Convex estimation and prediction for "generalized" generalized linear model

January 28, 2022

Contents

1 Introduction

The generalized linear model (GLM) provides a flexible family of predictive models for capturing the general notion of mapping input/response variables to predictors typically through a non-linear *link function*. GLM is an important type of model in statistics and machine learning (see, e.g., [4, 1]). In general, we can choose the link function from a wide range of selections, e.g., induced by exponential family distributions. Some common GLMs include the Bernoulli model (where the link function is a sigmoid function), the categorical distribution model, and the Poisson model (where the link function is the Poisson distribution). GLM is also related to the prediction problem in machine learning, where the non-linear link function can also choose to be, for instance, the ReLu (rectifier linear unit) function and the soft-max function (which can be viewed as the link function of the categorical distributions). The model fitting of GLMs is largely based on two approaches, the maximum likelihood (ML) and the least-square (LS) formulation. However, the LS model fitting can result in a non-convex optimization problem, which is usually computationally intractable and lacks performance guarantees.

In this paper, we present a computational framework for recovering parameters of the "generalized" generalized linear model (GGLM). When the link functions are monotone (as it was assumed when presenting a GGLM model), the resulting procedure is computationally efficient, which is not always the case for more traditional recovery procedures using classical loss functions or based on Maximum Likelihood. To this end, we use a *Variational Inequality* (VI) oriented procedure, which results in convex formulation [3]. Here we focus on the spatio-temporal models as an instance of GGLM: after rearranging the observation in the model, the problem can be cast into a GGLM model; however, in doing so, the observations are dependent. To address the challenges in recovering spatio-temporal models, we adopt two strategies. First, the VI formulation can avoid the non-convexity caused by non-linear link functions, since the structural assumptions on the model are weaker than those resulting from ML-based or LS-based problems. Second, to address spatio-temporal dependencies in the proposed GGLM, we couple the analysis with martingale concentration inequalities to quantify the recovery error.

The rest of the paper is organized as follows. Section 2 introduces the GGLM model, recovery algorithm, and performance metrics. Section 3 presents performance guarantees

for recoveries based on solving monotone operator based variational inequalities. Section 4 discusses how to use the framework to a special case: spatio-temporal processes, followed by two representative examples of Categorial processes (Section 4.1) and Poisson processes (Section 4.2). Finally, Section concludes the paper with discussions.

Notation: The notations of the paper is standard. All linear spaces we are working in are $\mathbf{R}^n$'s, with different dimensions n, equipped with the standard Euclidean structure. Sometimes the vectors from $\mathbf{R}^n$ in question have additional structure – they are represented as one-, or two-, or even three-dimensional arrays of blocks of common size. In all cases, we assume that the entries in the arrays under consideration independently of the block structure, if any, are assigned serial numbers, so that the arrays can be represented as column vectors from appropriate $\mathbf{R}^n$. This allows us to write inner products as $x^T y$, to identify a linear mapping from $\mathbf{R}^n$ into $\mathbf{R}^m$ with $m \times n$ matrix, its conjugate – with the transpose of this matrix, etc.

2 "Generalized" Generalized Linear Model (GGLM)

We start by presenting the main modeling framework, which we will demonstrate using examples of its wide applicability. In statistics, the generalized linear model (GLM) [4] is a flexible generalization of the Gaussian model: where the response variable is related to the predictor by a linear function contaminated by additive Gaussian noise. The GLM generalizes linear regression by allowing the response variable to be related to the predictors via a non-linear link function, which determines the mean and variance (dispersion) of the response.

2.1 Situation and goal

Consider the "Generalized" Generalized Linear Model (GGLM) as follows

> At time $t = 1, 2, ...$ we observe random pair $\upsilon_t = (\zeta_t, H_t)$, where random vector ζ_t representing the response takes values in a known subset $\mathcal{Z}_t$ of some $\mathfrak{Z}_t = \mathbf{R}^{n_t}$, H_t is random $\kappa \times n_t$ matrix, representing the predictor, taking values in a known subset $\mathcal{H}_t$ of the space $\mathbf{R}^{\kappa \times n_t}$ of $\kappa \times n_t$ matrices. For every

$t = 1, 2, ...$ the conditional, $v^{t-1} = (v_1, ..., v_{t-1})$ *and* H_t given, distribution of ζ_t has expectation

$$\mathbf{E}_{|v^{t-1}, H_t}\{\zeta_t\} = \Phi_t(H_t^T \beta), \tag{1}$$

where

- $\Phi_t(\cdot) : \mathrm{Dom}\, \Phi_t \to \mathfrak{Z}_t$ — the t-th link function — is a continuous monotone mapping defined on a closed convex domain $\mathrm{Dom}\, \Phi_t \subset \mathfrak{Z}_t$, monotonicity meaning that

$$[\Phi_t(x) - \Phi_t(y)]^T [x - y] \geq 0 \ \forall x, y \in \mathrm{Dom}\, \Phi_t;$$

- β is the vector of parameters taking values in a convex compact subset $\mathcal{B}$ of $\mathfrak{B} = \mathbf{R}^\kappa$;
- $\mathcal{H}_t$ and $\mathfrak{B}$ are such that $H_t^T x \in \mathrm{Dom}\, \Phi_t$ for all $H_t \in \mathcal{H}_t$, $x \in \mathfrak{B}$, and t.

We call the situation *stationary*, if $\mathfrak{Z}_t$, $\mathcal{Z}_t$, $\mathcal{H}_t$ and Φ_t are independent of t. Our goal is to recover β from observations $v^N = (v_1, ..., v_N)$, where N is a given time horizon.

Note that the stationary version of our model resembles the generalized linear model (see, e.g., [4]) where H_t is the regressor, and ζ_t is the outcome of time instant t, $t = 1, 2, ...$; the extension lies in the fact that we do not assume the pairs $v_t = (H_t, \zeta_t)$, $t = 1, 2, ...$, to be i.i.d.

2.2 Parameter recovery

The problem we address here is recovering the parameters β of GGLM model from observations. To this end, we intend to use a *Variational Inequality* (VI) oriented procedure; when the link functions are monotone (as it was assumed when presenting a GGLM model), the resulting procedure is computationally efficient, which not always is the case for more traditional recovery procedures using classical loss functions or based on Maximum Likelihood.

Monotone vector fields and variational inequalities. We start by introducing necessary background information.

Definition 2.1 (Monotone vector field.). *A vector field $F : \mathcal{X} \to \mathbf{R}^N$ defined on a nonempty convex subset $\mathcal{X}$ of $\mathbf{R}^N$ is called monotone, if $\langle F(x)-F(y), x-y\rangle \geq 0$ whenever $x, y \in \mathcal{X}$.*

A domain $\mathcal{X}$ and a monotone on $\mathcal{X}$ vector field F specify the *Variational Inequality*

$$\text{Find } z \in \mathcal{X} : \langle F(w), w - z\rangle \geq 0, \quad \forall w \in \mathcal{X}. \qquad \text{VI}[F, \mathcal{X}]$$

Vectors z satisfying the requirement just defined are called weak solutions to VI$[F, \mathcal{X}]$. A vector $z \in \mathcal{X}$ is called a *strong solution* to VI$[F, \mathcal{X}]$. when

$$\langle F(z), w - z\rangle \geq 0, \quad \forall w \in \mathcal{X}.$$

From monotonicity of F it follows that every strong solution is a weak one. When the monotone vector field F is continuous, the weak solutions to VI$[F, \mathcal{X}]$ are the same as strong solutions. Weak solutions definitely exist whenever $\mathcal{X}$ is a nonempty convex compact set.

Some basic example of monotone vector fields are
— the gradient field $\nabla f(x)$ of continuously differentiable convex function $f : \mathcal{X} \to \mathbf{R}$;
— the vector field $F(u, v) = [\nabla_u f(u, v); -\nabla_v f(u, v)]$ of continuously differentiable convex in u and concave in v function $f(u, v) : \mathcal{X} := U \times V \to \mathbf{R}$.
In these cases, the weak≡strong solutions to VI$[F, \mathcal{X}]$ are,respectively, the minimizers of f on $\mathcal{X}$ and the saddle points (min in u, max in v) of f on $U \times V$.

Application to GGLM case. Given a GGLM from Section 2, we associate with an observation $\upsilon^N = \{\upsilon_\tau = (\zeta_\tau, H_\tau), 1 \leq \tau \leq N\} \in \overbrace{[\mathcal{Z}_1 \times \mathcal{H}_1]}^{\Upsilon_1} \times \ldots \times \overbrace{[\mathcal{Z}_N \times \mathcal{H}_N]}^{\Upsilon_N}$ a collection of weights $\lambda_t > 0$, $1 \leq t \leq N$ which should be deterministic functions of (υ^{t-1}, H_t): $\lambda_t = \Lambda_t(\upsilon^{t-1}, H_t)$ and two vector fields. First, the observable field

$$F_{\upsilon^N}(z) = \frac{1}{N} \sum_{t=1}^{N} \Lambda_t^{-1}(\upsilon^{t-1}, H_t) \left[H_t \Phi_t\left(H_t^T z\right) - H_t \zeta_t\right] : \mathcal{B} \to \mathfrak{B},$$

and second, the unobservable field which is a "population" version of $F_{\upsilon^N}(\cdot)$,

$$\overline{F}_{\upsilon^N}(z) = \frac{1}{N}\sum_{t=1}^{N} \Lambda_t^{-1}(\upsilon^{t-1}, H_t)\left[H_t\Phi_t\left(H_t^T z\right) - H_t\Phi_t\left(H_t^T\beta\right)\right] : \mathcal{B} \to \mathfrak{B}.$$

Note that in the situation of Section 2.1 these vector fields are well-defined and continuous on $\mathcal{B}$; they clearly are monotone along with Φ.[1] In addition, $\mathcal{B}$ is compact. Consequently the VI:

$$\text{Find } z \in \mathcal{B} : \langle F_{\upsilon^N}(x), x - z\rangle \geq 0, \ \forall x \in \mathcal{B} \qquad \text{VI}(F_{\upsilon^N}, \mathcal{B})$$

has solutions, and these weak solutions are the same as the strong solutions to the VI. Our estimate of β will be the weak≡strong solution to the variational inequality $\text{VI}(F_{\upsilon^N}, \mathcal{B})$.

Special case. When the monotone vector mappings $\Phi_t : \text{Dom}\,\Phi_t \to \mathfrak{Z}_t$ are the gradient fields of convex functions $\mathfrak{F}_t(\cdot) : \text{Dom}\,\Phi_t \to \mathbf{R}$, the vector field F_{υ^N} is the gradient field of the convex function

$$\mathfrak{F}_{\upsilon^N}(x) = \frac{1}{N}\sum_{t=1}^{N} \Lambda_t^{-1}(\upsilon^{t-1}, H_t)\left[\mathfrak{F}_t(H_t^T x) - x^T H_t\zeta_t\right] \tag{2}$$

so that the solutions to $\text{VI}(F_{\upsilon^N}, \mathcal{B})$ are exactly the optimal solutions to the convex minimization problem

$$\min_{x\in\mathcal{B}} \mathfrak{F}_{\upsilon^N}(x). \tag{3}$$

In particular, in the case of the identity link $\Phi_t(\eta_t) \equiv \eta_t$, $\eta_t \in \mathfrak{Z}_t$, we have $\Phi_t(\eta_t) = \nabla\left[\frac{1}{2}\eta_t^T\eta_t\right]$, and the proposed recovery becomes just the Least Squares (LS).

An alternative: Maximum Likelihood estimation. Assuming that the density, taken w.r.t. some reference measure, of the conditional, υ^{t-1} and H_t given, distribution of ζ_t is a known function of the form $\psi_t(\zeta_t, H_t^T\beta)$, (e.g., specified by some exponential family distributions) we could try to recover β by maximizing likelihood of what we have

[1] We have used the elementary observation that if $x \mapsto Ax + b$ is affine mapping which maps a convex domain $X \subset \mathbf{R}^N$ into the domain $\text{Dom}\,G \subset \mathbf{R}^M$ of a monotone vector field G, then the vector field $A^T G(Ax + b)$ is monotone on X.

observed, that is, by solving the optimization problem

$$\max_{z\in\mathcal{B}} \frac{1}{N}\sum_{t=1}^{N} \ln(\psi_t(\zeta_t, H_t^T x)).$$

In order for this approach to be computation-friendly, the problem should be convex. The maximum likelihood problems is usually convex, when the link function is derived from the exponential family. In such cases, we can use the gradient of the log-likelihood function as the monotone operator in VI.

3 Performance Guarantees

We now establish results underlying theoretical guarantees for parameter recovery. We focus on the results for parameters recovered using VI. We start with presenting a few general results which will help our subsequent analysis and guarantee.

3.1 Concentration of empirical vector field

An important observation is as follows: β is the root of $\overline{F}_{\upsilon^N}(\cdot)$ in $\mathcal{B}$: $\overline{F}_{\upsilon^N}(\beta) = 0$, and

$$F_{\upsilon^N}(\beta) = F_{\upsilon^N}(\beta) - \overline{F}_{\upsilon^N}(\beta) = \frac{1}{N}\sum_{t=1}^{N} \Lambda_t^{-1}(\upsilon^{t-1}, H_t) H_t \left[\Phi_t\left(H_t^T\beta\right) - \zeta_t\right].$$

What is under the summation, is *martingale-difference*, since

$$\mathbf{E}_{|\upsilon^{t-1},H_t}[F_{\upsilon^N}(\beta) - \overline{F}_{\upsilon^N}(\beta)] = \frac{1}{N}\Lambda_t^{-1}(\upsilon^{t-1}, H_t) H_t \cdot \mathbf{E}_{|\upsilon^{t-1},H_t}\left[\Phi_t\left(H_t^T\beta\right) - \zeta_t\right] = 0;$$

we have used (1).

Proposition 3.1 (Concentration of empirical vector field)**.** *Assume that for every t, every $\upsilon^{t-1} \in \Upsilon_1 \times \ldots \times \Upsilon_{t-1}$ and every $H_t \in \mathcal{H}_t$ the conditional, υ^{t-1}, H_t given, distribution $P_{|\upsilon^{t-1},H_t}$ of ζ_t satisfies*

$$\ln\left(\mathbf{E}_{\zeta_t\sim P_{|\upsilon^{t-1},H_t}}\left\{\exp\{h^T\zeta_t\}\right\}\right) \le h^T\Phi_t\left(H_t^T\beta\right) + \Psi_{t,\upsilon^{t-1},H_t}(\|P_t h\|_1),\ \forall h \in \mathfrak{Z}_t \qquad (4)$$

where the rate functions $\Psi_{t,\upsilon^{t-1},H_t}(r):\mathbf{R}_+\to\mathbf{R}$ are continuous nondecreasing functions, and P_t are known deterministic matrices.

- *Let L_t, $t=1,2,...,$ be $m\times n_t$ matrices which are deterministic functions of υ^{t-1},H_t: $L_t=L_t(\upsilon^{t-1},H_t)$.*
- *Let the "average one-step prediction error" be*

$$G_{\upsilon^N}=\frac{1}{N}\sum_{t=1}^{N}L_t[\Phi_t(H_t^T\beta)-\zeta_t]$$

Note that when $L_t=\Lambda_t^{-1}(\upsilon^{t-1},H_t)H_t$, we get $G_{\upsilon^N}=F_{\upsilon^N}(\beta)$.

Then

(i) *For $\upsilon_\tau\in\Upsilon_\tau$, $1\le\tau\le N$, $1\le j\le m$, let $\Theta_{sj}(\upsilon^{s-1},H_s)$ be the $\|\cdot\|_1$-norm of j-th column in $P_sL_s^T$. Given a finite set $\Gamma=\{\alpha_i>0,1\le i\le K\}$ on the positive ray, and a tolerance $\epsilon\in(0,1)$, for every $j\le m$ the probability of the event*

$$\left\{\upsilon^N:|[G_{\upsilon^N}]_j|>\underbrace{\min_{i\le K}\left[\alpha_i\ln(2K/\epsilon)+\alpha_i\sum_{t=1}^{N}\Psi_{t,\upsilon^{t-1},H_t}(\alpha_i^{-1}N^{-1}\Theta_{tj}(\upsilon^{t-1},H_t))\right]}_{\delta_j(\upsilon^N)}\right\}\tag{5}$$

is at most ϵ.

(ii) *Assume that $\Psi_{t,\upsilon^{t-1},H_t}(\cdot)\le\Psi(\cdot)$ for all t, where $\Psi(\cdot):\mathbf{R}_+\to\mathbf{R}_+$ is a continuous nondecreasing "worst case" rate function, and that for some $\Theta<\infty$, the $\|\cdot\|_1$-norms of columns in $P_tL_t^T$ do not exceed Θ whenever $\upsilon_s\in\Upsilon_s$, $1\le s\le t$, for all t. Then for every $\delta\ge0$ one has*

$$\mathrm{Prob}\{\|G_{\upsilon^N}\|_\infty>\delta\}\le 2m\exp\{-N\Psi_*(\delta/\Theta)\},\tag{6}$$

where for $s\ge0$

$$\Psi_*(s)=\sup_{\alpha\ge0}[\alpha s-\Psi(\alpha)]\ge0,\tag{7}$$

is the Fenchel-Legendre transformation of $\Psi(\cdot)$.

Remark. Assume that we are in the situation of item (ii) and are given $\varepsilon\in(0,1)$, and let us look at the upper bound of the $(1-\varepsilon)$-quantile of $\|G_{\upsilon^N}\|_\infty$ we could get from (6).

This bound is (substitute $\gamma = \Theta/(N\alpha)$ into the definition of Ψ_*)

$$\begin{array}{rcl}\widehat{\delta} & = & \inf\left\{\delta > 0 : \exists \alpha > 0 : \left[\delta - \alpha N \Psi(N^{-1}\Theta/\alpha) - \alpha \ln\left(2m/\varepsilon\right)\right] \geq 0\right\} \\ & = & \inf_{\alpha>0}\left[\alpha \ln\left(2m/\varepsilon\right) + \alpha N \Psi(N^{-1}\Theta/\alpha)\right].\end{array}$$

On the other hand, we are in the situation when $\Theta_{tj}(\upsilon^{t-1}, H_t) \leq \Theta$ for all t, j and all $\upsilon^{t-1} \in \Upsilon^{t-1}$, $H_t \in \mathcal{H}_t$. It follows that given $\delta > \widehat{\delta}$, selecting $\bar{\alpha} > 0$ in such a way that

$$\left[\bar{\alpha} \ln\left(2m/\varepsilon\right) + \bar{\alpha} N \Psi(N^{-1}\Theta/\bar{\alpha})\right] \leq \delta$$

and specifying $\Gamma = \{\bar{\alpha}\}$, $\epsilon = \varepsilon/m$, the right hand side in the inequality in (5) would, for every $j \leq m$ be $\leq \delta$, implying, via the union bound, that the probability of the event

$$\|G_{\upsilon^N}\|_\infty > \delta$$

is $\leq \varepsilon$. Thus, in the case of (ii), the result of (i) with properly selected singleton Γ would result in the upper bound on the $(1-\varepsilon)$-quantile of $\|G_{\upsilon^N}\|_\infty$ which can be made arbitrarily close to the bound yielded by (ii). The rationale beyond (i) is that it results in *online bounds* $\delta_j(\upsilon^N)$ on the $(1-\epsilon)$-quantile of $\|G_{\upsilon^N}\|_\infty$, meaning that *when* $\Psi_{t,\upsilon^{t-1},H_t}$ *is observable*, the bound is an observable function of υ^N such that the probability for $|[G_{\upsilon^N}]_j|$ to exceed this bound is at most ϵ. This online bound adjusts itself to the observed rate functions $\Psi_{t,\upsilon^{t-1},H_t}(\cdot)$ and observed values of $\Theta_{tj}(\upsilon^{t-1}, H_t)$, in contrast to the deterministic bound $\widehat{\delta}$ oriented on the worst case rate function and a priori upper bound Θ on $\Theta_{tj}(\upsilon^{t-1}, H_t)$'s. Note that the price for cardinality K of Γ (the larger K, the stronger "adaptive abilities" of the online bound $\delta_j(\cdot)$) is low, since K is under the log.

Proof of Proposition 3.1. Let us set

$$\begin{array}{rcl}S_t(\upsilon^t) & = & \sum_{s=1}^t \frac{1}{N} L_s\left[\Phi_s\left(H_s^T\beta\right) - \zeta_s\right], \\ R_t(\upsilon^t; g) & = & \sum_{s=1}^t \Psi_{s,\upsilon^{s-1},H_s}(N^{-1}\|P_s L_s^T g\|_1) \quad [g \in \mathbf{R}^m],\end{array}$$

and let $\mathbf{E}_t$ stand for expectation over υ^t, and $\mathbf{E}_t^+$ stand for expectation over (υ^t, H_{t+1}).

(i): Let us fix j, and let $g = \chi e_j$, where e_j is j-th standard basic orth in $\mathbf{R}^m$ and

$\chi = \pm 1$, and $\gamma \geq 0$, and let

$$\begin{array}{rcl} \Sigma_t(\upsilon^t) & = & \sum_{s=1}^t \left[\frac{1}{N} \gamma g^T L_s \left[\Phi_s \left(H_s^T \beta \right) - \zeta_s \right] - \Psi_{s,\upsilon^{s-1},H_s}(\|\gamma N^{-1} P_s L_s^T g\|_1) \right] \\ & = & \gamma g^T S_t(\upsilon^t) - R_t(\upsilon^t; \gamma g). \end{array} \tag{8}$$

We have

$$\begin{array}{l} \mathbf{E}_{t+1}\left\{\exp\{\Sigma_{t+1}(\upsilon^{t+1})\}\right\} \\ = \mathbf{E}_t^+ \left\{ \mathbf{E}_{\zeta_{t+1} \sim P_{|\upsilon^t, H_{t+1}}} \left\{ \exp\left\{ \Sigma_t(\upsilon^t) - \Psi_{t+1,\upsilon^t,H_{t+1}}(\|\gamma N^{-1} P_{t+1} L_{t+1}^T g\|_1) \right.\right.\right. \\ \qquad \left.\left.\left. + \gamma N^{-1} [L_{t+1}^T g]^T [\Phi_{t+1}\left(H_{t+1}^T \beta\right) - \zeta_{t+1}] \right\} \right\} \right\} \\ = \mathbf{E}_t^+ \left\{ \exp\left\{ \Sigma_t(\upsilon^t) - \Psi_{t+1,\upsilon^t,H_{t+1}}(\|\gamma N^{-1} P_{t+1} L_{t+1}^T g\|_1) \right\} \right. \\ \qquad \left. \times \underbrace{\mathbf{E}_{\zeta_{t+1} \sim P_{|\upsilon^t, H_{t+1}}} \left\{ \exp\left\{ \gamma N^{-1} [L_{t+1}^T g]^T [\Phi_{t+1}\left(H_{t+1}^T \beta\right) - \zeta_{t+1}] \right\} \right\}}_{\leq \exp\{\Psi_{t+1,\upsilon^t,H_{t+1}}(\|\gamma N^{-1} P_{t+1} L_{t+1}^T g\|_1)\} \text{ by (4)}} \right\} \\ \leq \mathbf{E}_t^+ \left\{\exp\{\Sigma_t(\upsilon^t)\}\right\} = \mathbf{E}_t \left\{\exp\{\Sigma_t(\upsilon^t)\}\right\}. \end{array}$$

This reasoning works for all $t \geq 0$, with the empty sum $\Sigma_0(\upsilon^0)$, as usual, identically equal to 0. Consequently, $\mathbf{E}_{\upsilon^N}\left\{\exp\{\Sigma_N(\upsilon^N)\}\right\} \leq 1$, or, which is the same (note that $G_{\upsilon^N} = S_N(\upsilon^N)$):

$$\mathbf{E}_{\upsilon^N}\left\{\exp\{\gamma g^T G_{\upsilon^N} - R_N(\upsilon^N; \gamma g)\}\right\} \leq 1,$$

whence for every $\delta \geq 0$ one has

$$\text{Prob}\left\{\gamma g^T G_{\upsilon^N} - R_N(\upsilon^N; \gamma g) > \gamma\delta\right\} \leq \exp\{-\gamma\delta\}.$$

Setting $\gamma = 1/\alpha$ and recalling what R_N is, we get

$$\text{Prob}\left\{ g^T G_{\upsilon^N} > \alpha \sum_{t=1}^N \Psi_{t,\upsilon^{t-1},H_t}(\|\alpha^{-1} N^{-1} P_t L_{t-1}^T g\|_1) + \delta \right\} \leq \exp\{-\delta/\alpha\},$$

whence, due to monotonicity of $\Psi_{t,\upsilon^{t-1},H_t}(\cdot)$ and the fact that $\|P_t L_t^T g\|_1 \leq \Theta_{tj}(\upsilon^{t-1}, H_t)$,

we have

$$\mathrm{Prob}\left\{g^T G_{\upsilon^N} > \alpha \sum_{t=1}^{N} \Psi_{t,\upsilon^{t-1},H_t}(\alpha^{-1}N^{-1}\Theta_{tj}(\upsilon^{t-1},H_t)) + \delta\right\} \le \exp\{-\delta/\alpha\}.$$

Using the union bound, we get

$$\mathrm{Prob}\left\{|[G_{\upsilon^N}]_j| > \delta + \alpha \sum_{t=1}^{N} \Psi_{t,\upsilon^{t-1},H_t}(\alpha^{-1}N^{-1}\Theta_{tj}(\upsilon^{t-1},H_t))\right\} \le 2\exp\{-\delta/\alpha\}. \qquad (9)$$

For $1 \le i \le K$, let us set $\delta_i = \alpha_i \ln(2K/\epsilon)$, so that by (9) the probability of the event

$$\mathcal{E}_i = \{\upsilon^N : |[G_{\upsilon^N}]_j| > \alpha_i \ln(2K/\epsilon) + \alpha_i \sum_{t=1}^{N} \Psi_{t,\upsilon^{t-1},H_t}(\alpha_i^{-1}N^{-1}\Theta_{tj}(\upsilon^{t-1},H_t))\}$$

is at most ϵ/K. Consequently, the probability of the event $\cup_{i\le K}\mathcal{E}_i$, which is exactly the event (5), is at most ϵ, as claimed in (i).

(ii): From the definition of Θ it follows that $\|P_t L_t^T g\|_1 \le \Theta\|g\|_1$ for every $g \in \mathbf{R}^m$ and every t and $\upsilon^t \in \Upsilon^t$. Now let $g \in \mathbf{R}^m$ have $\|g\|_1 = 1$, and let $\gamma \ge 0$. We have

$$\begin{aligned}
&\mathbf{E}_{t+1}\left\{\exp\{\gamma g^T S_{t+1}(\upsilon^{t+1})\}\right\}\\
&= \mathbf{E}_t^+\left\{\exp\{\gamma g^T S_t(\upsilon^t)\}\mathbf{E}_{\zeta_{t+1}\sim P_{|\upsilon^t,H_{t+1}}}\left\{\exp\{[\gamma N^{-1}[L_{t+1}^T g]^T[\Phi_{t+1}\left(H_{t+1}^T\beta\right) - \zeta_{t+1}]\}\right\}\right\}\\
&\le \mathbf{E}_t^+\left\{\exp\{\gamma g^T S_t(\upsilon^t) + \Psi(\|\gamma N^{-1}P_{t+1}L_{t+1}^T g\|_1)\}\right\} \text{ [by (4) and due to } \Psi_{t,\upsilon^{t-1},H_t}(\cdot) \le \Psi(\cdot)]\\
&\le \mathbf{E}_t^+\left\{\exp\{\gamma g^T S_t(\upsilon^t)\}\right\}\exp\{\Psi(\gamma\Theta/N)\},\\
&\qquad\qquad \text{[since } \Psi \text{ is nondecreasing and } \|P_{t+1}L_{t+1}^T g\|_1 \le \Theta\|g\|_1 = \Theta]\\
&= \mathbf{E}_t\left\{\exp\{\gamma g^T S_t(\upsilon^t)\}\right\}\exp\{\Psi(\gamma\Theta/N)\},
\end{aligned}$$

whence

$$\ln\left(\mathbf{E}_{\upsilon^N}\left\{\gamma g^T S_N(\upsilon^N)\right\}\right) \le N\Psi(\gamma\Theta/N).$$

Therefore, for every $\delta > 0$ one has, using Markov inequality

$$\mathrm{Prob}\{g^T S_N(\upsilon^N) > \delta\} \le \exp\{N\Psi(\gamma\Theta/N) - \gamma\delta\},$$

whence finally

$$\begin{aligned} \mathrm{Prob}\{g^T S_N(\upsilon^N) > \delta\} &\leq \exp\{-N\Psi_*(\delta/\Theta)\}, & (10)\\ \Psi_*(s) &= \sup_{\alpha\geq 0}\left[\alpha s - \Psi(\alpha)\right]. & (11)\end{aligned}$$

Taking into account that $G_{\upsilon^N} = S_N(\upsilon^N)$, selecting g as $\pm$ basic orths in $\mathbf{R}^m$ and using the union bound, we arrive at (6). (ii) is proved. □

3.2 Quality of parameter recovery

Now we present a model parameter recovery performance bound, as a consequence of the results in Section 3.1. Assume that we are under the premise of Proposition 3.1.ii and that the conditional given the past distributions of ζ_t are "light tail" ones, so that the worst case rate function $\Psi(\cdot)$ is such that $\Psi(s) \leq Cs$ for all $s \in [0, c]$, with properly selected $c > 0$ and $C < \infty$. Specifying

$$m = \kappa,\ L_t = \Lambda_t^{-1}(\upsilon^{t-1}, H_t)H_t, \tag{12}$$

we get

$$G_{\upsilon^N} = F_{\upsilon^N}(\beta).$$

In this case under the premise of Proposition 3.1.ii the typical $\|\cdot\|_\infty$ norm of $F_{\upsilon^N}(\beta)$ for large N is of order of $\Theta\sqrt{\ln(\kappa)N}$. Thus, when there are reasons to believe that F_{υ^N} is, typically, strongly monotone on $\mathcal{B}$ with the parameter of strong monotonicity not deteriorating as N grows, we may hope that the solution to $\mathrm{VI}(F_{\upsilon^N}, \mathcal{B})$ will be at the distance $O(N^{-1/2})$ from β.

To formulate the precise statement, let us define the p-modulus of strong monotonicity $\theta_p(\upsilon^N)$ of $F_{\upsilon^N}(\cdot)$ on $\mathcal{B}$, $p \in [1, \infty]$, as

$$\theta_p(\upsilon^N) = \max\left\{\theta : \langle F_{\upsilon^N}(x) - F_{\upsilon^N}(y), x - y\rangle \geq \theta\|x - y\|_p^2,\ \forall x, y \in \mathcal{B}\right\}.$$

Note that strict positivity of $\theta_p(\upsilon^N)$ is independent of the value of p.

Proposition 3.2 (Parameter recovery guarantee)**.** *Let υ^N be such that $\theta_p(\upsilon^N) > 0$,*

$p \in [1,\infty]$. Then the weak≡strong solution $\widehat{\beta}(\upsilon^N)$ to the variational inequality $\mathrm{VI}(F_{\upsilon^N}, \mathcal{B})$ *is unique, and*

$$\|\widehat{\beta}(\upsilon^N) - \beta\|_p \le \|F_{\upsilon^N}(\beta)\|_\infty / \sqrt{\theta_p(\upsilon^N)\theta_1(\upsilon^N)}. \tag{13}$$

Proof. We know that in the situation of Section 2.1 weak solution $\hat{\beta} = \hat{\beta}(\upsilon^N)$ to $\mathrm{VI}(F_{\upsilon^N}, \mathcal{B})$ exists and is a strong solution; besides this, it is well known that under the premise of Proposition (sorry which proposition is referred to here?)The proposition we are proving! this solution is unique. Since $\hat{\beta}$ is a strong solution, we have $\langle F_{\upsilon^N}(\hat{\beta}), x - \hat{\beta}\rangle \ge 0$ for all $x \in \mathcal{B}$, and in particular $\langle F_{\upsilon^N}(\hat{\beta}), \beta - \hat{\beta}\rangle \ge 0$. As a result,

$$\|F_{\upsilon^N}(\beta)\|_\infty \|\beta - \hat{\beta}\|_1 \ge \langle F_{\upsilon^N}(\beta), \beta - \hat{\beta}\rangle = \underbrace{\langle F_{\upsilon^N}(\beta) - F_{\upsilon^N}(\hat{\beta}), \beta - \hat{\beta}\rangle}_{\ge \sqrt{\theta_p(\upsilon^N)\theta_1(\upsilon^N)}\|\beta-\hat{\beta}\|_p \|\beta-\hat{\beta}\|_1} + \underbrace{\langle F_{\upsilon^N}(\hat{\beta}), \beta - \hat{\beta}\rangle}_{\ge 0},$$

and (13) follows. □

To extract from Propositions 3.1 information on performance guarantees of our estimate, we need to understand how to lower-bound the modulus of strong monotonicity. Related theoretical analysis would require heavy assumptions on inter-dependence of the subsequent "regressors" H_t, $t = 1, 2, \ldots$. This is an issue we do not want to touch here. Instead, let us focus on how to lower-bound this modulus "online." An immediate observation is that if Φ_t are strongly monotone, with positive moduli γ_t, on their domains:

$$\langle \Phi_t(u) - \Phi_t(v), u - v\rangle \ge \gamma_t \langle u - v, u - v\rangle, \ \forall u, v \in \mathrm{Dom}\, \Phi_t,$$

then

$$\begin{array}{rl} & \langle F_{\upsilon^N}(x) - F_{\upsilon^N}(y), x - y\rangle \\ = & \frac{1}{N}\sum_{t=1}^N \Lambda_t^{-1}(\upsilon^{t-1}, H_t)\langle H_t\Phi_t(H_t^T x) - H_t\Phi_t(H_t^T y), x - y\rangle \\ \ge & \frac{1}{N}\sum_{t=1}^N \gamma_t\Lambda_t^{-1}(\upsilon^{t-1}, H_t)\langle H_t^T(x - y), H_t^T(x - y)\rangle \\ = & (x-y)^T \underbrace{\left[\frac{1}{N}\sum_{t=1}^N \gamma_t\Lambda_t^{-1}(\upsilon^{t-1}, H_t) H_t H_t^T\right]}_{\Gamma(\upsilon^N)} (x - y), \end{array}$$

whence

$$\theta_2(\upsilon^N) \ge \lambda_{\min}(\Gamma(\upsilon^N)), \tag{14}$$

where $\lambda_{\min}(Q)$ is the minimal eigenvalue of symmetric matrix Q. We see that $\theta_2(\upsilon^N)$

admits an "online observable" lower bound. We have also

$$\theta_p(\upsilon^N) \geq \min_{h:\|h\|_p=1} h^T\Gamma(\upsilon^N)h = \left[\max_{g:\|g\|_{p_*}\leq 1} g^T\Gamma^{-1}(\upsilon^N)g\right]^{-1} \qquad [p_* = \tfrac{p}{p-1}].$$

When $1 \leq p \leq 2$, the right hand side in this relation can be efficiently lower-bounded, the lower bound being tight within absolute constant [5] (equal to $\pi/2$ when $p = 1$).

3.2.1 The case of affine Φ_t's

Assume that the links $\Phi_t(\cdot)$ are affine; in this situation the vector field $F_{\upsilon^N}(\cdot)$ is affine as well. Assume also that the functions $\Psi_{t,\upsilon^{t-1},H_t}$ are observable. In this case observations υ^N imply $(1-\kappa\epsilon)$-reliable upper bound $\delta(\upsilon^N)$ on $\|F_{\upsilon^N}(\beta)\|_\infty$, specifically, the bound $\delta(\upsilon^N) := \max_{j\leq\kappa}\delta_j(\upsilon^N)$ with $\delta_j(\upsilon^N)$ given by (5) as applied with $L_t = \Lambda_t^{-1}(\upsilon^{t-1}, H_t)H_t$ (recall that this choice of L_t results in $G_{\upsilon^N} = F_{\upsilon^N}(\beta)$). As a result, the observable under our assumptions convex set

$$\Delta(\upsilon^N) = \{x \in \mathcal{B} : \|F_{\upsilon^N}(x)\|_\infty \leq \delta(\upsilon^N)\} \tag{15}$$

is "$(1-\kappa\epsilon)$-confidence interval" for β – this set contains β with probability $\geq 1-\kappa\epsilon$. In particular, given a norm $\|\cdot\|$ on $\mathbf{R}^\kappa$, the quantity

$$d(\upsilon^N) = \max_{x\in\Delta(\upsilon^N)} \|x - \widehat{\beta}\|$$

is an $(1-\kappa\epsilon)$-reliable upper bound on the $\|\cdot\|$-error of our estimate $\widehat{\beta}$. Whether this online error bound can or cannot be efficiently computed, it depends on the norm $\|\cdot\|$; e.g., for the uniform norm computing the bound is easy - it reduces to 2κ maximizations of linear forms over $\Delta(\upsilon^N)$.

Note that under our present assumptions every policy $\mathcal{R}$ for generating row vectors $L_t \in \mathbf{R}^{1\times n_t}$ as deterministic functions of υ^{t-1}, H_t, $t = 1, ..., N$, induces an observable, given υ^N, affine real-valued function

$$G^{\mathcal{R}}_{\upsilon^N}(x) = \frac{1}{N}\sum_{t=1}^{N} L_t\left[\Phi_t(H_t^T x) - \zeta_t\right] : \mathcal{B} \to \mathbf{R}$$

along with observable quantity $\delta_\epsilon^{\mathcal{R}}(\upsilon^N)$, given by (5), such that

$$\text{Prob}\{|G^{\mathcal{R}}_{\upsilon^N}(\beta)| \geq \delta^{\mathcal{R}}(\upsilon^N)\} \leq \epsilon.$$

Thus, the pair of observable linear inequalities $|G^{\mathcal{R}}_{\upsilon^N}(\beta)| \leq \delta^{\mathcal{R}}(\upsilon^N)$ on β do hold true with probability $\geq 1-\epsilon$, and we can add these inequalities to the description of $\Delta(\upsilon^N))$ as given by (15), thus refining the confidence interval, and we can implement several refinements of this type stemming from several policies $\mathcal{R}$. For an instructive application example, see Section 4.2.1.

4 Illustrations: Spatio-Temporal Processes

Assume that there are L spatial locations; at discrete time t, the state of location $k \in \{1, ..., L\}$ is described by realization $\omega_{t,k}$ of random variable taking values in a set $Z \subset \mathbf{R}^\mu$. Our observation at time t is the block vector

$$\zeta_t = [\omega_{t,1}; ...; \omega_{t,L}] \in \mathcal{Z} = Z \times ... \times Z \subset \mathfrak{Z} = \mathbf{R}^{\mu L}.$$

We assume that the conditional given what happened prior to time t expectation of ζ_t depends solely on the collection

$$(\zeta_{t-d}, \zeta_{t-d+1}, ..., \zeta_{t-1}) = \{\omega_{s,\ell} : t-d \leq s \leq t-1, 1 \leq \ell \leq L\},$$

where $d \geq 1$ is the "memory depth" of our process. Specifically, setting

$$\zeta_\tau^t = (\zeta_\tau, \zeta_{\tau+1}, ..., \zeta_t) = \{\omega_{s,\ell} \in \mathbf{R}^\mu : \tau \leq s \leq t, 1 \leq \ell \leq L\}, \zeta^t = \zeta_{-d+1}^t.$$

Assume that our observations start at time $-d+1$ and that for every $t \geq 1$ the conditional, given ζ^{t-1}, expectation $\mathbf{E}_{|\zeta^{t-1}}$ of ζ_t is

$$\mathbf{E}_{|\zeta^{t-1}}\{\zeta_t\} = \Phi(\eta^T(\zeta_{t-d}^{t-1})\beta), \tag{16}$$

where

- β is the vector of parameters of the process. This vector is obtained by writing down in once for ever prescribed order the entries of the elements of the collection

$$\{\beta^0 \in \mathbf{R}^{\mu L}, \beta^s \in \mathbf{R}^{\mu L \times \mu L}, 1 \le s \le d\}$$

of "actual parameters," and

$$\eta^T(\zeta_{t-d}^{t-1})\beta = \beta^0 + \sum_{s=1}^{b} \beta^s \zeta_{t-1} \in \mathbf{R}^{\mu L}.$$

In other words, denoting by

$$\kappa = \mu L + d\mu^2 L^2$$

the dimension of β, $\eta(\gamma)$ is matrix-valued function, taking values in $\mathbf{R}^{\kappa \times \mu L}$, of $L \times d$ array $\gamma = \{\gamma_{s\ell}, 1 \le s \le d, 1 \le \ell \le L\}$ of vectors $\gamma_{s\ell} \in \mathbf{R}^{\mu}$, and this function is uniquely defined by the requirement that

$$\forall x \in \mathfrak{B} : \eta^T(\gamma)x = x^0 + \sum_{s=1}^{d} x^s[\gamma_{t-s,1}; ...; \gamma_{t-s,L}],$$

where $\mathfrak{B} = \mathbf{R}^{\kappa}$ is the space of column vectors x obtained by arranging in a prescribed order the entries in a collection $\{x^0 \in \mathbf{R}^{\mu L}, x^s \in \mathbf{R}^{\mu L \times \mu L}, 1 \le s \le d\}$.

We always assume that β resides in a known in advance convex compact subset $\mathcal{B}$ of $\mathfrak{B}$;

- $\Phi(\cdot) : \mathrm{Dom}\,\Phi \to \mathfrak{Z}$ is the link function – a continuous and monotone vector field on a closed convex domain $\mathrm{Dom}\,\Phi \subset \mathfrak{Z}$. We always assume that whenever $\gamma \in \mathcal{Z} \times ... \times \mathcal{Z}$ and $x \in \mathcal{B}$, we have $\eta^T(\gamma)x \in \mathrm{Dom}\,\Phi$.

A spatio-temporal process we just have described can be modeled by stationary GGLM, where matrix component H_t of observations at time t is known deterministic function of $\zeta_{t-d}, \zeta_{t-d+1}, ..., \zeta_{t-1}$:

$$H_t = \eta(\zeta_{t-d}^{t-1}),$$

and all other components of the model are readily given by the above description. Consequently, we can apply the machinery from Section 2.2 to recover the parameters β from

observations ζ_t on time horizon $-d+1 \le t \le N$.

In this paper, we are especially interested in the case of *categorial* and *Poisson* spatio-temporal processes.

4.1 Categorial spatio-temporal process

Spatio-temporal categorial process with $\mu+1$ states models the situation when at time t k-th location can be in one of $\mu+1$ states — *ground state* 0 and *nontrivial states* $1, ..., \mu$. We encode these states by vectors from R^μ, with the ground state encoded by the origin, and nontrivial state i encoded by i-th basic orth in $\mathbf{R}^\mu$. As a result, Z becomes the finite set – the vertices of the simplex $\{z \in \mathbf{R}^\mu_+ : \sum_i z_i \le 1\}$, $\mathcal{Z}$ becomes the set of Boolean block-vectors with L blocks of dimension μ each, and at most one nonzero entry in every block. Next, we assume that the states ω_{tk} are conditionally given ζ^{t-1} independent across k. Finally, the values of Φ are block vectors with L blocks of dimension μ each; denoting by $\Phi_{ik}(z)$ i-th entry of k-th block in $\Phi(z)$, in order to meet (16), the quantity $\Phi_{ik}(\eta^T(\zeta_{t-d}^{t-1})\beta)$ should be the conditional given ζ^{t-1} probability for k-th location at time t to be at active state i; thus, it should be nonnegative, and the sum of these quantities should be ≤ 1. Consequently, when speaking about categorial spatio-temporal processes, we always assume that we are given a closed convex set $\overline{Z} \subset \mathrm{Dom}\,\Phi$ such that

$$\forall z \in \overline{Z} : \Phi(z) \ge 0 \ \& \sum_{i=1}^{\mu} \Phi_{ik}(z) \le 1,\ k \le L,$$

and that

$$\eta^T(\zeta_1, ..., \zeta_d)x \in \overline{Z}\ \forall (x \in \mathcal{B}, (\zeta_1, ..., \zeta_d) \in \mathcal{Z} \times ... \times \mathcal{Z}.$$

Note that the conditional given ζ^{t-1} probability for k-th location at time t to be in ground state is

$$1 - \sum_{i=1}^{\nu} \Phi_{ik}(\eta^T(\zeta_{t-d}^{t-1})\beta).$$

Categorial spatio-temporal process of the outlined type and VI-based techniques for parameter recovery in these processes were considered in [2].

Observe that in the resulting stationary GGLM for every $t \ge 1$ and $h \in \mathfrak{Z}$ setting

$\nu = \nu_t(\upsilon^{t-1}, H_t) = \Phi(H_t^T\beta)$ we get

$$\nu \geq 0, \|\nu\|_\infty \leq 1$$

and

$$\begin{array}{rcl} \mathbf{E}_{|\upsilon^{t-1},H_t}\left\{\exp\{h^T\zeta_t\}\right\} & = & \exp\left\{\sum_{k=1}^L\left[1+\sum_{i=1}^\mu \nu_{ik}[\mathrm{e}^{h_{ik}}-1]\right]\right\} \\ \Rightarrow & & \\ \ln\left(\mathbf{E}_{|\upsilon^{t-1},H_t}\left\{\exp\{h^T\zeta_t\}\right\}\right) & = & h^T\nu + \left[\sum_{k=1}^L \ln\left(1+\sum_{i=1}^\mu \nu_{ik}[\mathrm{e}^{h_{ik}}-1]\right) - h^T\nu\right]. \end{array} \tag{17}$$

The function

$$f_\nu(g) = \sum_{k=1}^L \ln\left(1+\sum_{i=1}^\mu \nu_{ik}[\mathrm{e}^{g_{ik}}-1]\right) - g^T\nu$$

of $g \in \mathfrak{Z}$ is convex, and therefore its maximum over the set $\{g : \|g\|_1 \leq r := \|h\|_1\}$ is achieved at a vertex, where all but one entries of g are zero, and the remaining entry is $\pm r$. As a result, setting

$$\begin{array}{l} \alpha_t(\upsilon^{t-1}, H_t) = \max\limits_{x\in\mathcal{B}} \|\Phi(H_t^T x)\|_\infty \in [\|\nu_t\|_\infty, 1], \\ \Psi^{\mathrm{cat}}_{t,\upsilon^{t-1},H_t}(s) := \max\limits_\alpha \left\{\max\left[\ln\left(1+\alpha[\mathrm{e}^s-1]\right)-\alpha s, \ln\left(1+\alpha[\mathrm{e}^{-s}-1]\right)+\alpha s\right] : 0 \leq \alpha \leq \alpha_t(\upsilon^{t-1}, H_t)\right\} \\ \leq \Psi^{\mathrm{cat}}(r) := \max\limits_\alpha \left\{\max\left[\ln\left(1+\alpha[\mathrm{e}^s-1]\right)-\alpha s, \ln\left(1+\alpha[\mathrm{e}^{-s}-1]\right)+\alpha s\right] : 0 \leq \alpha \leq 1\right\} \end{array}$$

we get continuous convex and even function of $s \in \mathbf{R}$, with $\Psi_{t,\upsilon^{t-1},H_t}(\cdot)$ observable at time t, such that

$$\forall(h \in \mathfrak{Z}, t \geq 1, \upsilon_s \in \Upsilon_s, s \leq t) : \mathbf{E}_{|\upsilon^{t-1},H_t}\left\{\exp\{h^T\zeta_t\}\right\} \leq h^T\Phi(H_t^T\beta) + \Psi^{\mathrm{cat}}_{t,\upsilon^{t-1},H_t}(\|h\|_1),$$

that is, relation (4) holds true with $\Psi_{t,\upsilon^{t-1},H_t} = \Psi^{\mathrm{cat}}_{t,\upsilon^{t-1},H_t}$ and the unit matrix in the role of P_t.

Observe also that, as it is immediately seen, whenever ζ_s, $1 \leq s \leq d$, are Boolean vectors (which definitely is the case when $\zeta_s \in \mathcal{Z}$), $s \leq d$, the matrix $\eta^T(\zeta_1, ..., \zeta_d)$ is Boolean, and every column of it has at most one nonzero entry, that is, the columns in all realizations of H_t are of ℓ_1-norm not exceeding 1. As a result, when applying Proposition 3.1 to a spatio-temporal categorial process, we can use the $\mu L \times \mu L$ identity matrix in the role of P_t's, and set $\Psi_{t,\upsilon^{t-1},H_t} \equiv \Psi^{\mathrm{cat}}_{t,\upsilon^{t-1},H_t}$. Further, when specifying L_t according to

(12), resulting in $G_{\upsilon^N} = F_{\upsilon^N}$, we can set $\Theta_{tj}(\upsilon^{t-1}, H_t) = \Lambda_t^{-1}(\upsilon^{t-1}, H_t)$ in (i); this choice works also whenever the $\|\cdot\|_1$-norms of all columns in L_t are bounded by 1. When, in addition, $\Lambda_t(\cdot) \equiv 1$, the premise of item (ii) is satisfied with $\Theta = 1$ and $\Psi = \Psi^{\rm cat}$.

4.2 Poisson spatio-temporal process

Spatio-temporal Poisson process models the situation where the state ω_{tk} of the k-th location is a nonnegative integer (that is, $\mu = 1$ and $Z = \{0, 1, 2, ...\}$), and the conditional given the past distribution of $\zeta_t = [\omega_{t1}; \omega_{t2}; ...; \omega_{tL}]$ is the distribution with independent across k Poisson, with parameters $\Phi_k(\eta^T(\zeta_{t-d}^{t-1})\beta)$, components ω_{tk}. In order to meet (16), we should assume now that we are given a closed convex set $\overline{Z} \subset \mathrm{Dom}\,\Phi$ such that

$$\Phi(z) \geq 0 \, \forall z \in \overline{Z}$$

and that for all nonnegative integral vectors $z_s \in \mathbf{R}^L$, $1 \leq s \leq d$, it holds

$$\eta^T(z_1, ..., z_d)x \in \overline{Z} \ \forall x \in \mathcal{B}.$$

Observe that in the resulting stationary GGLM for all $t \geq 1$ and all $h = [h_1; ...; h_L] \in \mathfrak{Z}$, setting $\nu_t = \Phi(H_t^T\beta)$ we have

$$\begin{array}{rcl} \mathbf{E}_{|\upsilon^{t-1}, H_t}\left\{\exp\{h^T\zeta_t\}\right\} & = & \exp\left\{\sum_{k=1}^L [\nu_t]_k(\exp\{h_k\} - 1)\right\} \\ \Rightarrow & & \\ \ln\left(\mathbf{E}_{|\upsilon^{t-1}, H_t}\left\{\exp\{h^T\zeta_t\}\right\}\right) & = & h^T\nu_t + \sum_{k=1}^L [\nu_t]_k\left[\exp\{h_k\} - h_k - 1\right] \\ & = & h^T\Phi(H_t^T\beta) + \sum_{k=1}^L [\nu_t]_k\left[\exp\{h_k\} - h_k - 1\right] \end{array} \tag{18}$$

Now let us set

$$\chi_t(\upsilon^{t-1}, H_t) = \max_{u \in \mathcal{B}} \|\Phi(H_t^T u)\|_\infty, \ \ f(s) = \exp\{s\} - s - 1$$

Then $\|\nu_t\|_\infty \leq \chi_t(\upsilon^{t-1}, H_t)$ and

$$\begin{array}{l}\sum_{k=1}^L [\nu_t]_k \left[\exp\{h_k\} - h_k - 1\right] \\ \leq \|\nu_t\|_\infty \|[f(h_1); f(h_2); ...; f(h_L)]\|_1 \leq \chi_t(\upsilon^{t-1}, H_t) \max_g \left\{E(g) : \|g\|_1 \leq \|h\|_1\right\}, \\ E(g) = \|[f(g_1); ...; f(g_L)]\|_1.\end{array}$$

Since f is nonnegative and convex on $\mathbf{R}$, the function $E(g)$ is convex, and therefore its maximum over the ℓ_1-ball $\{g : \|g\|_1 \leq \|h\|_1\}$ is achieved at a vertex, implying that

$$\max_g \{E(g) : \|g\|_1 \leq \|h\|_1\} = f(\|h\|_1).$$

Thus, setting

$$\Psi_{t,\upsilon^{t-1},H_t}(r) = \chi_t(\upsilon^{t-1}, H_t)\Psi(r),\ \Psi(r) := \exp\{r\} - r - 1,\ P_t = I_L,$$

and invoking (18), we conclude that

$$\mathbf{E}_{|\upsilon^{t-1},H_t}\left\{\exp\{h^T\zeta_t\}\right\} \leq h^T\Phi_t(H_t^T\beta) + \Psi_{t,\upsilon^{t-1},H_t}(\|P_t h\|_1),$$

as required by (4). Note that $\Psi_{t,\upsilon^{t-1},H_t}(\cdot)$, is, modulo computational aspects, observable at time t.

4.2.1 Numerical illustration

In the experiments to be reported, we dealt with the identity link function, so that the process was

$$[\zeta_t]_k \sim \text{Poisson}\left(\beta_k^0 + \sum_{s=1}^d \sum_{\ell=1}^L \beta_{k\ell}^s [\zeta_{t-s}]_\ell\right),\ t \geq 1, 1 \leq k \leq L$$

where d is the memory depth, L is the number of locations, and $[\zeta_t]_k$ is the k-th coordinate of response at time t.

Generating parameter β. Since Φ is the identity, selection of β should ensure nonnegativity of the quantities

$$\beta_k^0 + \sum_{s=1}^{d}\sum_{\ell=1}^{L} \beta_{k\ell}^s [\zeta_{t-s}]_\ell,\ k \le L,$$

whatever be nonnegative integral $\zeta_s \in \mathbf{R}^L$. To this end, we restrict β to be nonnegative. Besides this, with positive β_i^0, in order for the process not to explode we should have

$$\sum_{s=1}^{d}\sum_{\ell=1}^{L} \beta_{k\ell}^s < 1,\ k \le L.$$

Our generation was governed by two parameters $a > 0$ and $b \le 1$, specifically, we ensured that

$$\beta \ge 0\ \&\ \beta_k^0 = a\ \&\ \sum_{s=1}^{d}\sum_{\ell=1}^{L} \beta_{k\ell}^s = b,\ k \le L. \tag{19}$$

Generation was organized as follows: given d, L, we selected at random a "target" $\widehat{\beta} = \{\widehat{\beta}_k^0, \widehat{\beta}_{k\ell}^s, 1 \le s \le d, 1 \le k, \ell \le L\}$, and specified β as the closest to the target in ?? norm collection satisfying (19.

The set $\mathcal{B}$ – a priori localized for β – also was given by parameters a, b according to

$$\mathcal{B} = \{x = \{x_k^0, x_{k\ell}^s, 1 \le s \le d, 1 \le k, \ell \le L\} : x]geq0, x_k^0 \le 1.1a, \sum_{s=1}^{s} um_{\ell=1}^{L} x_{k\ell}^s \le 1.1b\,\forall k.\}$$

Building estimate $\widehat{\beta}$. Our estimate was exactly as explained in Special case of Section 2.2 with $\Lambda_t \equiv 1$, that is, $\widehat{\beta}(\upsilon^N)$ was an optimal solution to the Least Squares problem

$$\min_{x\in\mathcal{B}} \left[\frac{1}{2N}\sum_{t=1}^{N} \|H_t^T x - \zeta_t\|_2^2 = \frac{1}{2N}\sum_{t=1}^{N}\sum_{k=1}^{L}\left([\zeta_t]_k - x_k^0 - \sum_{s=1}^{d}\sum_{\ell=1}^{L} x_{k\ell}^s [\zeta_{t-s}]_\ell\right)^2\right].$$

On-line error bounds. The major goal of our experimentation was to get online upper bounds on the recovery errors. As it turned out, bounds stemming from Proposition 3.2 under the circumstances were quite poor – the values of the resulting error bounds typically were larger than the corresponding sizes of our a priori parameter's localizer

$\mathcal{B}$. To ge meaningful error bounds, we used the approach described in Section 3.2.1. Specifically, we

1. selected M of policies $\mathcal{R}^\iota$, $1 \le \iota \le M$, of generating L-dimensional row vectors L_t as deterministic functions of v^{t-1} and H_t, resulting in M affine functions

$$G^\iota_{v^N}(x) = \frac{1}{N}\sum_{t=1}^{N} L^\iota_t[H_t x - \xi_t]$$

   (L^ι_t are yielded by $\mathcal{R}^\iota$)

2. used Proposition 3.1.i and the preceding results of this section to build online $(1-\epsilon/M)$-reliable upper bounds on the quantities $|G^\iota_{v^N}(\beta$, specifically, the bounds

$$\begin{array}{c}\delta^\iota(v^N) = \min_i \left[\alpha_i \ln(2KM/\epsilon) + \sum_{t=1}^N \alpha_i \Psi_t(\alpha_i^{-1} N^{-1}\|L^\iota_t\|_1)\right] \\ \Psi_t(r) = \chi_t[\exp\{r\} - r - 1], \\ \chi_t = \max\limits_{x\in\mathcal{B}} \|H_t^T x\|_\infty\end{array} \tag{20}$$

   on $|G^\iota_{v^N}(\beta)|$

3. specified $(1-\epsilon)$-confidence interval

$$\Delta(v^N) = \{x \in \mathcal{B} : |G^\iota_{v^N}(x)| \le \delta^\iota(v^N), 1 \le \iota \le M\}$$

   – under the circumstances, this is a convex compact set which contains β with probability at least $1-\epsilon$

4. built induced confidence intervals Δ^0_k, $\Delta^s_{k\ell}$ on the entries in β, the endpoints of these intervals being the minima and the maxima of the corresponding entries in x taken over $x \in \Delta(v^N)$.
   By construction, the probability for all entries in β to be covered by the corresponding intervals is at least $1-\epsilon$, so that the maximal, over points from the intervals, deviations of the points from the corresponding entries in $\widehat{\beta}$ form a vector which is an $(1-\epsilon)$-reliable upper bound on the vector of magnitudes of the recovery errors.

The essence of the matter is how to specify the policies $\mathcal{R}^\iota$. The first κ of our policies were to take, as L_t^ι, the ι'th row in H_t, resulting in

$$[G^1_{\upsilon^N}(x); ...; G^\kappa_{\upsilon^N}(x)] = F_{\upsilon^N}(x).$$

The remaining policies were more sophisticated and inspired by the following consideration.

> Given index $k \leq \kappa$ of an entry in β and a "scale parameter" $\theta > 0$, let us impose on L_t's the restriction $\|L_t\|_1 \leq \kappa$, thus imposing upper bounds $\Psi_t(\alpha_i^{-1}N^{-1}\theta)$ on the terms $\Psi_t(\alpha_i^{-1}N^{-1}\|L_t^\iota\|_1)$ in (20), and under this assumption let us try to select L_t's in order to make the slope of the associated affine function $G_{\upsilon^N}(\cdot)$ as close as possible to k-th basic orth in $\mathbf{R}^\kappa$. Assume for a moment that we fully succeeded and the slope is exactly this basic orth. Then we get at our disposal inequality $|s - c| \leq \delta$ in variable s with hopefully small δ with is with ovewhelming probability satisdied by k-th entry of β.
>
> The outlined methodology reduces to finding reals L-dimensional row vectors L_t of $\|\cdot\|_1$-norm not exdceeding a given θ such that the vector
>
> $$\frac{1}{N}\sum_{t=1}^{T} H_t L_t^T$$
>
> is as close as possible to k-th basic orth e_k in $\mathbf{R}^\kappa$. The difficulty here is that L_t should be "non-anticipating" – they should be specified in terms of ν^{t-1} and H_t. The simplest policy of this type is the greedy policy
>
> $$\begin{array}{rcl} f_1 & = & e_k \\ L_t & \in & \mathrm{Argmin}_g \{\|f_t - N^{-1}H_t g\| : \|g\|_1 \leq \theta\} \\ f_{t+1} & = & f_t - N^{-1}L_t \end{array}$$
>
> In our experiments, the only one of the standard norms which worked in this greedy policy was $\|\cdot\|_2$, and these were the policies we used. To implement such a policy, we need to specify θ. This parameter is responsible for the tradeoff between the closeness of $e_k(\upsilon^N) := \frac{1}{N}\sum_{t=1}^{T} H_t L_t^T$ to our target e_k and

online upper bound

$$\delta = \min_i \left[\alpha_i \ln(2KM/\epsilon) + \sum_{t=1}^{N} \alpha_i \Psi_t(\alpha_i^{-1} N^{-1} \theta) \right]$$

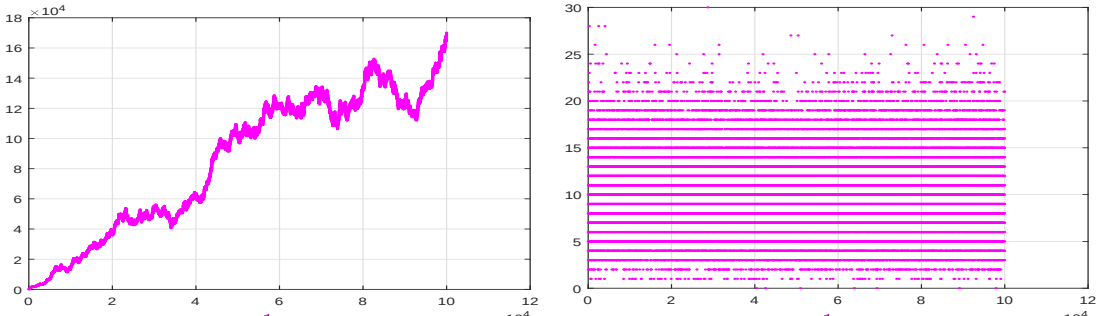
on $|e_k^T(\upsilon^n)\beta - c_k(\upsilon^N)|$ we end up with. Resolving this tradeoff analytically seems to be an impossible task, but we can use the same computation-oriented approaxh as with optimizing in the scale parameter α, specifically, select in advance a finite grid $\Theta = \{\theta_k : k \leq K\}$ of values of θ and run K greedy policies parameterized by $\theta \in \Theta$

In our implementation, we selected a K-element set Θ on the positive ray and, on the top of already explained κ policies $\mathcal{R}^\iota$, $\iota \leq \kappa$, given by $L_t = \mathrm{Row}_\iota(H_t)$, used κK policies more, associating with every $k \leq \kappa$ K aforementioned greedy policies with common target e_k and different values of θ running through Θ.

Implementation and numerical results. In the experiments to be reported, we used $d = L = 5$, resulting in $\kappa = 130$, $N = 100,000$, $\epsilon = 0.01$, and

$$\Gamma = \{10^{0.25i-4}, 0 \leq i \leq 36\}, \ \Theta = \{0.5, 0.75, 1.00, 1.25, 2\}$$

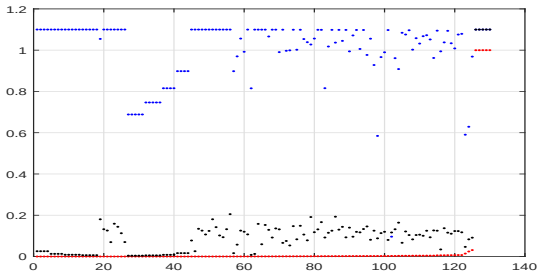
Numerical results. We present numerical results for two (in fact, quite representative) experiments. In the first (Experiment A) we used $a = 1$, $b = 1$, resulting in self-exciting process with growing with time t total, over locations, number $\omega_t^{\text{total}} := \sum_{k=1}^{L} \omega_{tk}$ of events at time t. In the second (Experiment B) we used $a = 1$, $b = 0.5$, resulting in stable process:



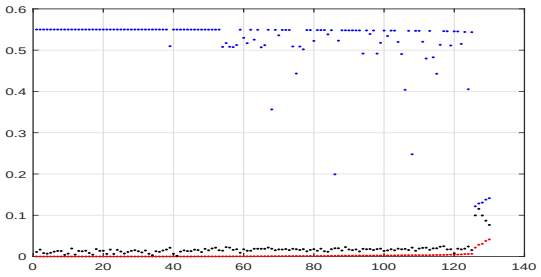
Experiment A, $\frac{1}{LN}\sum_t \omega_t^{\text{total}} \approx 1.7\text{e}4$ Experiment B, $\frac{1}{LN}\sum_t \omega_t^{\text{total}} = 2.00$

Sample trajectories of ω_t^{total} vs. t

Recovery errors. Here is the graphical and numerical data on actual recovery errors and online bounds on these errors:



Experiment A



Experiment B

Coordinate-wise recovery errors and online upperbounds on these errors
red dots: actual coordinatewise recovery errors; blue dots: basic online bounds;
black dots: advanced online bounds. Ordering of coordinates makes the errors nondecreasing

	min	median	mean	max
actual errors	0.0000/0.0000	0.0005/0.0004	0.0402/0.0018	1.0000/0.0312
bounds, basic	0.0962/0.0962	1.0775/1.0725	1.0045/1.0007	1.100/1.100
bounds, advanced	0.0041/0.0041	0.0983/0.0932	0.1215/0.0824	1.100/0.2053

Experiment A, coordinate-wise errors and their online bounds

	min	median	mean	max
actual errors	0.0000/0.0000	0.0007/0.0006	0.0024/0.0013	0.0413/0.0064
bounds, basic	0.1216/0.1991	0.5485/0.5487	0.5140/0.5293	0.5500/0.5500
bounds, advanced	0.0018/0.0018	0.0161/0.0158	0.0183/0.0152	0.1155/0.0249

Experiment B, coordinate-wise errors and their online bounds

On the plots and in the tables, "basic" online bounds stem from the κ initial policies $\mathcal{R}^\iota$, $\iota \leq \kappa$, while "advanced" bounds stem from all $\kappa(K+1)$ policies described above.

In the tables, the first number in a cell represents the "column quantity", say, the median, taken over all $\kappa = 130$ coordinate-wise errors and their upper bounds, while the second number corresponds to recovery errors/bounds on the errors for 125 "location to location influence" parameters $\beta^s_{k\ell}$. The reason for this distinction is that in Experiment A the "birth rates" β^0_k do not admit nontrivial estimates. Indeed, with our setup, these birth rates are equal to 1, while the typical observation in the unstable case is of order of 10^4; of course, there is absolutely no way to recover birthrates of order of 1 on the "background" of magnitude 10^4. In Experiment A the estimates of birth rates were nearly zero; this quite natural fact is responsible for the "large" – equal to 1 – uniform recovery error in

observed in this experiment. As we see from the corresponding table, the uniform norm of recovering all but 5 "birth rate" coefficients in β is 0.03 - 30 times smaller than the uniform norm of recovering the entire β. Similar albeit less profound, phenomenon takes place in Experiment B.

From the above data it is clear that utilizing greedy policies improved significantly the quality of online error bounds. While even after this improvement the bounds are much larger than the actual errors, these bounds seem to yield some information.

Upper-bounding $\|F_{\upsilon^N}(\beta)\|_\infty$. Our experiments show that the online upper bound on $\|F_{\upsilon^N}(\beta)\|_\infty$ is within factor 5-10 of the actual value of the quantity.

Predictive power. One way to utilize an estimate $\widehat{\beta}$ of β is to use the observations obtained so far to predict the states in the cells in the future. Specifically, given observations $\omega^t := \{\omega_{\tau k} : \tau \le t, k \le L\}$, and prediction step $p \ge 1$, we can write a simple recurrence specifying the conditional, ω^t given, expectations $\bar{\omega}_{t+p,k}$ of the states $\omega_{t+p,k}$ and use these expectations to build confidence intervals for $\omega_{t+p,k}$ *as if* the conditional, ω^t given, distributions of these random variables were Poisson($\bar{\omega}_{t+p,k}$) ("as if" reflects the fact that the actual conditional distributions in question provably are Poisson($\bar{\omega}_{t+p,k}$) only when $p = 1$). To write down the aforementioned recurrence, we need to know the true parameters β of the process; replacing these parameters with their estimates $\widehat{\beta}$, we get "empirical" confidence intervals for future states. This is how this prediction worked at testing sample:

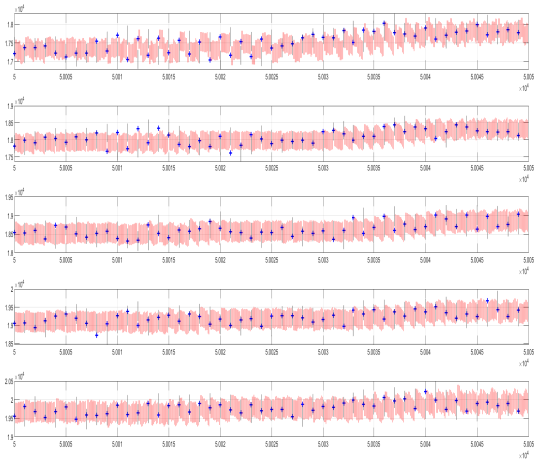
p	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
1	0.049	0.050	0.050	0.048	0.050
2	0.049	0.051	0.050	0.050	0.050
3	0.088	0.054	0.054	0.054	0.068
4	0.088	0.056	0.057	0.058	0.113
5	0.135	0.063	0.063	0.061	0.116
6	0.137	0.069	0.068	0.080	0.142
7	0.166	0.075	0.070	0.084	0.162
8	0.170	0.080	0.077	0.093	0.170
9	0.194	0.087	0.081	0.106	0.191
10	0.198	0.093	0.088	0.112	0.205
	0.322	0.156	0.182	0.184	0.223

Experiment A

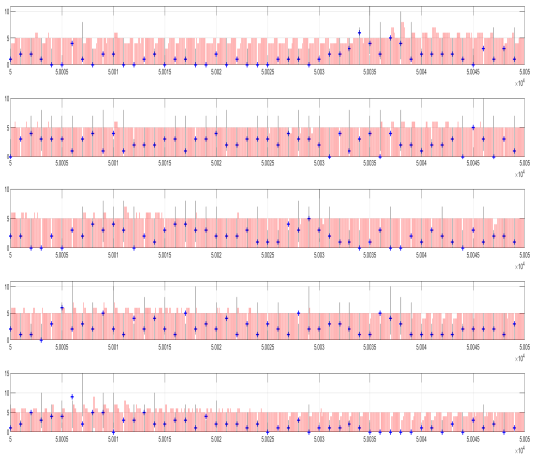
p	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
1	0.015	0.015	0.014	0.014	0.014
2	0.016	0.015	0.014	0.015	0.014
3	0.016	0.015	0.015	0.015	0.020
4	0.017	0.015	0.015	0.016	0.021
5	0.021	0.016	0.015	0.017	0.021
6	0.021	0.017	0.018	0.018	0.023
7	0.021	0.018	0.018	0.019	0.024
8	0.023	0.018	0.018	0.019	0.025
9	0.025	0.019	0.019	0.019	0.026
10	0.025	0.019	0.019	0.019	0.026
	0.202	0.183	0.180	0.181	0.203

Experiment B

Average, over 100,000-element testing sample, frequency for p-step-ahead empirical 0.95-confidence intervals not to cover the actual states of locations k, $1 \le k \le 5$. Last row: trivial prediction with $\bar{\omega}_{t+p,k}$ set to ω_{tk} ("tomorrow will be same as today")



Experiment A



Experiment B

Prediction on time interval $50,000 \le t \le 50,050$. Top to bottom: locations. Blue crosses: actual states; 10 orange vertical segments: in-between the crosses: 0.95-confidence intervals with prediction steps 1,2,...,10; black vertical segments: confidence intervals yielded by trivial prediction

5 More illustrations: Nonlinear dynamics

5.1 Discrete time nonlinear dynamics

The initial form of the problem we are interested in is as follows:

> At times $t = 1, 2, ...$ we observe the states $x_t \in \mathbf{R}^d$ of a discrete time dynamical system evolving according to
>
> $$\mathbf{E}_{|x^{t-1}}\{x_t\} = \Phi_t(\sum_{\ell=1}^{\kappa} \beta_\ell \phi_\ell^t(x^{t-1})), t = 1, 2, ... \tag{21}$$
>
> where $\Phi_t(z) : \mathbf{R}^d \to \mathbf{R}^d$ are given continuous monotone vector fields, $x^s = (x_1, ..., x_s)$, $\phi_\ell^t(\cdot) : \mathbf{R}^{(t-1)d} \to \mathbf{R}^d$ are known mappings, $\mathbf{E}_{|x^s}$ is the conditional, given x^s, expectation, and $\beta = \{\beta_\ell : 1 \le \ell \le \kappa\}$ is the vector of unknown parameters known to belong to given convex compact set $\mathcal{B} \subset \mathbf{R}^\kappa$. Given observations on time horizon $t \le N$, we want to recover β.

Setting $\zeta_s = x_s$, $H_t = [\phi_1^t(x^{t-1}); \phi_2^t(x^{t-1}); ...; \phi_\kappa^t(x^{t-1})]$, the problem reduces to GGLM here the predictors H_t are known deterministic functions of responses $x_1, ..., x_{t-1}$.

5.2 Discreteized continuous time nonlinear dynamics

Here the initial form of the problem we are interested in is as follows:

> At times $t = 1, 2, ...$ we observe the states $x_t \in \mathbf{R}^d$ of a discrete time dynamical system evolving according to
>
> $$\mathbf{E}_{|x^{t-1}}\{x_t\} = x_{t-1} + \Phi_t(\sum_{\ell=1}^{\kappa} \beta_\ell \psi_\ell^t(x^{t-1})), t = 1, 2, ... \tag{22}$$
>
> where x_0 is given, $\Phi_t(z) : \mathbf{R}^d \to \mathbf{R}^d$ are given continuous monotone vector fields, $x^s = (x_0, ..., x_s)$, $\psi_\ell^t(\cdot) : \mathbf{R}^{td} \to \mathbf{R}^d$ are known mappings, $\mathbf{E}_{|x^s}$ is the conditional, given x^s, expectation, and $\beta = \{\beta_\ell : 1 \le \ell \le \kappa\}$ is the vector of unknown parameters known to belong to given convex compact set $\mathcal{B} \subset \mathbf{R}^\kappa$. Given observations on time horizon $t \le N$, we want to recover β.

The rationale behind the model is as follows: In the nature there exists a continuous time dynamical system given, informally speaking, by

$$\mathbf{E}_{|y[0,\tau]}\{\dot{y}(\tau)\} = \Psi(\tau, \sum_{\ell=1}^{\kappa} \beta_\ell \psi_\ell(\tau, y[0, \tau])),$$

where $\tau \ge 0$ is continuous time, $y(\tau) \in \mathbf{R}^d$ is the state at time τ, $y[0, \tau]$ is the trajectory of states on the time window $0 \le s \le \tau$, $\psi_\ell(\tau, y[0, \tau])$ are given mappings taking values in $\mathbf{R}^d$, and $\Psi(\tau, y) : \mathbf{R}_+ \times \mathbf{R}^d \to \mathbf{R}^d$ is monotone in y. System (22) is the Euler discretization of the latter continuous time system, so that $x(t) = y(th)$, $\Phi_t(\cdot) = h\Psi(th, \cdot)$, and $\phi_\ell^t(x^{t-1})$ are approximations of $\psi_\ell(th, y[0, th])$.

Coming back to (22), let us treat as observation at time $t = 1, 2, ...$ the vector $\zeta_t = x_t - x_{t-1}$. Then (22) becomes

$$\mathbf{E}_{|\zeta^{t-1}}\{\zeta_t\} = \Phi_t(\sum_{\ell} \beta_\ell \overline{\phi}_\ell^t(\zeta^{t-1})),\ \overline{\phi}_\ell^t(\zeta^{t-1}) = \phi_\ell^t((x_0 + \zeta_1, x_0 + \zeta_1 + \zeta_2, ..., x_0 + \zeta_1 + ... + \zeta_{t-1})),$$

and we, same as in the previous section, arrive at GGLM where the predictors H_t are deterministic functions of the responses $\zeta_1, ..., \zeta_{t-1}$.

Note that in the special case where Φ_t is proportional to the identity mapping: $\Phi_t(z) =$

$h_t z$ with some $h_t > 0$, which still allows (22) to be a highly nonlinear dynamical system, a simpler reduction to GGLM is possible: setting $\beta_0 = 1$ and $\phi_0^t(x^{t-1}) = h_t^{-1} x_{t-1}$, (22) becomes

$$\mathbf{E}_{|x^{t-1}}\{x_t\} = \sum_{\ell=0}^{\kappa} \beta_\ell \overline{\phi}_\ell^t(x^{t-1})), t = 1, 2, ...$$

where $\overline{\phi}_\ell^t(\cdot) = h_t \phi_\ell^t(\cdot)$, $0 \le \ell \le \kappa$, which, up to the range of index ℓ, is (21), and we already know how to pose the latter model as a GGLM.

5.3 Illustration: Recovering parameters of Lotka-Volterra model

The classical Lotka-Volterra predator-prey model reads

$$\begin{array}{rcl} \dot{x}(\tau) & = & zx(\tau) - bx(\tau)y(\tau) \\ \dot{y}(\tau) & = & -cy(\tau) + dx(\tau)y(\tau) \end{array}$$

where $x(\tau)$ and $y(\tau)$ are sizes of prey and predator populations, and a, b, c, d are positive parameters. Consider randomly perturbed discrete time version of this model:

$$\begin{array}{rcl} x_0 & = & \bar{x},\ y_0 = \bar{y} \\ x_t & = & x_{t-1} + h\left[a_t x_{t-1} - b_t x_{t-1} y_{t-1}\right] \\ y_t & = & y_{t-1} + h\left[-c_t y_{t-1} + d_t x_t y_t\right] \end{array}$$

where $h \ll 1$ is resolution in continuous time, and $\{a_t, b_t, c_t, d_t\}_{t \ge 1}$ are random nonnegative sequences such that the conditional, $\bar{x}$, $\bar{y}$, $\{a_s, b_s, c_s, d_s\}_{1 \le s < t}$ given, expectations of a_t, b_t, c_t, d_t are, for every t, some a, b, c, d. We are given $[\bar{x}, \bar{y}]$ in advance, and at time t, $1 \le t \le N$, observe x_t, y_t; our goal is to recover a, b, c, d. Setting

$$\zeta_t = [x_t; y_t], H_t = \left[\begin{array}{cc} x_{t-1} & 0 \\ x_{t-1}y_{t-1} & 0 \\ 0 & x_{t-1}y_{t-1} \\ 0 & y_{t-1} \end{array}\right],$$

$$\beta = [1 + ah; -bh; ch; dh],\ \mathcal{B} = \{\beta \in \mathbf{R}^4 : 0 \le \beta_i \le B, 1 \le i \le 4\}$$

where $x_0 = \bar{x}$, $y_0 = \bar{y}$ and B is an a priori upper bound on the entries in the just defined β, and specifying observation of time t as $\upsilon_t = (\zeta_t, H_t)$, for every $t \geq 1$ the conditional, υ^{t-1}, H_t given, expectation of response ζ_t is $H_t^T \beta$. We arrive at GGLM with the identity link functions.

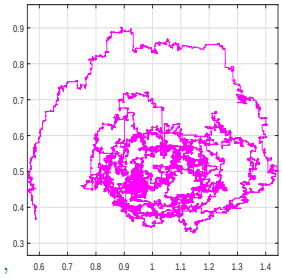
Numerical illustration. In the illustration to follow, we used $\bar{x} = \bar{y} = 1$, $a = 2/3$, $b = 4/3$, $c = d = 1$, $h = 0.0005$, $N = 100,000$, and

$$a_t = a\exp\{-0.5\theta^2 + \theta\xi_{t,a}\},\ b_t = b\exp\{-0.5\theta^2 + \theta\xi_{t,b}\},\ c_t = c\exp\{-0.5\theta^2 + \theta\xi_{t,c}\},\ d_t = d\exp\{-0.5\theta^2 + \theta\xi_{t,d}\}$$

with $\theta = 1.5$ and independent of each other and across t standard (zero mean, unit variance) Gaussian $\xi_{t,a}, \xi_{t,b}, \xi_{t,c}, \xi_{t,d}$. Sample trajectory and recovery errors are presented below.

	a	b	c	d
mean	0.0238	0.0445	0.0368	0.0329
median	0.0164	0.0371	0.0329	0.0255
max	0.1080	0.1870	0.1057	0.1120

Recovery errors, data over 100 simulations



Sample trajectory

References

[1] Alan Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

[2] Anatoli Juditsky, Arkadi Nemirovski, Liyan Xie, and Yao Xie. Convex parameter recovery for interacting marked processes. *IEEE Journal on Selected Areas in Information Theory*, 2020.

[3] Anatoli B Juditsky and AS Nemirovski. Signal recovery by stochastic optimization. *Automation and Remote Control*, 80(10):1878–1893, 2019.

[4] Peter McCullagh. *Generalized linear models*. Routledge, 2018.

[5] Yu Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optimization Methods and Software*, 9(1-3):141–160, 1998.