Sequential Poisson Regression in Diffusion Networks

Kamil Dedecius and Radomír Žemlička

***Abstract*—The Poisson regression is a popular model for positive integer random variables determined by known explanatory variables. This letter studies the problem of its collaborative Bayesian *sequential* estimation under potentially slowly time-varying regression coefficients. We assume networks where agents share their information about the inferred quantities with adjacent neighbors in order to improve the overall estimation performance. The communication strategy is the information diffusion, i.e., only one information exchange per time instant is allowed.**

***Index Terms*—Diffusion, distributed estimation, collaborative estimation, Poisson regression.**

I. Introduction

DISTRIBUTED inference of unknown variables in networks of collaborating agents has become an established discipline in the signal processing domain. It finds applications in sensor networks, smart grids and microgrids, Internet of Things, social networks, and other networked systems [1]–[3].

Generally, three communication and information processing strategies can be distinguished: the incremental strategy, consensus, and diffusion [4], [5]. In this letter, we focus on the diffusion strategy, where the information exchange runs on a single time scale and within one network hop distance [6]. Many popular sequential inference algorithms have found their more or less modified diffusion counterparts. To name only a few: the LMS [2], [7]–[10], RLS [6], Kalman filters [11], [12], Bernoulli filters [13], particle filters [14]–[16], or the quasi-Bayesian mixture estimation algorithm [17]. A unifying Bayesian framework for diffusion inference of a wide class of models was designed in [18] and [19].

This letter focuses 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 [20], [21]. Generally, high counts can be approximated by continuous data models. However, these fail if the counts are small and include many zeros [22]. In particular, we focus on the Poisson regression model, propose its low-cost real-time sequential estimation to deal with streaming data, and devise a method for its distributed inference in networks of collaborating agents (sensors). We are not aware of any existing sequential distributed or non-distributed alternative. The non-distributed Poisson models rely on computationally intensive optimization techniques [23], [24], complicating their use in online tasks.

We face several major difficulties. First, the sequential estimation of the Poisson model is generally impossible due to its functional form. We show that a way towards the solution provides the Bayesian paradigm and a couple of approximations providing stable and analytically tractable results. There are several novel points in the proposed algorithm. We recast the static Bayesian Poisson regression [25] into an algorithm capable of *online* estimation of potentially slowly time-varying regression coefficients. Then, a rule for the combination of these estimates in diffusion networks is devised.

Manuscript received May 14, 2020; revised May 14, 2020. First published May 14, 2020; current version published May 14, 2020. The work of R. Žemlička was supported by the internal grant program 'VýLet' of the Faculty of Information Technology of the Czech Technical University in Prague. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wei Li.

K. Dedecius is with the Czech Academy of Sciences, the Institute of Information Theory and Automation, Pod Vodárenskou věží 4, 182 00 Prague 8, Czech Republic. R. Žemlička is with the Faculty of Information Technology, Czech Technical University in Prague, Thákurova 9, 160 00 Prague 6, Czech Republic (email: dedecius@utia.cas.cz, zemlirad@fit.cvut.cz).

Digital Object Identifier XXX/XXX

II. Sequential Inference of the Poisson Model

We consider a discrete-time modeling of a stochastic process $\{Y_t; t = 1, 2, \ldots\}$ with mutually independent observations $y_t \in \mathbb{N}$. The random variables Y_t are 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 generalized linear model (GLM) [23]

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

where $g(\cdot)$ is a known link function, and the product $\beta_t^\intercal x_t$ is called the linear predictor. Recall, that the identity function $g(\cdot)$ provides the ordinary linear regression model. In our case of $y_t \in \mathbb{N}$, the role of the link function plays the natural logarithm,

$$g(\mathbb{E}[Y_t|x_t, \beta_t]) = \log(\mathbb{E}[Y_t|x_t, \beta_t]) = \beta_t^\intercal x_t, \tag{2}$$

resulting in the Poisson regression model

$$Y_t \sim Po(\lambda_t) = Po\left(g^{-1}(\beta_t^\intercal x_t)\right) = Po\left(\exp\left(\beta_t^\intercal x_t\right)\right). \tag{3}$$

The expected value and the variance are

$$\mathbb{E}[Y_t|x_t, \beta_t] = \mathrm{var}(Y_t|x_t, \beta_t) = \lambda_t = \exp(\beta_t^\intercal x_t). \tag{4}$$

The probability density function (pdf) of the model reads

$$f(y_t|x_t, \beta_t) = \frac{\lambda_t^{y_t} e^{-\lambda_t}}{y_t!} = \frac{e^{\beta_t^\intercal x_t y_t} e^{-\exp(\beta_t^\intercal x_t)}}{y_t!}. \tag{5}$$

Except for the linear regression model, direct Bayesian inference of GLMs is analytically intractable due to the lack of convenient conjugate prior distributions. The inference thus mostly relies on MCMC methods [24], which are prohibitive

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 [26], or the three-stage approximation Poisson $\rightarrow$ loggamma $\rightarrow$ normal pdf in the static Poisson model [25]. This idea will be adopted below to propose the sequential estimator.

A. Stabilization of variance

A frequently encountered problem of the Poisson models evident from (4) is that the variance varies with the expected value. Therefore, a variance-stabilizing transformation by a convenient function $h(\cdot)$ may improve the estimation quality. It is well known that the square-root transform of the Poisson variable $Y_t \sim Po(\lambda_t)$ yields a variable that is approximately normal and has a constant variance, $\widetilde{Y}_t = \sqrt{Y_t} \sim \mathcal{N}(\sqrt{\lambda_t}, \frac{1}{4})$ [27], [28]. Although there are several other favored (and sometimes better) transformations of the Poisson variable, summarized, e.g., in [23] or [29], we will stick with the square root for its general simplicity and effectiveness.

The pdf of $\widetilde{Y}_t$ due to the change of variables theorem is

$$f(\widetilde{y}_t|x_t,\beta_t) = f\left(h^{-1}(\widetilde{y}_t)\right)\left|\frac{\mathrm{d}h^{-1}(\widetilde{y}_t)}{\mathrm{d}\widetilde{y}_t}\right| \tag{6}$$

$$= \frac{\lambda_t^{\widetilde{y}_t^2} e^{-\lambda_t}}{\widetilde{y}_t^2!} \cdot 2\widetilde{y}_t \tag{7}$$

$$= \frac{e^{\beta_t^\intercal x_t \widetilde{y}_t^2} e^{-\exp(\beta_t^\intercal x_t)}}{\widetilde{y}_t^2!} \cdot 2\widetilde{y}_t \tag{8}$$

$$= \frac{2}{\Gamma(\widetilde{y}_t^2)} e^{\beta_t^\intercal x_t \widetilde{y}_t^2} e^{-\exp(\beta_t^\intercal x_t)} \cdot \frac{1}{\widetilde{y}_t}, \tag{9}$$

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

Although the variance-stabilized variable $\widetilde{Y}_t$ is approximately normal, the problem of nonexistent conjugate prior persists. Hence instead of using the normal pdf, we will stick with the functional form (9) in the subsequent steps.

B. Approximate sequential estimation of β

Let us now devise the Bayesian sequential estimator. We introduce the prior pdf $\pi(\beta_t|x_{0:t-1},\widetilde{y}_{0:t-1})$ where $x_{0:t-1} = [x_0,\dots,x_{t-1}]$ and $\widetilde{y}_{0:t-1} = [\widetilde{y}_0,\dots,\widetilde{y}_{t-1}]$. It 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 symbolize the knowledge (pseudo-observations) available at the very beginning of the modeling, e.g., from historical observations, or given by an expert.

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

$$\pi(\beta_t|x_{0:t},\widetilde{y}_{0:t}) = \mathcal{C}\cdot f(\widetilde{y}_t|x_t,\beta_t)\pi(\beta_t|x_{0:t-1},\widetilde{y}_{0:t-1}), \tag{10}$$

where $\mathcal{C}$ is the normalizing constant ensuring that the resulting left-hand side posterior function is a pdf that integrates to one.

Generally, the Bayesian update (10) does not yield posterior distributions in closed forms. An important exception in this respect is the case of models belonging to the exponential family of distributions estimated with conjugate prior distributions [30]. In particular, an exponential family distribution of a random variable $\widetilde{Y}_t$ with a parameter β_t is characterized by a pdf of the (non-unique) form

$$f(\widetilde{y}_t|x_t,\beta_t) = k(x_t,\widetilde{y}_t)l(\beta_t)e^{\eta(\beta_t)^\intercal T(x_t,\widetilde{y}_t)}, \tag{11}$$

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 encompasses 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. The prior distribution for the estimation of β_t conjugate to the model (11) 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

$$\pi(\beta_t|\Xi_{t-1},\nu_{t-1}) = m(\Xi_{t-1},\nu_{t-1})l(\beta_t)^{\nu_{t-1}}e^{\eta(\beta_t)^\intercal \Xi_{t-1}}, \tag{12}$$

where $m(\Xi_{t-1},\nu_{t-1})$ is a known function, and $l(\beta_t)$ is the same function as in (11). The Bayesian update (10) multiplying the model (11) with the prior pdf (12) then results in the posterior pdf of the same functional type as the prior, characterized by the posterior hyperparameters

$$\begin{aligned}\Xi_t &= \Xi_{t-1} + T(x_t,\widetilde{y}_t),\\ \nu_t &= \nu_{t-1} + 1.\end{aligned} \tag{13}$$

We stress that this result allows for efficient sequential estimation of β_t, as the Bayesian update (10) is equivalent to simple summations, and the functional form of the posterior density is the same as that of the prior density. Therefore, the posterior pdf can serve as the prior pdf for the next time instant.

The confrontation of the true data model (9) with the exponential family form (11) reveals that there cannot exist a convenient conjugate prior distribution of the form (12). For the one-shot static Poisson regression without the variance stabilization a workaround is suggested in [25]. It consists of approximating the likelihood of the (time-invariant) β for all the observed data $x_{1:t}$,

$$f(y_{1:t}|\beta,x_{1:t}) = \prod_\tau f(y_\tau|\beta,x_\tau) \tag{14}$$

in the posterior distribution by a normal distribution, in the sense proposed originally by Bartlett and Kendall [28]. Although we deal with the transformed variable $\widetilde{y}_t = \sqrt{y_t}$ and our aim is to update the estimate of β_t sequentially with the incoming observations, we can adopt a similar philosophy.

Let us once again focus on the Bayesian update (10). The model (9) acts in the posterior pdf as a function of β_t with x_t and $\widetilde{y}_t$ fixed. It is known that the density functions of a real random variable u that are proportional to $\exp(uz)\exp(-\exp(u))$ where z stands for a parameter, can be approximated by a normal distribution $\mathcal{N}(\log z, z^{-1})$, provided that z is large enough [28]. In Equation (9) with x_t and $\widetilde{y}_t$ fixed, this leads to the approximation by $\mathcal{N}(\log\widetilde{y}_t^2, \widetilde{y}_t^{-2})$. In order to compensate the approximation error under low values of $\widetilde{y}_t$, we suggest the following moment matching-based calibration: the bias of both the approximating mean value and the standard deviation can be with sufficient accuracy predicted and suppressed using the regression models with $\widetilde{y}_t$

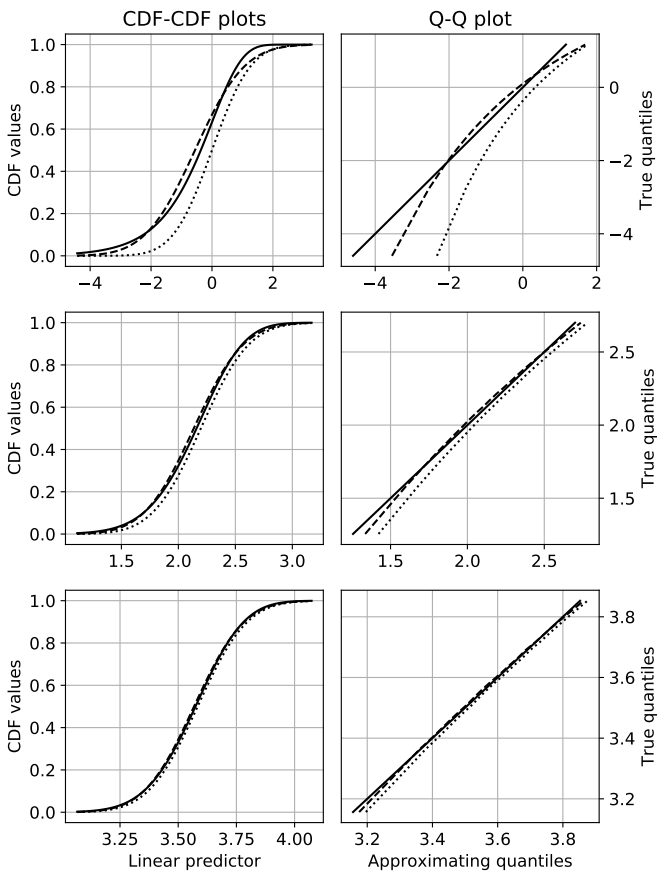

Fig. 1. Approximation of the true distribution of $\widetilde{y}_t$ for the values 1 (top), 3 (middle), and 6 (bottom), i.e., $y_t = 1$, 9, and 36. The solid line depicts the true distribution, the dashed line is used for the calibrated distribution, and the dotted line for the noncalibrated distribution. Apparently, the calibrated approximation by $\mathcal{N}(\mu_c, \sigma_c^2)$ has a better accuracy than $\mathcal{N}(\log \widetilde{y}_t^2, \widetilde{y}_t^{-2})$. With increasing $\widetilde{y}_t$ the approximating distributions tend to the true one.

in the role of the regressand. The calibrated normal distribution removing the bias has the mean and standard deviation

$$\begin{aligned}\mu_c &= \log \widetilde{y}_t^2 - \frac{0.5574}{\widetilde{y}_t^2},\\ \sigma_c &= \frac{1}{\widetilde{y}_t} + \frac{0.0724}{\widetilde{y}_t^2} + \frac{0.2121}{\widetilde{y}_t^4},\end{aligned} \tag{15}$$

where the coefficients were obtained from the ordinary least squares over the values $\widetilde{y}_t^2 = 1, \ldots, 100$. Fig. 1 compares the true distribution of $\widetilde{y}_t$, the calibrated, and noncalibrated normal approximations. We remark that since $y_t \in \mathbb{N}$, it is possible to use a table of precomputed values of μ_c and σ_c for low y_t.

C. The posterior distribution

The posterior distribution $\pi(\beta_t | x_{0:t}, \widetilde{y}_{0:t})$ in (10) now consists of the normal distribution $\mathcal{N}(\mu_c, \sigma_c^2)$ defined by the moments (15) and a prior distribution $\pi(\beta_t | x_{0:t-1}, \widetilde{y}_{0:t-1})$. The normal distribution belongs to the exponential family and can be written in the form (11),

$$f(y_t | x_t, \beta_t) = \frac{1}{\sqrt{2\pi\sigma_{c,t}^2}} \exp\left\{ -\frac{1}{2\sigma_{c,t}^2} ||\mu_{c,t} - \beta_t^\mathsf{T} x_t||^2 \right\}$$

$$\propto \exp\left\{ -\frac{1}{2} \operatorname{Tr}\left(\underbrace{\begin{bmatrix}-1\\ \beta_t\end{bmatrix}\begin{bmatrix}-1\\ \beta_t\end{bmatrix}^\mathsf{T}}_{\eta \equiv \eta(\beta_t)} \underbrace{\begin{bmatrix}\mu_{c,t}\\ x_t\end{bmatrix}\begin{bmatrix}\mu_{c,t}\\ x_t\end{bmatrix}^\mathsf{T} \sigma_{c,t}^{-2}}_{T(x_t, \widetilde{y}_t)} \right)\right\}, \tag{16}$$

where, for the sake of simplicity, we use a slight abuse of notation and avoid vectorizations. This reveals that the appropriate conjugate prior distribution is the normal distribution with the mean vector b_{t-1} and the covariance matrix P_{t-1} with the pdf in the compatible form

$$\pi(\beta | b_{t-1}, P_{t-1}) \propto \exp\left\{ -\frac{1}{2} \operatorname{Tr}\left(\underbrace{\begin{bmatrix}-1\\ \beta\end{bmatrix}\begin{bmatrix}-1\\ \beta\end{bmatrix}^\mathsf{T}}_{\eta \equiv \eta(\beta)} \underbrace{\begin{bmatrix}b_{t-1}^\mathsf{T}\\ I\end{bmatrix} P_{t-1}^{-1} \begin{bmatrix}b_{t-1}^\mathsf{T}\\ I\end{bmatrix}^\mathsf{T}}_{\Xi_{t-1}} \right)\right\}, \tag{17}$$

where I is the $n \times n$ identity matrix. The posterior distribution following from the Bayes' rule (10) in terms of the update of the prior hyperparameters (13) is given by

$$\Xi_t = \Xi_{t-1} + T(x_t, \widetilde{y}_t). \tag{18}$$

A little algebra reveals that the posterior 'conventional' normal hyperparameters are

$$\begin{aligned} b_t &= (\sigma_{c,t}^{-2} x_t x_t^\mathsf{T} + P_{t-1}^{-1})^{-1} (P_{t-1}^{-1} b_{t-1}^\mathsf{T} + \sigma_{c,t}^{-2} x_t \mu_{c,t}),\\ P_t &= (\sigma_{c,t}^{-2} x_t x_t^\mathsf{T} + P_{t-1}^{-1})^{-1}. \end{aligned} \tag{19}$$

D. Time-varying β_t

The case of constant model parameters is rather an exception than a rule. Although there is no explicit model for the evolution $\beta_{t-1} \to \beta_t$ in many situations, we still may proceed by means of *forgetting* if the variations are slow. It means heuristic discounting of potentially outdated information about β_t from the posterior distribution. The most basic yet prevailing procedure is the exponential forgetting, flattening the prior pdf by its exponentiation [31],

$$\pi(\beta_t | x_{0:t-1}, y_{0:t-1}) = [\pi(\beta_{t-1} | x_{0:t-1}, y_{0:t-1})]^\alpha, \quad \alpha \in [0, 1]. \tag{20}$$

In conjugate priors this amounts to

$$\nu_{t-1} \leftarrow \alpha \nu_{t-1}, \qquad \Xi_{t-1} \leftarrow \alpha \Xi_{t-1}. \tag{21}$$

For more elaborate forgetting methods, see, e.g., [32] and the references therein.

III. Diffusion estimation

Let us now assume a network consisting of a set $\mathcal{I}$ of agents that independently observe the processes $Y_t^{(i)} \sim Po(\exp(\beta_t^\mathsf{T} x_t^{(i)}))$ where β_t is global, i.e., identical for all $i \in \mathcal{I}$, while the observations $y_t^{(i)}$ and potentially the regressors $x_t^{(i)}$ are local. These situations occur, e.g., in particle detection, where each agent (sensor) 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. Furthermore assume that at each time instant t the agents may perform one mutual exchange of the posterior pdfs with all their adjacent neighbors within 1 network hop

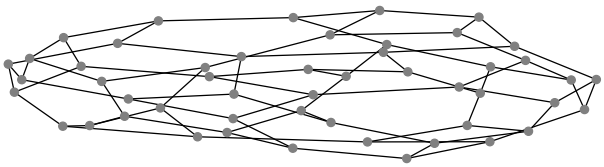

Fig. 2. Network topology used in the simulation.

distance. We denote by $\mathcal{I}^{(i)}$ the set of i's adjacent neighbors, and let $i \in \mathcal{I}^{(i)}$ too. The pdfs are fully represented by $\Xi_t^{(i)}$.

A careful inspection of the Bayesian update (13) reveals 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 Sec. II-B. The combination of the posterior pdfs in terms of the hyperparameter averaging

$$\bar{\Xi}_t^{(i)} = \frac{1}{\mathrm{card}(\mathcal{I}^{(i)})} \sum_{j \in \mathcal{I}^{(i)}} \Xi_t^{(j)} \tag{22}$$

where $\mathrm{card}(\mathcal{I}^{(i)})$ denotes the cardinality of $\mathcal{I}^{(i)}$ thus amounts to the uniformly weighted Bayesian update by neighbors' observations. Moreover, from a little algebra it follows that

$$\bar{P}_t^{(i)} = \left(\frac{1}{\mathrm{card}(\mathcal{I}^{(i)})} \sum_{j \in \mathcal{I}^{(i)}} P_t^{-1,(j)} \right)^{-1}, \tag{23}$$

$$\bar{b}_t^{(i)} = \bar{P}_t^{(i)} \left(\frac{1}{\mathrm{card}(\mathcal{I}^{(i)})} \sum_{j \in \mathcal{I}^{(i)}} P_t^{-1,(j)} b_t^{\intercal,(j)} \right), \tag{24}$$

which is known as the *covariance intersection*. The first author shows in [18] that this result is Kullback-Leibler-optimal. Although (22) corresponds to uniformly weighted averaging of neighbors' knowledge about β_t, the covariance matrices $P_t^{(j)}$ effectively reflect the uncertainty about the individual estimates, see (24). More elaborate strategies are proposed, e.g., in [33].

Algorithm 1 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 α. 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. (21).
3) Update the prior hyperparameter, Eq. (18)

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. (22), or in terms of $b_t^{(j)}$ and $P_t^{(j)}$, Eq. (24).
3) Evaluate the point the estimate $\bar{b}_t^{(i)}$ and the covariance matrix $\bar{P}_t^{(i)}$ from $\bar{\Xi}_t^{(i)}$.

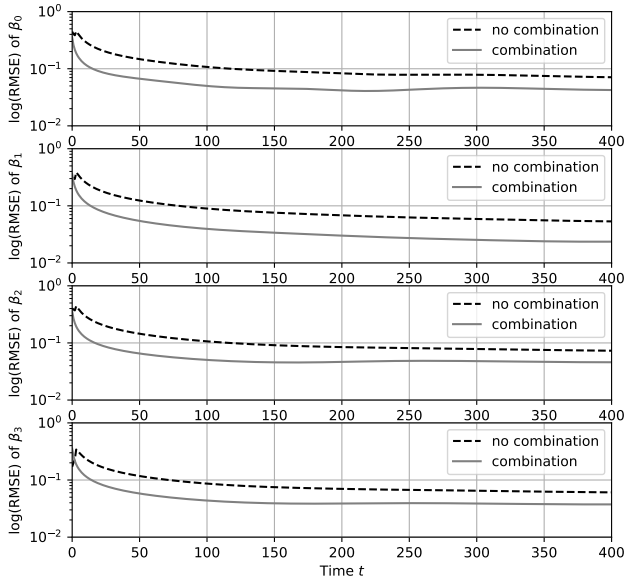

Fig. 3. Evolution of the RMSE averaged over all network nodes.

IV. SIMULATION EXAMPLE

This example demonstrates the performance of the proposed method. As the authors are not aware of any alternative method for sequential modeling of counts in diffusion networks, two scenarios are compared: (i) the 'combination' scenario using Algorithm 1, and (ii) the isolated 'no combination' scenario.

Fig. 2 depicts the randomly generated diffusion network of 50 nodes. They observe independently generated outcomes of a Poisson regression process simulated with

$$\beta_t = \begin{pmatrix} 0.7 + 0.1 \cdot \sin\left(4\pi \cdot \frac{t}{500}\right) \\ 0.5 + 0.05 \cdot \cos\left(2\pi \cdot \frac{t}{500}\right) \\ -0.2 \\ 0.05 \cdot \cos\left(\pi \cdot \frac{t}{500}\right) \end{pmatrix}, \quad t = 1, \ldots, 400,$$

and randomly generated regressors $x_t^{(i)} \sim U(0,5)^4$. For all the nodes $i \in \{1, \ldots, 10\}$, 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 depicts the root mean-squared error (RMSE) evolution averaged over all nodes. The collaborative estimation improves the estimation quality, particularly in terms of the convergence rate. When the estimates stabilize, the RMSE may slightly vary due to the time-varying nature of β_t. The results show that the proposed method yields good estimation performance. Our observations also indicate that the estimates of the time-varying β_t are more stable in terms of smoothness.

V. CONCLUSION

We proposed a method for sequential distributed modeling of counts using the Poisson model. The parameters are locally estimated using a calibrated stabilized estimation procedure. Then, the posterior pdfs are combined in the network. The future work will focus on count processes with specific properties, e.g., the zero inflation or overdispersion [24], [34], and on the full adapt-then-combine diffusion strategy [4], [5].

REFERENCES

[1] M. H. Cintuglu and D. Ishchenko, "Secure distributed state estimation for networked microgrids," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8046–8055, Oct. 2019.

[2] S. Ghazanfari-Rad and F. Labeau, "Formulation and analysis of LMS adaptive networks for distributed estimation in the presence of transmission errors," *IEEE Internet of Things J.*, vol. 3, no. 2, pp. 146–160, Apr. 2016.

[3] Y. Chen, S. Kar, and J. M. F. Moura, "The Internet of Things: Secure distributed inference," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 64–75, Sep. 2018.

[4] A. H. Sayed, "Diffusion Adaptation over Networks," in *Academic Press Library in Signal Processing*, R. Chellapa and S. Theodoridis, Eds. Academic Press, Elsevier, May 2014, vol. 3, pp. 323–454.

[5] ——, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

[6] F. S. Cattivelli, C. G. Lopes, and A. H. Sayed, "Diffusion recursive least-squares for distributed estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1865–1877, May 2008.

[7] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.

[8] J. Plata-Chaves, N. Bogdanović, and K. Berberidis, "Distributed diffusion-based LMS for node-specific adaptive parameter estimation," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3448–3460, Jul. 2015.

[9] J. Plata-Chaves, M. H. Bahari, M. Moonen, and A. Bertrand, "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)*. IEEE, Mar. 2016, pp. 4159–4163.

[10] D. Liu, X. Yang, W. Huang, and S. Chen, "Diffusion LMS with component-wise variable step-size over sensor networks," *IET Signal Processing*, vol. 10, no. 1, pp. 37–45, Feb. 2016.

[11] F. S. Cattivelli and A. H. Sayed, "Diffusion strategies for distributed Kalman filtering and smoothing," *IEEE Trans. Autom. Control*, vol. 55, no. 9, pp. 2069–2084, 2010.

[12] J. Hu, L. Xie, and C. Zhang, "Diffusion Kalman filtering based on covariance intersection," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 891–902, Feb. 2012.

[13] S. S. Dias and M. G. S. Bruno, "Distributed Bernoulli filters for joint detection and tracking in sensor networks," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 3, pp. 260–275, 2016.

[14] M. G. S. Bruno and S. S. Dias, "Collaborative emitter tracking using Rao-Blackwellized random exchange diffusion particle filtering," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 19, Feb. 2014.

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

[16] W. Li, Z. Wang, Y. Yuan, and L. Guo, "Particle filtering with applications in networked systems: A survey," *Complex & Intelligent Systems*, vol. 2, no. 4, pp. 293–315, 2016.

[17] K. Dedecius, J. Reichl, and P.M. Djurić, "Sequential estimation of mixtures in diffusion networks," *IEEE Signal Processing Lett.*, vol. 22, no. 2, pp. 197–201, 2015.

[18] K. Dedecius and P. M. Djurić, "Sequential estimation and diffusion of information over networks: A Bayesian approach with exponential family of distributions," *IEEE Trans. Signal Process.*, vol. 65, no. 7, pp. 1795–1809, Apr. 2017.

[19] K. Dedecius and V. Sečkárová, "Factorized estimation of partially shared parameters in diffusion networks," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5153–5163, Oct. 2017.

[20] N. Bosowski, V. K. Ingle, and D. Manolakis, "Generalized linear models for count time series," in Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 4272–4276.

[21] L. Wang and Y. Chi, "Stochastic approximation and memory-limited subspace tracking for Poisson streaming data," *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 1051–1064, Feb. 2018.

[22] D. Manolakis, N. Bosowski and V. K. Ingle, "Count time-series analysis: A signal processing serspective," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 64–81, May 2019.

[23] P. McCullagh and J. A. Nelder, *Generalized Linear Models, Second Edition*, Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.

[24] R. H. Myers, D. C. Montgomery, G. G. Vining, and T. J. Robinson, *Generalized Linear Models*, Wiley Series in Probability and Statistics. Hoboken, NJ, USA: Wiley & Sons., Mar. 2010.

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

[26] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal densities," *Journal of the American Statistical Association*, vol. 81, no. 393, pp. 82–86, Mar. 1986.

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

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

[29] F. A. Haight, *Handbook of the Poisson Distribution*. New York: John Wiley & Sons, 1967.

[30] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*. Harvard University Press, Jan. 1961.

[31] V. Peterka, "Bayesian approach to system identification," in *Trends and Progress in System Identification*, P. Eykhoff, Ed. Oxford, U.K.: Pergamon Press, 1981, pp. 239–304.

[32] K. Dedecius, I. Nagy, and M. Kárný, "Parameter tracking with partial forgetting method," *International Journal of Adaptive Control and Signal Processing*, vol. 26, no. 1, pp. 1–12, 2012.

[33] D. Jin, J. Chen, C. Richard, J. Chen, and A. H. Sayed, "Affine combination of diffusion strategies over networks," *IEEE Trans. Signal Process.*, (early access), pp. 1–16, 2020.

[34] D. B. Hall, "Zero-inflated Poisson and binomial regression with random effects: A case study," *Biometrics*, vol. 56, no. 4, pp. 1030–1039, Dec. 2000.