = \begin{cases} 0 & \text{with probability } p_i + (1-p_i)e^{-\lambda_i}, \\ k & \text{with probability } (1-p_i)e^{-\lambda_i}\frac{\lambda_i^k}{k!}, \quad k = 1, 2, \dots \end{cases} \tag{4.2}$$

Another topic which can be discussed is overdispersion [5]. The Poisson model is equidispersed, meaning the mean and the variance have the same value. In many cases this is very limiting, as the actual random variable is overdispersed and the variance may differ from the mean. One of the models that solve this problem is the negative binomial model. The most common implementation is the NB2 model [44]. Its pdf reads as

$$f(y|\mu,\alpha) = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}, \tag{4.3}$$

where $\alpha \geq 0$ and $y = 0, 1, 2, \ldots$ Note that the negative binomial distribution reduces to the Poisson distribution if $\alpha = 0$.

4.2 Distributed estimation

In Chapter 2, an algorithm for diffusion Poisson regression was presented. However, it only utilizes the combination phase of the estimation by diffusion presented in [24], and skips the adaptation phase. During the adaptation phase, every node in the network gathers observations from all of its neighbors and performs the Bayesian update similar to Equation (1.72),

$$\Xi_t^{(i)} = \Xi_{t-1}^{(i)} + \sum_{j \in \mathcal{I}^{(i)}} c_{i,j} T(x_t^{(i)}, \widetilde{y}_t^{(i)}), \tag{4.4}$$

where $c_{i,j}$ are adaptation weights assigned to neighbors of the node i. If the observation is considered to be an outlier, then $c_{i,j} = 0$. Otherwise $c_{i,j} = 1$. An algorithm which makes use of the adaptation phase, resulting in the full ATC (adapt-then-combine) scenario, can be devised in the future [10, 11].

Another interesting topic is the estimation of heterogeneous parameters, meaning every node observes slightly (or completely) different process. In an isolated scenario, this is naturally not a problem. However, in the case of cooperation, the complexity is significantly high under the lack of knowledge which parameters are shared and among which agents. In [25], a framework for estimation of heterogeneous parameters in diffusion networks is presented, and a simulated example shows that the collaboration improves estimation performance of both the shared and strictly local parameters.

Conclusion

The aim of this thesis was to compile an overview of the GLMs and the Poisson regression model, focus on El-Sayyad's approach to its Bayesian estimation and propose a sequential variant. Another objective was to propose methods for stabilization of the estimation procedure and study their behavior on convenient examples, and, if possible, suggest a use case of the proposed modeling approach in the signal processing domain.

In Chapter 1, an overview of GLMs and the Poisson regression model was elaborated and a method for stabilization of the estimation was proposed. Then, an algorithm for the sequential estimation was devised. In Chapter 2, a method for sequential distributed modeling of counts using the Poisson model was proposed. The parameters are locally estimated using a calibrated stabilized estimation procedure. Then, the posterior pdfs are combined in the network. In Chapter 3, several sets of simulation examples were presented to demonstrate the efficiency of the proposed methods and the effect of different hyperparameter values and network configurations on the estimation quality. Finally, Chapter 4 discusses a few interesting topics that can be explored in the future, such as the zero inflation or overdispersion [43, 5] and the full ATC diffusion strategy [10, 11].

Bibliography

[1] Bosowski, N.; Ingle, V.; et al. Generalized Linear Models for count time series. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4272–4276.

[2] Wang, L.; Chi, Y. Stochastic Approximation and Memory-Limited Subspace Tracking for Poisson Streaming Data. *IEEE Transactions on Signal Processing*, volume 66, no. 4, 2018: pp. 1051–1064.

[3] Manolakis, D.; Bosowski, N.; et al. Count Time-Series Analysis: A Signal Processing Perspective. *IEEE Signal Processing Magazine*, volume 36, no. 3, 2019: pp. 64–81.

[4] McCullagh, P.; Nelder, J. A. *Generalized Linear Models, Second Edition (Monographs on Statistics & Applied Probability)*. Chapman and Hall, second edition, Aug. 1989, ISBN 0412317605.

[5] Myers, R. H.; Montgomery, D. C.; et al. *Generalized linear models (Wiley Series in Probability and Statistics)*. John Wiley & Sons, Mar. 2010, ISBN 9780470454633.

[6] Cintuglu, M. H.; Ishchenko, D. Secure Distributed State Estimation for Networked Microgrids. *IEEE Internet of Things Journal*, volume 6, no. 5, 2019: pp. 8046–8055.

[7] Ghazanfari-Rad, S.; Labeau, F. Formulation and analysis of LMS adaptive networks for distributed estimation in the presence of transmission errors. *IEEE Internet of Things Journal*, volume 3, no. 2, 2015: pp. 146–160.

[8] Ratner, B. *Statistical and Machine-Learning Data Mining, Third Edition: Techniques for Better Predictive Modeling and Analysis of Big Data, Third Edition*. Chapman & Hall/CRC, third edition, 2017, ISBN 1498797601.

[9] Chen, Y.; Kar, S.; et al. The internet of things: Secure distributed inference. *IEEE Signal Processing Magazine*, volume 35, no. 5, 2018: pp. 64–75.

[10] Sayed, A. H. Diffusion adaptation over networks. In *Academic Press Library in Signal Processing*, volume 3, Elsevier, 2014, pp. 323–453.

[11] Sayed, A. H.; et al. Adaptation, learning, and optimization over networks. *Foundations and Trends® in Machine Learning*, volume 7, no. 4-5, 2014: pp. 311–801.

[12] Cattivelli, F. S.; Lopes, C. G.; et al. Diffusion recursive least-squares for distributed estimation over adaptive networks. *IEEE Transactions on Signal Processing*, volume 56, no. 5, 2008: pp. 1865–1877.

[13] Cattivelli, F. S.; Sayed, A. H. Diffusion LMS Strategies for Distributed Estimation. *IEEE Transactions on Signal Processing*, volume 58, no. 3, 2010: pp. 1035–1048.

[14] Plata-Chaves, J.; Bogdanović, N.; et al. Distributed Diffusion-Based LMS for Node-Specific Adaptive Parameter Estimation. *IEEE Transactions on Signal Processing*, volume 63, no. 13, 2015: pp. 3448–3460.

[15] Plata-Chaves, J.; Bahari, M. H.; et al. Unsupervised diffusion-based LMS for node-specific parameter estimation over wireless sensor networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4159–4163.

[16] Huang, W.; Yang, X.; et al. Diffusion LMS with component-wise variable step-size over sensor networks. *IET Signal Processing*, volume 10, no. 1, 2016: pp. 37–45.

[17] Cattivelli, F. S.; Sayed, A. H. Diffusion Strategies for Distributed Kalman Filtering and Smoothing. *IEEE Transactions on Automatic Control*, volume 55, no. 9, 2010: pp. 2069–2084.

[18] Hu, J.; Xie, L.; et al. Diffusion Kalman Filtering Based on Covariance Intersection. *IEEE Transactions on Signal Processing*, volume 60, no. 2, 2012: pp. 891–902.

[19] Dias, S. S.; Bruno, M. G. S. Distributed Bernoulli Filters for Joint Detection and Tracking in Sensor Networks. *IEEE Transactions on Signal and Information Processing over Networks*, volume 2, no. 3, 2016: pp. 260–275.

[20] Bruno, M. G.; Dias, S. S. Collaborative emitter tracking using Rao-Blackwellized random exchange diffusion particle filtering. *Eurasip Journal on Advances in Signal Processing*, volume 2014, no. 1, 2014: p. 19.

[21] Dedecius, K.; Djurić, P. M. Diffusion filtration with approximate Bayesian computation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 3207–3211.

[22] Li, W.; Wang, Z.; et al. Particle filtering with applications in networked systems: a survey. *Complex & Intelligent Systems*, volume 2, no. 4, 2016: pp. 293–315.

[23] Dedecius, K.; Reichl, J.; et al. Sequential Estimation of Mixtures in Diffusion Networks. *IEEE Signal Processing Letters*, volume 22, no. 2, 2015: pp. 197–201.

[24] Dedecius, K.; Djurić, P. M. Sequential Estimation and Diffusion of Information Over Networks: A Bayesian Approach With Exponential Family of Distributions. *IEEE Transactions on Signal Processing*, volume 65, no. 7, 2017: pp. 1795–1809.

[25] Dedecius, K.; Sečkárová, V. Factorized Estimation of Partially Shared Parameters in Diffusion Networks. *IEEE Transactions on Signal Processing*, volume 65, no. 19, 2017: pp. 5153–5163.

[26] El-Sayyad, G. Bayesian and classical analysis of Poisson regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, volume 35, no. 3, 1973: pp. 445–451.

[27] Dedecius, R., Kamil; Žemlička. Sequential Poisson Regression in Diffusion Networks. *IEEE Signal Processing Letters*, volume 27, no. 1, 2020: pp. 625–629.

[28] Montgomery, D. *Introduction to linear regression analysis*. Hoboken, NJ: Wiley A John Wiley & Sons, Inc, 2012, ISBN 0470542810.

[29] Tierney, L.; Kadane, J. B. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, volume 81, no. 393, 1986: pp. 82–86.

[30] Myung, I. J. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, volume 47, no. 1, 2003: pp. 90–100.

[31] Bartlett, M. S.; Kendall, D. G. The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Supplement to the Journal of the Royal Statistical Society*, volume 8, no. 1, 1946: pp. 128–138.

[32] Raiffa, H.; Schlaifer, R. *Applied statistical decision theory*. Harvard University Press, 1961.

[33] Haight, F. A. *Handbook of the Poisson distribution*. Wiley, 1967.

[34] Cramér, H. Mathematical Methods of Statistics. *Princeton University Press., Princeton, N.J.*, 1946.

[35] Curtiss, J. H. On transformations used in the analysis of variance. *The Annals of Mathematical Statistics*, volume 14, no. 2, 1943: pp. 107–122.

[36] Bartlett, M. S. The Use of Transformations. *Biometrics*, volume 3, no. 1, 1947: pp. 39–52, ISSN 0006341X, 15410420. Available from: `http://www.jstor.org/stable/3001536`

[37] Silverman, B. W. *Density estimation for statistics and data analysis*. London New York: Chapman and Hall, 1986, ISBN 9780412246203.

[38] Fink, D. A Compendium of Conjugate Priors. 1997. Available from: `https://www.johndcook.com/CompendiumOfConjugatePriors.pdf`

[39] Peterka, V. Chapter 8 - BAYESIAN APPROACH TO SYSTEM IDENTIFICATION. In *Trends and Progress in System Identification*, edited by P. EYKHOFF, Pergamon, 1981, ISBN 978-0-08-025683-2, pp. 239 – 304, doi:https://doi.org/10.1016/B978-0-08-025683-2.50013-2. Available from: `http://www.sciencedirect.com/science/article/pii/B9780080256832500132`

[40] Dedecius, K.; Nagy, I.; et al. Parameter tracking with partial forgetting method. *International Journal of Adaptive Control and Signal Processing*, volume 26, no. 1, 2012: pp. 1–12.

[41] Jin, D.; Chen, J.; et al. Affine Combination of Diffusion Strategies Over Networks. *IEEE Transactions on Signal Processing*, volume 68, 2020: pp. 2087–2104.

[42] Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, volume 34, no. 1, 1992: pp. 1–14.

[43] Hall, D. B. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, volume 56, no. 4, 2000: pp. 1030–1039.

[44] Cameron, A. C.; Trivedi, P. K. *Regression Analysis of Count Data*. Econometric Society Monographs, Cambridge University Press, second edition, 2013, doi:10.1017/CBO9781139013567.

Appendix A

Acronyms

ATC adapt-then-combine.

GLM generalized linear model.

IoT Internet of Things.

KDE kernel density estimation.

LMS least mean squares.

MCMC Markov chain Monte Carlo.

MLE maximum likelihood estimation.

OLS ordinary least squares.

pdf probability density function.

RLS recursive least squares.

RMSE root mean square error.

Appendix B

Contents of enclosed CD

```
readme.txt ........................ the file with CD contents description
src ..................................... the directory of source codes
  implementation .............. implementation sources of simulations
  thesis ............... the directory of LaTeX source codes of the thesis
text .......................................... the thesis text directory
  thesis.pdf .......................... the thesis text in PDF format
```