Dynamic Survival Analysis with Controlled Latent States

Linus Bleistein [*1 2 3] **Van-Tuan Nguyen** [*1 4 5] **Adeline Fermanian** [4] **Agathe Guilloux** [1 2]

Abstract

We consider the task of learning individual-specific intensities of counting processes from a set of static variables and irregularly sampled time series. We introduce a novel modelization approach in which the intensity is the solution to a controlled differential equation. We first design a neural estimator by building on neural controlled differential equations. In a second time, we show that our model can be linearized in the signature space under sufficient regularity conditions, yielding a signature-based estimator which we call CoxSig. We provide theoretical learning guarantees for both estimators, before showcasing the performance of our models on a vast array of simulated and real-world datasets from finance, predictive maintenance and food supply chain management.

1. Introduction

Time-to-event data is ubiquitous in numerous fields such as meteorology, economics, healthcare and finance. We typically want to predict when an event - which can be a catastrophic earthquake, the burst of a housing bubble, the onset of a disease or a financial crash - will occur by using some prior historical information (Ogata, 1988; Bacry et al., 2015; Bussy et al., 2019). This general problem encompasses many settings and in particular survival analysis, where every individual experiences at most one event (Cox, 1972).

For an individual i, we have access to several event times $T_1^i < T_2^i < \dots$ and features $\mathbf{W}^i \in \mathbb{R}^s$ measured at time 0. For instance, in neurology, one might consider the onset times of a series of seizures (Rasheed et al., 2020) and $\mathbf{W}^i$ summarizes unchanging characteristics of the individual (age, gender, ethnicity, ...). The physician's goal is to determine whether an individual has a high probability to experience a seizure at time t given their characteristics. Such a task is most often addressed by modelling the individual specific intensity of a counting process of the form $\sum_{j\geq 1} \mathbb{1}_{T_j^i \leq t}$, using for instance Cox models (Cox, 1972; Aalen et al., 2008; Kvamme et al., 2019) or Hawkes processes in the case of self-exciting processes (Bacry et al., 2015). Recent advances in the field have also enriched these models using deep architectures (Mei & Eisner, 2017; Kvamme et al., 2019; Omi et al., 2019; Chen et al., 2021; Groha et al., 2020; Shchur et al., 2021; De Brouwer et al., 2022; Tang et al., 2022). Once learnt, the intensity of the process can be used to predict occurrence times of future events or rank individuals based on their relative risks.

Learning with Time-dependent Data. More realistically, in addition to the static features $\mathbf{W}^i$, we often have access to time-dependent features along with their sampling times

$$\mathbf{X}^i =: \{(\mathbf{X}^i(t_1), t_1), \dots, (\mathbf{X}^i(t_K), t_K)\} \in \mathbb{R}^{d\times K},$$

where $D = \{t_1, \dots, t_K\} \subset [0, \tau]$ is a set of measurement times and τ the end of study. Taking again the example of seizure prediction, the time-dependent features may represent some measurements made by a wearable device, as done for instance by Dumanis et al. (2017). Taking both the static and time-dependent information into account is crucial when making predictions. This setting calls for highly flexible models of the intensity which take into account the stream of information carried by the longitudinal features.

From joint models to ODE-based methods. This problem has been tackled by the bio-statistics community, in particular using joint models that concurrently fit parametric models to the trajectory of the longitudinal features and the intensity of the counting process (Ibrahim et al., 2010; Crowther et al., 2013; Proust-Lima et al., 2014; Long & Mills, 2018). Popular implementations include `JMBayses` (Rizopoulos, 2016). While being highly interpretable, they do not scale to high-dimensional and fre-

[*]Equal contribution [1]Inria Paris, F-75015 Paris, France [2]Centre de Recherche des Cordeliers, INSERM, Université de Paris, Sorbonne Université, F-75006 Paris, France [3]LaMME, UEVE and UMR 8071, Paris Saclay University, F-91042, Evry, France [4]LOPF, Califrais' Machine Learning Lab, Paris, France [5]Laboratoire de Probabilités, Statistique et Modélisation, LPSM, Univ. Paris Cité, F-75005, Paris, France. Correspondence to: Linus Bleistein <linus.bleistein@inria.fr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

quently measured data, despite some recent algorithmic advances (Hickey et al., 2016; Murray & Philipson, 2022; Rustand et al., 2024) adapted to moderate dimension (up to $\simeq 5$ longitudinal features).

Modern deep methods, that can encode complex and meaningful patterns from complex data in latent states, offer a particularly attractive alternative for this problem. However, the literature bridging the gap between deep learning and survival analysis is scarce. Notably, Lee et al. (2019) tackle this problem by embedding the time-dependent data through a recurrent neural network combined with an attention mechanism. They then use this embedding in a discrete-time setting to maximize the likelihood of dying in a given time-frame conditional on having survived until this time. Moon et al. (2022) combine a probabilistic model with a continuous-time neural network, namely the ODE-RNNS of Rubanova et al. (2019) in a similar setup.

Modelling Time Series with Controlled Latent States. Building on the increasing momentum of differential equation-based methods for learning (Chen et al., 2018; De Brouwer et al., 2019; Rubanova et al., 2019; Chen et al., 2021; Moon et al., 2022; Marion et al., 2022), we propose a novel modelling framework in which the unknown intensity of the counting process is parameterized by a latent state driven by a controlled differential equation (CDE). Formally, we let the unknown intensity of the counting process of individual i depend on their covariates $\mathbf{W}^i$ and an unobserved process $x^i : [0, \tau] \to \mathbb{R}^d$ that is the continuous unobserved counterpart of the time series $\mathbf{X}^i$, i.e., $(\mathbf{X}^i(t), t) = x^i(t)$ for all $t \in D$. We model the intensity (i.e. the instantaneous probability of experiencing an event — see Section 2.2) by setting

$$\lambda_\star^i\big(t \,|\, \mathbf{W}^i, (x^i(s))_{s\le t}\big) = \exp\big(z_\star^i(t) + \beta_\star^\top \mathbf{W}^i\big), \qquad (1)$$

where the dynamical latent state $z_\star^i(t) \in \mathbb{R}$ is the solution to the CDE

$$dz_\star^i(t) = \mathbf{G}_\star\big(z_\star^i(t)\big)^\top dx^i(t) \qquad (2)$$

with initial condition $z_\star^i(0) = 0$ driven by x^i. Here, the vector field $\mathbf{G}_\star : \mathbb{R} \to \mathbb{R}^d$ and $\beta_\star \in \mathbb{R}^s$ are both unknown. This means that the latent dynamics are common between individuals, but are driven by individual-specific data, yielding individual-specific intensities. Such a modelling strategy is reminiscent of state space models, which embed times series through linear controlled latent differential equations (Gu et al., 2022; Cirone et al., 2024). Our framework is introduced in more detail later.

Contributions. In an effort to provide scalable and efficient models for event-data analysis, we propose two novel estimators. We first leverage neural CDEs (Kidger et al., 2020), which directly approximate the vector field $\mathbf{G}_\star$ with a neural vector field $\mathbf{G}_\psi$. In a second time, following Fermanian et al. (2021) and Bleistein et al. (2023), we propose to linearize the unknown dynamic latent state $z_\star^i(\cdot)$ in the signature space. Informally, this means that at any time t, we have the simplified expression

$$z_\star^i(t) \approx \alpha_{\star,N}^\top \mathbf{S}_N(x_{[0,t]}^i)$$

where $\alpha_{\star,N}$ is an unknown finite-dimensional vector and $\mathbf{S}_N(x_{[0,t]}^i)$ is a deterministic transformation of the time series x^i observed up to time t called the *signature transform*. Notice that in this form, the vector $\alpha_{\star,N}$ does not depend on t and can hence be learned at any observation time. We obtain theoretical guarantees for both models ; for the second model in particular, we state a precise decomposition of the variance and the discretization bias of our estimator, which crucially depends on the coarseness of the sampling grid D. Finally, we benchmark both methods on simulated and real-world datasets from finance, healthcare and digital food retail, in a survival analysis setting. Our signature-based estimator provides state-of-the-art results.

Summary. Section 2 details our theoretical framework. In Section 3, we state theoretical guarantees for our model. Lastly, we conduct a series of experiments in Section 4 that displays the strong performances of our models against an array of benchmarks. All proofs are given in the appendix. The code is available at `https://github.com/LinusBleistein/signature_survival`.

2. Modelling Point Processes with Controlled Latent States

2.1. The Data

In practice, an individual can be censored (for example after dropping out from a study) or cannot experience more than a given number of events. To take this into account, we introduce $Y^i : [0, \tau] \to \{0, 1\}$ the at-risk indicator function, which equals 1 when the individual i is still at risk of experiencing an event. Together with Y^i, we define

$$N^i(t) := \sum_{j\ge 1} \mathbb{1}_{T_j^i \le t} Y^i(T_j^i)$$

as the stochastic process counting the number of events experienced by individual i up to time t and while $Y^i(T_j^i) = 1$. Our dataset

$$\mathcal{D}_n := \{\mathbf{X}^i, \mathbf{W}^i, Y^i(t), N^i(t),\, 0 \le t \le \tau\}$$

consists of n i.i.d. historical observations up to time τ. Our setup can be extended to individual-dependent grids $(D^i)_{i=1}^n$, but we choose to focus on the former setting for the sake of clarity. The individual specific time series are only observed as long as the individual is at risk. We first make an assumption on the time series.

Assumption 1. For every individual $i = 1, \dots, n$, there exists a continuous path of bounded variation $x^i : [0, \tau] \to \mathbb{R}^d$ satisfying, for all $0 \le s < t \le \tau$,

$$\left\|x^i\right\|_{1\text{-var},[s,t]} := \sup_D \sum_k \left\|x^i(t_{k+1}) - x^i(t_k)\right\| \le L_x|t-s|$$

where $\|\cdot\|$ is the Euclidean norm and the supremum is taken over all finite dissections $D = \{s = t_1 < \dots < t_K = t\}$. The time series $\mathbf{X}^i$ is a discretization of x^i on the grid D.

Remark that this assumption implies that the paths are L_x-Lipschitz. We now state a supplementary assumption on the static features.

Assumption 2. There exists a constant $B_{\mathbf{W}} > 0$ such that for every $i = 1, \dots, n$, $\left\|\mathbf{W}^i\right\|_2 \le B_{\mathbf{W}}$.

2.2. Modelling Intensities with Controlled Differential Equations

Intensity of a counting process. We define the individual-specific intensity $\lambda_\star^i\big(t \,|\, \mathbf{W}^i, x^i_{[0,t]}\big)$ of the underlying counting process, which we will simply write $\lambda_\star^i(t)$ in the following, as

$$\lambda_\star^i(t) := \lim_{h \to 0^+} \frac{1}{h} \mathbb{E}\big(N^i(t+h) - N^i(t) \,|\, \mathcal{F}_t^i\big)$$

where $\mathcal{F}_t^i$ is the past information at time t which includes $\mathbf{W}^i$ and $x^i_{[0,t]}$ (Aalen et al., 2008).

Controlled Dynamics. Controlled differential equations are a theoretical framework that allows to generalize ODEs beyond the non-autonomous regime (Lyons et al., 2007). Recall that a non-autonomous ODE is the solution to

$$dz(t) = \mathbf{F}(z(t), t)dt$$

with a given initial value $z(0) = z_0 \in \mathbb{R}^p$. Here, the vector field $\mathbf{F} : \mathbb{R}^p \times [0, +\infty[\to \mathbb{R}^p$ depends explicitly on $t \ge 0$, allowing for time-varying dynamics unlike autonomous ODEs whose dynamics remain unchanged through time. Controlled differential equations can be seen as a generalization of non-autonomous ODEs. They allow for the vector field to depend explicitly on the values of another $\mathbb{R}^d$-valued function $x : [0, 1] \to \mathbb{R}^d$ through

$$dz(t) = \tilde{\mathbf{F}}(z(t), x(t))dt$$

thus encoding even richer dynamics. Formally, a CDE writes

$$\begin{aligned} dz(t) &= \mathbf{G}\big(z(t)\big)dx(t) \\ z(0) &= z_0 \in \mathbb{R}^p \end{aligned}$$

where $\mathbf{G}$ is a $\mathbb{R}^{p\times d}$-valued vector field. Existence and uniqueness of the solution is ensured under regularity conditions on $\mathbf{G}$ and x by the Picard-Lindelhöf Theorem (see Theorem A.1). The following assumption is needed in order to ensure that the function

$$\lambda_\star^i(t) = \exp\left(z_\star^i(t) + \beta_\star^\top \mathbf{W}^i\right),$$

where the dynamical latent state $z_\star^i(t) \in \mathbb{R}$ is the solution to the CDE

$$dz_\star^i(t) = \mathbf{G}_\star\big(z_\star^i(t)\big)^\top dx^i(t)$$

with initial condition $z_\star^i(0) = 0$ driven by x^i is well-defined.

Assumption 3. The vector field $\mathbf{G}_\star : \mathbb{R} \to \mathbb{R}^d$ defining $\lambda_\star^i$ in Equation (2) is $L_{\mathbf{G}_\star}$-Lipschitz; $\beta_\star$ is such that $\|\beta_\star\|_2 \le B_{\beta,2}$, $\|\beta_\star\|_1 \le B_{\beta,1}$ and $\|\beta_\star\|_0 \le B_{\beta,0}$, where $B_{\beta,2}, B_{\beta,1}, B_{\beta,0} > 0$ are constants.

Under these assumptions, the intensity is bounded at all times.

Lemma 2.1 (A bound on the intensity)**.** *For every individual $i = 1, \dots, n$ and all $t \in [0, \tau]$, the log intensity $\log \lambda_\star^i(t)$ is upper bounded by*

$$B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x t \exp\left(L_{\mathbf{G}_\star} L_x t\right)$$

almost surely.

This is a direct consequence of Lemma 3.3 in Bleistein & Guilloux (2024). Remark that $\|\mathbf{G}_\star(0)\|_{\text{op}} < \infty$ since the vector field is Lipschitz and hence continuous.

Remark 2.2. By differentiation, one can see that the intensity itself satisfies a so-called controlled Volterra differential equation (Lin & Yong, 2020). Indeed, differentiating the intensity $\lambda_\star^i$ yields the CDE

$$d\lambda_\star^i(s) = \lambda_\star^i(s)\mathbf{G}_\star(z_\star^i(s))dx^i(s)$$

with initial condition $\lambda_\star^i(0) = \exp(\beta_\star^\top \mathbf{W}^i)$. Note that this CDE is path dependent, i.e., its vector field depends on the path $z_\star^i : [0, \tau] \to \mathbb{R}$.

Remark 2.3. This model enforces continuity of the intensity: indeed, the solution of a CDE inherits the regularity of its driving path. A possible solution to accommodate discontinuous intensity functions is to add a jump term to the generative CDE, which could then be learnt using neural jump ODEs (Jia & Benson, 2019).

2.3. Neural Controlled Differential Equations

Following the ideas of continuous time models, our first approach to learning the dynamics is to fit a parameterized intensity to this model by setting

$$\lambda_\theta^i(s) = \exp(\alpha^\top z_\theta^i(s) + \beta^\top \mathbf{W}^i),$$

where $z_\theta^i(s) \in \mathbb{R}^p$ is an embedding of the time series $\mathbf{X}^i$ parameterized by $\theta \in \mathbb{R}^v$ and $\alpha \in \mathbb{R}^p$ is a learnable parameter. We propose to use Neural Controlled Differential

Equations (NCDEs), a powerful tool for embedding irregular time series introduced by Kidger et al. (2020). NCDEs work by first embedding a time series $\mathbf{X}^i$ in the space of functions of bounded variation, yielding $x^{i,D} : [0,\tau] \to \mathbb{R}^d$, before defining a representation of the data through

$$dz_\theta(t) = \mathbf{G}_\psi\big(z_\theta(s)\big)dx^{i,D}(s)$$

with initial condition $z_\theta(0) = \mathbf{0}$. It is common practice to set $\mathbf{G}_\psi : \mathbb{R}^p \to \mathbb{R}^{p\times d}$ to be a small feed-forward neural network parameterized by ψ. The learnable parameters of this model are thus $\theta = (\alpha, \psi, \beta)$. In our setup, the embedding must be carefully chosen in order not to leak information from the future observations. Hence natural cubic splines, used in the original paper by Kidger et al. (2020), cannot be used and we resort to the piecewise constant interpolation scheme proposed by Morrill et al. (2021) and defined as $x^{i,D}(s) = (\mathbf{X}^i(t_k), s)$ for all $s \in [t_k, t_{k+1}[$. This yields a discretely updated latent state equal to

$$z_\theta^{i,D}(t_k) = z_\theta^{i,D}(t_{k-1}) + \mathbf{G}_\psi(z_\theta^{i,D}(t_{k-1}))\Delta\mathbf{X}^i(t_k)$$

where $\Delta\mathbf{X}^i(t_k) = \mathbf{X}^i(t_k) - \mathbf{X}^i(t_{k-1})$. This architecture has been studied under the name of *controlled ResNet* because of its resemblance with the popular ResNet (Cirone et al., 2023; Bleistein & Guilloux, 2024).

In order to provide theoretical guarantees, we restrict ourselves to a bounded set of NCDEs i.e. we consider a set of NCDE predictors

$$\Theta_1 = \{\theta \in \mathbb{R}^v \text{s.t.} \|\alpha\|_2 \le B_\alpha, \|\psi\| \le B_\psi, \|\beta\|_2 \le B_{\beta,2}\}$$

where the norm on ψ refers to the sum of ℓ_2 norms of the weights and biases of the neural vector field $\mathbf{G}_\psi$. This restriction is fairly classical in statistical learning theory (Bach, 2021).

2.4. Linearizing CDEs in the Signature Space

The Signature Transform. While neural controlled differential equations allow for great flexibility in representation of the time series, they are difficult to train and require significant computational resources. The signature is a promising and theoretically well-grounded tool from stochastic analysis, that allows for a parameter-free embedding of the time series. Mathematically, the signature coefficient of a function

$$x : t \in [0,\tau] \mapsto \big(x^{(1)}(t), \dots, x^{(d)}(t)\big)$$

associated to a word $I = (i_1, \dots, i_k) \in \{1, \dots, d\}^k$ of size k is the function

$$\mathbf{S}^I(x_{[0,t]}) := \int_{0<u_1<\dots<u_k<t} dx^{(i_1)}(u_1)\dots dx^{(i_k)}(u_k)$$

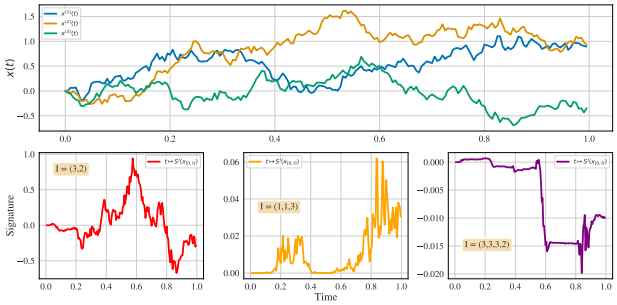

Figure 1: Sample path $x(t)$ of a 3-dimensional fractional Brownian motion on **top**, and three signature coefficients $\mathbf{S}^I(x_{[0,t]})$ associated to different words on the **bottom**.

which maps $[0,\tau]$ to $\mathbb{R}$. The integral is to be understood as the Riemann-Stieltjes integral. While the definition of the signature is technical, it can simply be seen as a feature extraction step. We refer to Figure 1 for an illustration. The truncated signature of order $N \ge 1$, which we write $\mathbf{S}_N(x_{[0,t]})$, is equal to the collection of all signature coefficients associated to words of size $k \le N$ sorted by lexicographical order. Finally, the infinite signature is the sequence defined through

$$\mathbf{S}(x_{[0,t]}) = \lim_{N\to+\infty} \mathbf{S}_N(x_{[0,t]}).$$

Learning with Signatures. Signatures are a prominent tool in stochastic analysis since the pioneering work of Chen (1958) and Lyons et al. (2007). They have recently found successful applications in statistics and machine learning as a feature representation for irregular time series (Kidger et al., 2019; Morrill et al., 2020; Fermanian, 2021; Salvi et al., 2021; Fermanian, 2022; Lyons & McLeod, 2022; Bleistein et al., 2023; Horvath et al., 2023) and a tool for analyzing residual neural networks in the infinite depth limit (Fermanian et al., 2021).

Signatures and CDEs. An appealing feature of signatures is their connection to controlled differential equations. Indeed, under sufficient regularity assumptions (Friz & Victoir, 2010; Fermanian et al., 2021; Bleistein et al., 2023; Cirone et al., 2023), the generative CDE (1) can be linearized in the signature space. Informally, this means that there exists a sequence $\alpha_\star$ such that for all $t \in [0,\tau]$ we have

$$z_\star^i(t) = \alpha_\star^\top \mathbf{S}(x^i_{[0,t]}).$$

The mathematical definition of $\alpha_\star$ is technical and we refer to Appendix A.4 for a formal statement and a discussion of the regularity assumptions. Hence, under the corresponding regularity conditions, the true intensity for individual i

writes

$$\lambda_\star^i(t) = \exp\left(\alpha_\star^\top \mathbf{S}(x^i_{[0,t]}) + \beta_\star^\top \mathbf{W}^i\right).$$

This motivates the use of the signature-based estimator

$$\lambda_\theta^{i,D}(t) := \exp\left(\alpha^\top \mathbf{S}_N(x^{i,D}_{[0,t]}) + \beta^\top \mathbf{W}^i\right),$$

where $\theta = (\alpha, \beta) \in \mathbb{R}^q \times \mathbb{R}^s$, $N \geq 1$ is treated as a hyperparameter and $x^{i,D}$ corresponds to the piecewise constant embedding of the observed time series $\mathbf{X}^i$ described previously. The integer $q = \frac{d^{N-1}-1}{d-1}$ is the size of the signature truncated at depth $N \geq 1$. The superscript in D emphasises the dependence of this estimator on the observation grid D.

Similarly to the NCDE-based estimator, we restrict ourselves to the bounded set of estimators

$$\Theta_2 = \{\theta \text{ s.t. } \|\alpha\| \leq B_\alpha, \|\beta\| \leq B_{\beta,2}\}.$$

2.5. Connections to Cox Models with Time-Varying Covariates

Cox models with time-varying covariates are the classical class of models (Therneau & Grambsch, 2000; Aalen et al., 2008; Zhang et al., 2018), where the individual specific hazard rate has the form $\lambda_\theta^i(t) = \lambda_0(t)\exp(\alpha_\star^\top \mathbf{X}^i(t) + \beta_\star^\top \mathbf{W}^i)$, where $\lambda_0 : [0,\tau] \to \mathbb{R}_+$ is called the *baseline hazard*.

For signature-based embeddings, recall that we compute the signature of a time-embedded time series $\mathbf{X}^i = \{(\mathbf{X}^i(t_1), t_1), \ldots, (\mathbf{X}^i(t_k), t_k)\}$. In fact, this amounts to

$$\begin{aligned}&\alpha^\top \mathbf{S}_N(x^{i,D}_{[0,t]})\\ &= \underbrace{\sum_{k=0}^N \alpha_k t^k}_{=\log\lambda_0(t)} + \underbrace{\alpha_{I_1}^\top \mathbf{X}^i(t) + \sum_{I\in I_2} \alpha_I \mathbf{S}^I(x^{i,D}_{[0,t]})}_{=\log \text{ of individual specific hazard rate}}\end{aligned}$$

where α_{I_1} is a subvector of α and $I_2 \subset \prod_{k=2}^N \{1,\ldots,d\}^k$. Hence our model can be interpreted as a generalized version of Cox models with time-varying covariates. A similar interpretation holds for NCDEs. We detail this link in Appendix A.6.

3. Theoretical Guarantees

3.1. The Learning Problem

For both models, the parameter θ can be fitted by likelihood maximization by solving

$$\hat\theta \in \arg\min_{\theta\in\Theta} \ell_n^D(\theta) + \text{pen}(\theta), \tag{3}$$

where $\Theta \in \{\Theta_1, \Theta_2\}$ depending on whether one uses signature or NCDE-based embeddings, pen : $\Theta \to \mathbb{R}_+$ is a penalty and $\ell_n^D(\theta)$ is equal to the negative log-likelihood of the sample $\mathcal{D}_n$ evaluated at θ.

Unless specified other, the following statements hold for both NCDEs and signature-based embeddings (up to different constants given explicitly in the proofs). Following Aalen et al. (2008), the negative log likelihood $\ell_n^D(\theta)$ of the sample writes

$$\frac{1}{n}\sum_{i=1}^n \int_0^\tau \lambda_\theta^{i,D}(s)Y^i(s)ds - \int_0^\tau \log\lambda_\theta^{i,D}(s)dN^i(s),$$

and we let

$$\ell_n^\star = \frac{1}{n}\sum_{i=1}^n \int \lambda_\star^i(s)Y^i(s)ds - \int \log\lambda_\star^i(s)dN^i(s)$$

be the true likelihood of the data. Our goal, in this section, is to obtain a bias-variance decomposition of the difference

$$\ell_n^D(\hat\theta) - \ell_n^\star$$

between the true likelihood and the likelihood of the learnt model.

3.2. A Risk Bound

Theorem 3.1 (Informal Risk Bound for the Signature Model)**.** *Consider the signature-based embedding. Let $\hat\theta$ be the solution of (3) with* $\text{pen}(\theta) = \eta_1\|\alpha\|_1 + \eta_2\|\beta\|_1$. *For any $N \geq 1$, we have with high probability and an appropriate choice of η_1, η_2 that*

$$\begin{aligned}\ell_n^D(\hat\theta) - \ell_n^\star \leq &\text{ Discretization bias} + \text{Approximation bias}\\ &+ \mathcal{O}\left(\sqrt{\frac{\log N d^N}{n}}\right) + \mathcal{O}\left(\sqrt{\frac{\log s}{n}}\right).\end{aligned}$$

For a formal statement, see Appendix B.4. We make a series of comments on this result.

1. This full risk bound can only be obtained for the signature-based model. It can also be extended to other types of penalty such as Ridge or Group Lasso (see for instance Nardi & Rinaldo (2008)) For NCDEs, we are able to give precise guarantees on the bias following Bleistein & Guilloux (2024), but a precise control of the variance term is out of reach.

2. The discretization bias is proportional to $|D| := \max_{i=1,\ldots,K} |t_i - t_{i-1}|$ and hence vanishes as sampling gets finer.

3. The approximation bias crucially depends on the regularity of the unknown tensor field $\mathbf{G}_\star$, and more precisely on the speed of decay of its derivatives, which can be seen as a measure of smoothness of the target function.

4. The regularity assumptions made on $\mathbf{G}_\star$ are not necessary to bound the approximation bias of the NCDE model: in this case, this bias term depends on the approximation capacities of the neural tensor field.

5. Remarkably, we obtain classical rates in $n^{-1/2}$ for the variance term. For signature based methods, fast rates in n^{-1} are yet to be obtained.

4. Experimental Evaluation

We now focus on the survival analysis setup. We hence let T^i be the unique time-of-event, which may eventually be censored, of individual i. Δ^i is the censorship indicator, equal to 1 if the individual experiences the event and to 0 otherwise.

4.1. Training Setup

We train on a dataset $\mathcal{D}_n$ of the same structure than described in Section 2.1 and learn the parameter $\hat{\theta}$ by solving the optimization problem (3). NCDEs are trained without penalization, while we use a mixture of elastic-net penalties

$$\text{pen}(\theta) := \eta_1 \text{pen}_{\text{EN}}(\alpha) + \eta_2 \text{pen}_{\text{EN}}(\beta)$$

for training the signature-based model, where $\text{pen}_{\text{EN}}(\cdot) = \gamma \|\cdot\|_1 + (1-\gamma)\|\cdot\|_2$. The hyperparameters (η_1, η_2, N) are chosen by cross-validation of a mixed metric equal to the difference between the C-index and the Brier score (see below) and we set $\gamma = 0.1$. We refer to Appendix C.1 for a detailed description of the training procedures. We evaluate our model's capacity to predict events in $[t, t+\delta t]$ by leveraging values of the longitudinal features up to t (see Figure 2) through a ranking metric and a calibration metric. This evaluation procedure is standard (Lee et al., 2019).

4.2. Metrics

We compute four metrics using the individual specific survival functions as estimated by our model with parameters θ. At time $t+\delta t$ for $\delta t > 0$ conditional on survival up to time t, and on observation of the longitudinal features up to time t, it is defined as

$$r^i_\theta(t, \delta t) = \mathbb{P}\Big(T^i > t + \delta t \,|\, T^i > t, (\mathbf{X}^i(s))_{\substack{s \le t \\ s \in D}}, \mathbf{W}^i\Big).$$

We describe its detailed computation in Appendix C.2.

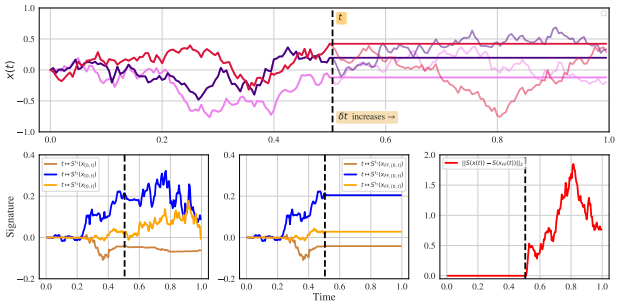

Figure 2: On the **top**, observed time series up to time t in bold colors and true time series in faded colors. When evaluating our models, we fill-forward the last observed value from t on. On the **bottom**, signatures of the true path (**left**), of the observed path (**center**) and difference in ℓ_2 norm (**right**) — $x_{FF}(t)$ denotes the filled-forward time series.

Time-dependent Concordance Index. Following Lee et al. (2019), we measure the discriminative power of our models by using a time-dependent concordance index $C(t, \delta t)$ that captures our models ability to correctly rank individuals on their predicted probability of survival. The concordance index $C(t, \delta t)$ is then finally computed as

$$\frac{\sum_{j=1}^{n}\sum_{i=1}^{n} \mathbb{1}_{r^i_\theta(t,\delta t) > r^j_\theta(t,\delta t)} \mathbb{1}_{T^i > T^j,\, T^j \in [t, t+\delta t], \Delta^j = 1}}{\sum_{j=1}^{n}\sum_{i=1}^{n} \mathbb{1}_{T^i > T^j,\, T^j \in [t, t+\delta t], \Delta^j = 1}}.$$

This metric captures the capacity of our model to discriminate between j and another individual i through the conditional probability of survival.

Brier Score. While the concordance index is a ranking-based measure, the Brier Score measures the accuracy in predictions by comparing the estimated survival function and the survival indicator function (Lee et al., 2019; Kvamme et al., 2019; Kvamme & Borgan, 2023). Formally, we define the Brier score $\text{BS}(t, \delta t)$ as

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{T^i \le t+\delta t, \Delta^i = 1} r^i_\theta(t, \delta t)^2 + \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{T^i > t+\delta t} (1 - r^i_\theta(t, \delta t))^2.$$

Contrarily to the C-index, the Brier score is a measure of calibration of the predictions: it measures the distance between the estimated survival function and the indicator function of survival on the interval $[t, t+\delta t]$.

Averaged performance. Additionally, we evaluate the average prediction performance of our models over a set

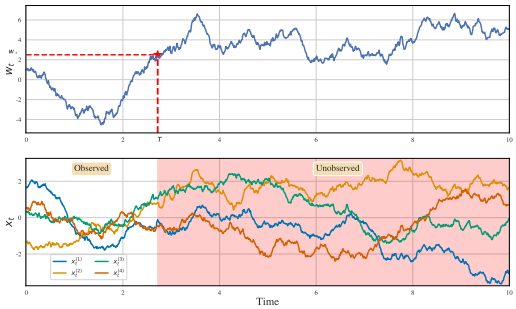

Figure 3: Time series $\mathbf{X}^i$ of a randomly picked individual on **bottom** and unobserved SDE $w^i(t)$ on the **top**. The red star indicates the first hitting time of the threshold value $w_\star = 2.5$.

of different prediction times. The averaged C-index and Brier score on the interval $[t_1, t_2]$ along with the window time δt are defined respectively as

$$\frac{1}{t_2 - t_1}\int_{t_1}^{t_2} C(s, \delta t)ds \quad \text{and} \quad \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} BS(s, \delta t)ds.$$

Comparison with static metrics. A crucial difference with static survival analysis metrics is that our metric only compares the individuals who experienced the event in this time window to all the ones who are still at risk at time t. This can lead to a C-index below 0.5 and Brier scores above 0.25 without the model being worse than random.

Additional metrics. We furthermore report AUC and weighted Brier score in Appendix D.

4.3. Methods

We propose three distinct methods. In addition to the signature-based model, which we call **CoxSig**, we also consider **CoxSig+** which adds the first value of the time series to the static features. This is motivated by the translation invariance of signatures (see discussion below). Our last method is the **NCDE** embedding of the longitudinal features. We benchmark our three models against a set of competing methods. All methods are detailed in Appendix C.1.

Time-Independent Cox Model. As a sanity check, we implement a simple Cox model with elastic-net penalty which uses the parameterized intensity $\lambda_\theta^i(t) = \lambda_0(t)\exp(\beta^\top \mathbf{W}^i)$ using `scikit-survival` (Pölsterl, 2020). This baseline allows to check whether our proposed methods can make use of the supplementary time-dependent information. If no static features are available, we use the first observed value of the time series, i.e., $\mathbf{W}^i = \mathbf{X}^i(0)$.

Name	n	d	Censoring	Avg. Times
Hitting time	500	5	Terminal (3.2%)	177
Tumor Growth	500	2	Terminal (8.4%)	250
Maintenance	200	17	Online (50%)	167
Churn	1043	14	Terminal (38.4%)	25

Table 1: Description of the 4 datasets we consider. The integer d is the dimension of the time series including the time channel. *Terminal* censoring means that the individuals are censored at the end of the overall observation period $[0, \tau]$ if they have not experienced any event. It is opposed to *online* censoring that can happen at any time in $[0, \tau]$. The reported percentage indicates the censoring level i.e. the share of the population that does not experience the event. The last column reports the average number of observations times over individuals.

Random Survival Forest (RSF). We use RSF (Ishwaran et al., 2008) with static features $\mathbf{W}^i$ as the only input. Similarly to our implementation of the Cox model, we use the first value of the time series as static features if no other features are available.

Dynamic DeepHit (Lee et al., 2019). DDH is a state-of-the-art method for dynamical survival analysis, that combines an RNN with an attention mechanism and uses both time dependent and static features.

SurvLatent ODE (Moon et al., 2022). SLODE is a recent deep learning framework for survival analysis that leverages an ODE-RNN architecture (Rubanova et al., 2019) to handle the time dependent features.

4.4. Synthetic Experiments

Hitting time of a partially observed SDE. Predicting hitting times is a crucial problem in finance — for instance, when pricing so-called catastrophe bounds triggering a payment to the holder in case of an event (Cheridito & Xu, 2015; Corcuera & Valdivia, 2016). Their relation to survival analysis is well documented, see e.g. (Lee & Whitmore, 2006). Building on this problem, we consider the Ornstein-Uhlenbeck SDE

$$dw^i(t) = -\omega(w^i(t) - \mu)dt + \sum_{j=1}^{d} dx^{(i,j)}(t) + \sigma dB^i(t)$$

where $d = 5$, $\sigma = 1$, $\mu = 0.1$ and $\omega = 0.1$ are fixed parameters. $x^i(t) = (x^{(i,1)}(t), \ldots, x^{(i,d-1)}(t))$ is a sample path of a fractional Brownian motion with Hurst parameter $H = 0.6$, and $B^i(t)$ is a Brownian noise term. In this setup, our data consists of $\mathbf{X}^i$ which is a downsampled version of x^i and the Brownian part is unobserved. Our goal is to predict the first hitting time $\min\{t > 0 \,|\, w_t \geq w_\star\}$

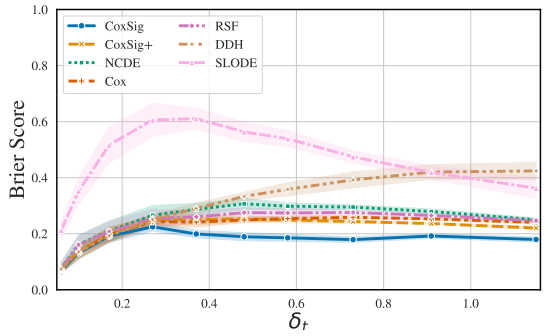

Figure 4: Brier score $\delta t \mapsto \mathrm{BS}(t, \delta t)$, evaluated at $t = 0.23$, for the partially observed SDE experiment. Confidence intervals indicate 1 standard deviation.

of a threshold value $w_\star = 2.5$. We train on $n = 500$ individuals. Figure 3 shows the sample paths and SDE of a randomly selected individual. This setup is close to a well-specified model since signatures linearize controlled differential equations.

Tumor Growth. We similarly aim at predicting the hitting time of a stochastic process modelling the growth of a tumor (Simeoni et al., 2004), where x^i represents a drug-intake. In this experiment, the time series $\mathbf{X}^i$ is very-low dimensional ($d = 2$, which includes the time channel).

4.5. Real-World Datasets

Predictive Maintenance. (Saxena et al., 2008) This dataset collects simulations of measurements of sensors placed on aircraft gas turbine engines run until a threshold value is reached. In this context, the time-to-event is the failure time. This dataset features a small sample size, considerable censoring rates and a high number of time channels.

Churn prediction. We use a private dataset provided by Califrais, a food supply chain company that delivers fresh products from Rungis to food professionals. The company has access to a variety of features observed through time for every customer. Its goal is for example to predict when the customer will churn. The time series in this setup are high dimensional but sampled at a low frequency.

Further details on all datasets are provided in Appendix D. Overall, our datasets are diverse in terms of sample size, size and length of the time series and censoring type.

4.6. Results

General performance of CoxSig. Overall, the signature-based estimators outperform competing methods. We observe that CoxSig and CoxSig+ improve over the strongest baselines in terms of Brier scores. Contrarily to the strong baseline DDH, this improvement is consistent over larger prediction windows $[t, t+\delta t]$ as δt increases (see Figure 4). They provide even stronger improvements in terms of C-indexes (see Figure 5 and Appendix D). This suggests that they are particularly well-tailored for ranking tasks, such as identifying the most-at-risk individual. Including the first observed value of the time series generally improves CoxSig's performance: this is possibly due to the fact that signatures are invariant by translation (i.e. the signature of $x : t \mapsto x(t)$ is equal to the signature of $x : t \mapsto x(t) + a$), and hence including the first value of the time series provides non-redundant information.

	Algorithms	Avg. C-Index ↑	IBS ↓
OU	CoxSig	**0.857±0.01**	**0.091±0.01**
	CoxSig+	0.857±0.01	0.095±0.01
	NCDE	0.517±0.04	0.103±0.01
	DDH	0.545±0.02	0.094±0.01
	SLODE	0.621±0.05	0.253±0.03
Tumor	CoxSig	0.696±0.02	0.138±0.01
	CoxSig+	0.797±0.03	0.137±0.01
	NCDE	0.827±0.02	**0.130±0.01**
	DDH	**0.941±0.05**	0.133±0.01
	SLODE	0.601±0.07	0.136±0.01
NASA	CoxSig	0.858±0.04	0.154±0.03
	CoxSig+	**0.867±0.04**	0.154±0.03
	NCDE	0.541±0.09	0.178±0.04
	DDH	0.813±0.06	0.156±0.02
	SLODE	0.438±0.14	**0.145±0.02**
Califrais	CoxSig	0.741±0.01	0.130±0.01
	CoxSig+	**0.751±0.01**	**0.129±0.01**
	NCDE	0.529±0.05	0.152±0.01
	DDH	0.570±0.03	0.139±0.01
	SLODE	0.542±0.03	0.193±0.03

Table 2: Averaged value of our metrics for 4 considered dataset over set of 10 different values of t chosen from the 5 to the 50th percentile of the distribution of event times. The values of δt for each dataset is chosen to be the same as that shown in Figure 5.

Performance on low-dimensional data. A notable exception is the tumor growth simulation, in which CoxSig is generally outperformed (see Figures 15 and 16 in the appendix). The competitive performance of signatures for moderate to high dimensional data streams and its below average performance on low dimensional data is a well-studied feature (see Fermanian (2021) for an empirical study). A possible solution to handle low-dimensional data is to use embeddings before computing signatures to make them more informative (Morrill et al., 2020).

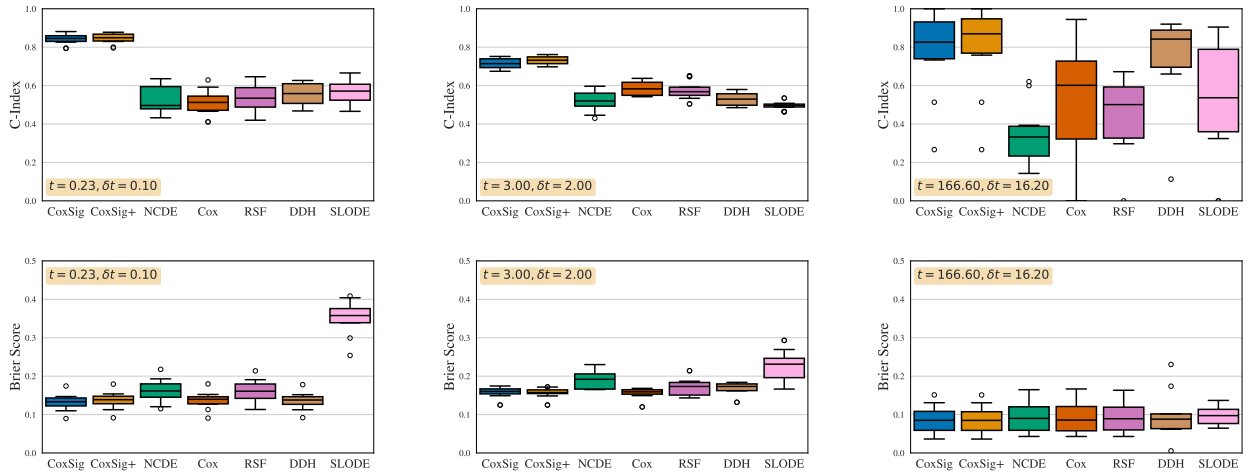

Figure 5: C-Index (*higher* is better) on **top** and Brier score (*lower* is better) on **bottom** for hitting time of a partially observed SDE (**left**), churn prediction (**center**) and predictive maintenance (**right**) evaluated at chosen points $(t, \delta t)$. t is chosen as the first decile of the event times i.e. 90% of the events occur after t. Hollow dots indicate outliers, and error bars indicate 80% of the interquartile range. We report detailed results for numerous points $(t, \delta t)$ in Appendix D.

NCDEs. On the other side, NCDEs generally tie or perform worse than competing methods. Notably, when considering C-indexes, they even perform worse than random on the predictive maintenance dataset. This stands in stark contrast to their good performances on classification or regression tasks (Kidger et al., 2020; Morrill et al., 2021; Vanderschueren et al., 2023).

Running times. Finally, we observe that our methods run in similar times than DDH, while including cross-validation (see Figure 13 in the appendix). Models that do not use time dependent features (RSF and Cox) are 2 orders of magnitude faster to train.

5. Conclusion

We have designed and analyzed a model for generic counting processes driven by a controlled latent state, which can be readily estimated using either NCDE or signature-based estimators. CoxSig in particular offers a parsimonious alternative to deep models and yields excellent performance for survival analysis. Future research efforts will be targeted at extending our model to competing risks and multimodal data.

Limitations. While our model shows competitive performance on moderate to high-dimensional data, one central limitation is its below average performance on low dimensional data. We also stress that the extension to very high dimensional time series is computationally prohibitive since the signature scales exponentially with the dimension of the time series. Finally, our experimental setup is limited to the survival analysis case: we plan on extending it to general counting processes in future work.

Acknowledgements

For this work, AG has benefited from the support of the National Agency for Research under the France 2030 program with the reference ANR-22-PESN-0016. LB and AG thank Killian Bakong-Epouné, who performed preliminary analysis of the NCDE model during an internship at Inria. LB, VTN and AF thank the Sorbonne Center for Artificial Intelligence (SCAI) and its team for their support. LB thanks Claire Boyer and Gérard Biau for supervising a previous internship that sparked his interest in signatures.

Impact statement

This paper presents work whose goal is to rank individuals or to predict their risk of event by relying on past information. This is a common goal in many diverse fields, including healthcare (which patient is most at risk of dying ?), insurance policy pricing (given a person's history, what is her probability of experiencing an event covered by her policy ?) and human resources management (which employee is most likely to quit ?).

Our proposed method can be used in any of these settings to personalize predictions, and hence target interventions. As any prediction algorithm, it may cause some disadvantaged groups to suffer from biased decisions.

References

Aalen, O., Borgan, O., and Gjessing, H. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

Bach, F. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384 – 414, 2010.

Bach, F. Learning theory from first principles. *Draft of a book, version of Sept*, 6:2021, 2021.

Bacry, E., Mastromatteo, I., and Muzy, J.-F. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

Bleistein, L. and Guilloux, A. On the generalization capacities of neural controlled differential equations. In *International Conference on Learning Representations*, 2024.

Bleistein, L., Fermanian, A., Jannot, A.-S., and Guilloux, A. Learning the dynamics of sparsely observed interacting systems. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 2603–2640. PMLR, 2023.

Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Bussy, S., Veil, R., Looten, V., Burgun, A., Gaïffas, S., Guilloux, A., Ranque, B., and Jannot, A.-S. Comparison of methods for early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework. *BMC medical research methodology*, 19:1–9, 2019.

Chen, K.-T. Integration of paths—a faithful representation of paths by non-commutative formal power series. *Transactions of the American Mathematical Society*, 89: 395–407, 1958.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 6572–6583. Curran Associates, Inc., 2018.

Chen, R. T. Q., Amos, B., and Nickel, M. Neural spatio-temporal point processes. In *International Conference on Learning Representations*, 2021.

Cheridito, P. and Xu, Z. Pricing and hedging cocos. *Available at SSRN 2201364*, 2015.

Cirone, N. M., Lemercier, M., and Salvi, C. Neural signature kernels as infinite-width-depth-limits of controlled resnets. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, pp. 25358–25425. PMLR, 2023.

Cirone, N. M., Orvieto, A., Walker, B., Salvi, C., and Lyons, T. Theoretical foundations of deep selective state-space models. *arXiv preprint arXiv:2402.19047*, 2024.

Corcuera, J. M. and Valdivia, A. Pricing cocos with a market trigger. In Benth, F. E. and Di Nunno, G. (eds.), *Stochastics of Environmental and Financial Economics*, pp. 179–209. Springer International Publishing, 2016.

Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Crowther, M. J., Abrams, K. R., and Lambert, P. C. Joint modeling of longitudinal and survival data. *The Stata Journal*, 13(1):165–184, 2013.

De Brouwer, E., Simm, J., Arany, A., and Moreau, Y. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 7379–7390. Curran Associates, Inc., 2019.

De Brouwer, E., Gonzalez, J., and Hyland, S. Predicting the impact of treatments over time with uncertainty aware neural differential equations. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 4705–4722. PMLR, 2022.

Dumanis, S. B., French, J. A., Bernard, C., Worrell, G. A., and Fureman, B. E. Seizure forecasting from idea to reality. outcomes of the my seizure gauge epilepsy innovation institute workshop. *Eneuro*, 4(6), 2017.

Fermanian, A. Embedding and learning with signatures. *Computational Statistics & Data Analysis*, 157:107148, 2021.

Fermanian, A. Functional linear regression with truncated signatures. *Journal of Multivariate Analysis*, 192: 105031, 2022.

Fermanian, A., Marion, P., Vert, J.-P., and Biau, G. Framing RNN as a kernel method: A neural ODE approach. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3121–3134. Curran Associates, Inc., 2021.

Friz, P. K. and Victoir, N. B. *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*, volume 120 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

Groha, S., Schmon, S. M., and Gusev, A. A general framework for survival analysis and multi-state modelling. *arXiv preprint arXiv:2006.04893*, 2020.

Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC medical research methodology*, 16:1–15, 2016.

Horvath, B., Lemercier, M., Liu, C., Lyons, T., and Salvi, C. Optimal stopping via distribution regression: a higher rank signature approach. *arXiv preprint arXiv:2304.01479*, 2023.

Ibrahim, J. G., Chu, H., and Chen, L. M. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796, 2010.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. Random survival forests. *The Annals of Applied Statistics*, 2(3):841 – 860, 2008.

Jia, J. and Benson, A. R. Neural jump stochastic differential equations. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, 2019.

Kidger, P. and Lyons, T. Signatory: differentiable computations of the signature and logsignature transforms, on both cpu and gpu. In *International Conference on Learning Representations*, 2020.

Kidger, P., Bonnier, P., Perez Arribas, I., Salvi, C., and Lyons, T. Deep signature transforms. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Kidger, P., Morrill, J., Foster, J., and Lyons, T. Neural controlled differential equations for irregular time series. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6696–6707. Curran Associates, Inc., 2020.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kvamme, H. and Borgan, Ø. The brier score under administrative censoring: Problems and a solution. *Journal of Machine Learning Research*, 24(2):1–26, 2023.

Kvamme, H., Ørnulf Borgan, and Scheel, I. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129): 1–30, 2019.

Lee, C., Yoon, J., and Van Der Schaar, M. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1): 122–133, 2019.

Lee, M.-L. T. and Whitmore, G. A. Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. 2006.

Lemler, S. Oracle inequalities for the Lasso in the high-dimensional Aalen multiplicative intensity model. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(2):981 – 1008, 2016.

Lin, P. and Yong, J. Controlled singular volterra integral equations and pontryagin maximum principle. *SIAM Journal on Control and Optimization*, 58(1):136–164, 2020.

Long, J. D. and Mills, J. A. Joint modeling of multivariate longitudinal data and survival data in several observational studies of huntington's disease. *BMC medical research methodology*, 18(1):1–15, 2018.

Lyons, T. and McLeod, A. D. Signature methods in machine learning. *arXiv preprint arXiv:2206.14674*, 2022.

Lyons, T., Caruana, M., and Lévy, T. *Differential Equations driven by Rough Paths*, volume 1908 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

Marion, P., Fermanian, A., Biau, G., and Vert, J.-P. Scaling resnets in the large-depth regime. *arXiv preprint arXiv:2206.06929*, 2022.

Mei, H. and Eisner, J. M. The neural hawkes process: A neurally self-modulating multivariate point process. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30, 2017.

Moon, I., Groha, S., and Gusev, A. Survlatent ode : A neural ode based time-to-event model with competing risks for longitudinal data improves cancer-associated venous thromboembolism (vte) prediction. In Lipton, Z., Ranganath, R., Sendak, M., Sjoding, M., and Yeung, S. (eds.), *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182, pp. 800–827. PMLR, 2022.

Morrill, J., Fermanian, A., Kidger, P., and Lyons, T. A generalised signature method for multivariate time series feature extraction. *arXiv preprint arXiv:2006.00873*, 2020.

Morrill, J., Kidger, P., Yang, L., and Lyons, T. Neural controlled differential equations for online prediction tasks. *arXiv preprint arXiv:2106.11028*, 2021.

Murray, J. and Philipson, P. A fast approximate em algorithm for joint models of survival and multivariate longitudinal data. *Computational Statistics & Data Analysis*, 170:107438, 2022.

Nardi, Y. and Rinaldo, A. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 0, 01 2008. doi: 10.1214/08-EJS200.

Ogata, Y. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, pp. 9–27, 1988.

Omi, T., ueda, n., and Aihara, K. Fully neural network based model for general temporal point processes. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, 2019.

Pölsterl, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.

Proust-Lima, C., Séne, M., Taylor, J. M., and Jacqmin-Gadda, H. Joint latent class models for longitudinal and time-to-event data: a review. *Statistical methods in medical research*, 23(1):74–90, 2014.

Rasheed, K., Qayyum, A., Qadir, J., Sivathamboo, S., Kwan, P., Kuhlmann, L., O'Brien, T., and Razi, A. Machine learning for predicting epileptic seizures using eeg signals: A review. *IEEE Reviews in Biomedical Engineering*, 14:139–155, 2020.

Reizenstein, J. F. and Graham, B. Algorithm 1004: The iisignature library: Efficient calculation of iterated-integral signatures and log signatures. *ACM Transactions on Mathematical Software*, 46(1):1–21, 2020.

Rizopoulos, D. The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software*, 72(7):1–46, 2016.

Rubanova, Y., Chen, R. T. Q., and Duvenaud, D. K. Latent ordinary differential equations for irregularly-sampled time series. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 5320–5330. Curran Associates, Inc., 2019.

Rustand, D., Van Niekerk, J., Krainski, E. T., Rue, H., and Proust-Lima, C. Fast and flexible inference for joint models of multivariate longitudinal and survival data using integrated nested laplace approximations. *Biostatistics*, 25(2):429–448, 2024.

Salvi, C., Lemercier, M., Liu, C., Horvath, B., Damoulas, T., and Lyons, T. Higher order kernel mean embeddings to capture filtrations of stochastic processes. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 16635–16647, 2021.

Saxena, A., Goebel, K., Simon, D., and Eklund, N. Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 International Conference on Prognostics and Health Management*, pp. 1–9, 2008.

Shchur, O., Türkmen, A. C., Januschowski, T., and Günnemann, S. Neural temporal point processes: A review. In Zhou, Z.-H. (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4585–4593, 2021.

Shorack, G. R. and Wellner, J. A. *Empirical processes with applications to statistics*. Society for Industrial and Applied Mathematics, 2009.

Simeoni, M., Magni, P., Cammia, C., De Nicolao, G., Croci, V., Pesenti, E., Germani, M., Poggesi, I., and Rocchetti, M. Predictive pharmacokinetic-pharmacodynamic modeling of tumor growth kinetics in xenograft models after administration of anticancer agents. *Cancer research*, 64(3):1094–1101, 2004.

Tang, W., Ma, J., Mei, Q., and Zhu, J. Soden: A scalable continuous-time survival model through ordinary differential equation networks. *Journal of Machine Learning Research*, 23(34):1–29, 2022.

Therneau, T. M. and Grambsch, P. M. *Modeling survival data: extending the Cox model*. Springer, New-York, 2000.

Van de Geer, S. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics*, pp. 1779–1801, 1995.

Vanderschueren, T., Curth, A., Verbeke, W., and Van Der Schaar, M. Accounting for informative sampling when learning to forecast treatment outcomes over time. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 34855–34874. PMLR, 2023.

Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., and Groothuis-Oudshoorn, C. G. Time-varying covariates and coefficients in cox regression models. *Annals of Translational Medicine*, 6(7), 2018.

Supplementary Material

A. Supplementary Mathematical Elements

A.1. Supplementary elements on survival analysis

The counting process associated with the observation of $T_1^i < T_2^i < \dots$ is denoted by $\tilde{N}^i$. The observed counting process is $t \to N^i(t) = \int_0^t Y^i(s) d\tilde{N}^i(s)$. The integral against the counting process N^i is to be understood is to be understood as a Stieltjes integral, i.e., $\int_0^t \lambda_\star^i(s) dN^i(s) = \sum_{T_i \leq t} \lambda_\star^i(T_i)$ — see Aalen et al. (2008, p.55-56). Its intensity writes $\lambda_\star^i(t \,|\, \mathbf{W}^i, (x^i(s))_{s \leq t}) Y^i(t)$, which we simply write $\lambda_\star^i(t) Y^i(t)$ to alleviate notations.

To the observations, we associate the filtration $\mathcal{F}$, with all σ-fields at $0 \leq t \leq \tau$ defined as

$$\mathcal{F}_t = \bigcup_{i=1,\dots,n} \mathcal{F}_t^i$$

where $\mathcal{F}_t^i = \sigma\big(x^i(s), \mathbf{W}^i, N^i(s), Y^i(s), 0 \leq s \leq t\big)$. We assume in addition that Y^i is $\mathcal{F}^i$-predictable.

Using the Doob-Meyer decomposition of counting processes - see Aalen et al. (2008, p. 52-60) - we have

$$N^i(t) = \int_0^t \lambda_\star^i(s) Y^i(s) ds + M^i(t) \tag{4}$$

where M^i is local square integrable martingale with respect to $\mathcal{F}^i$.

A.2. Picard-Lindelhöf Theorem

Theorem A.1. *Let $x : [0, \tau] \to \mathbb{R}^d$ be a continuous path of bounded variation, and assume that $\mathbf{G} : \mathbb{R}^p \to \mathbb{R}^{p \times d}$ is Lipschitz continuous. Then the CDE*

$$dz(t) = \mathbf{G}(z(t)) dx(t)$$

with initial condition $z_0 \in \mathbb{R}^p$ has a unique solution on $[0, \tau]$.

A full proof can be found in Fermanian et al. (2021, Theorem 4). Remark that since in our setting, NCDEs are Lipschitz since typical neural vector fields, such as feed-forward neural networks, are Lipschitz (Virmaux & Scaman, 2018). This ensures that the solutions to NCDEs are well defined.

A.3. Continuity of the Flow of CDEs

We state a result on the continuity of the flow adapted from Bleistein & Guilloux (2024), Theorem B.5.

Theorem A.2. *Let $\mathbf{F}, \mathbf{G} : \mathbb{R}^p \to \mathbb{R}^{p \times d}$ be two Lipschitz vector fields with Lipschitz constants $L_\mathbf{F}, L_\mathbf{G} > 0$. Let $x, r : [0, \tau] \to \mathbb{R}^d$ be either continuous or piecewise constant paths of total variations bounded by L_x and L_r. Consider the controlled differential equations*

$$dw(t) = \mathbf{F}(w(t)) dx(t) \;\; \textit{and} \;\; dv(t) = \mathbf{G}(v(t)) dr(t)$$

with initial conditions $w(0) = v(0) = 0$ respectively. It then follows that for any $t \in [0, \tau]$

$$\|w(t) - v(t)\| \leq \left(\|x - r\|_{\infty,[0,t]} \big(1 + L_\mathbf{F} L_r \mathcal{K}\big) + \max_{v \in \Omega} \|\mathbf{F}(v) - \mathbf{G}(v)\|_{\text{op}} L_r \right) \exp(L_\mathbf{F} L_x),$$

where

$$\mathcal{K} = \left[L_\mathbf{F} \big(\|\mathbf{F}(0)\|_{\text{op}} L_x \big) \exp(L_\mathbf{F} L_x) + \|\mathbf{F}(0)\|_{\text{op}} \right] \exp(L_\mathbf{F} L_x)$$

and

$$\Omega = \big\{ u \in \mathbb{R}^p \mid \|u\| \leq (\|\mathbf{G}(0)\|_{\text{op}} L_r) \exp(L_\mathbf{G} L_r) \big\}.$$

A.4. Linearization in the Signature Space

A.4.1. General Result

In this section, we give additional details on the linearization of CDEs in the signature space. We first define the differential product.

Definition A.3. Let $F, G : \mathbb{R}^p \to \mathbb{R}^p$ be two $\mathcal{C}^\infty$ vector fields and let $J(\cdot)$ be the Jacobian matrix. Their differential product $F \star G : \mathbb{R}^p \to \mathbb{R}^p$ is the smooth vector field defined for every $h \in \mathbb{R}^p$ by

$$(F \star G)(h) = \sum_{j=1}^{e} \frac{\partial G}{\partial h_j}(h) F_j(h) = J(G)(h)F(h).$$

We now consider a tensor field $\mathbf{F} : \mathbb{R}^p \to \mathbb{R}^{p \times d}$ which we write

$$\mathbf{F} = \begin{bmatrix} | & \dots & | \\ F^1 & \dots & F^d \\ | & \dots & | \end{bmatrix},$$

where for every $1 \leq i \leq d$, $F^i : \mathbb{R}^p \to \mathbb{R}^p$, and we define

$$\Gamma_k(\mathbf{F}) := \sup_{\|h\| \leq M,\, i_1 \leq \dots \leq i_k \leq d} \left\| F^{i_1} \star \dots \star F^{i_k}(h) \right\|_2.$$

Consider the solution $z : [0, \tau] \to \mathbb{R}^p$ to the CDE

$$\begin{aligned} dz(t) &= \mathbf{F}(z(t))dx(t) \\ z(0) &= \mathbf{0} \in \mathbb{R}^p \end{aligned} \tag{5}$$

where $x : [0, \tau] \to \mathbb{R}^d$ is a continuous path of finite total variation bounded by $L_x \tau > 0$. We recall the following result from Fermanian et al. (2021), Proposition 4.

Proposition A.4 (Fermanian et al. (2021), Proposition 4.)**.** *We have*

$$\left\| z_\star^i(t) - \alpha_{\star,N}^\top \mathbf{S}_N(x_{[0,t]}) \right\| \leq \frac{(dL_x t)^{N+1}}{(N+1)!} \Gamma_{N+1}(\mathbf{F})$$

As a consequence, we have the following theorem.

Theorem A.5. *Let $\mathbf{F} : \mathbb{R}^p \to \mathbb{R}^{p \times d}$ be a $\mathcal{C}^\infty$ tensor field. If*

$$\frac{(dL_x t)^{N+1}}{(N+1)!} \Gamma_{N+1}(\mathbf{F}) \to 0$$

as $N \to +\infty$, then the solution z to the CDE (5) can be written as

$$z(t) = \sum_{k \geq 1} \sum_{I \in \{1, \dots, d\}^k} \mathbf{S}^I(x_{[0,t]}) F^{i_1} \star \dots \star F^{i_k}(\mathbf{0}).$$

A.4.2. Application to our Model

Recall that we have defined our generative model through the CDE

$$dz_\star^i(t) = \mathbf{G}_\star(z_\star^i(t))dx^i(t)$$

with initial condition $z_\star^i(0) = 0$, where $\mathbf{G}_\star : \mathbb{R} \to \mathbb{R}^p$ is a $L_{\mathbf{G}_\star}$-Lipschitz vector field. Since in our case, the vector field $\mathbf{G}_\star$ maps $\mathbb{R}$ to $\mathbb{R}^d$, it can be written as

$$\mathbf{G}_\star(z) = \begin{bmatrix} G_\star^1(z) & \dots & G_\star^d(z), \end{bmatrix},$$

where for every $1 \leq i \leq d$, $G_\star^i : \mathbb{R} \to \mathbb{R}$. In this setup, for $1 \leq i_1, i_2 \leq d$ the differential product collapses to

$$(G_\star^{i_1} \star G_\star^{i_2})(h) = (G_\star^{i_2})'(h) \times G_\star^{i_1}(h) \in \mathbb{R}.$$

For $1 \leq i_1, i_2, i_3 \leq d$, it writes

$$\begin{aligned}(G_\star^{i_1} \star G_\star^{i_2} \star G_\star^{i_3})(h) &= (G_\star^{i_2} \star G_\star^{i_3})'(h) \times G_\star^{i_1}(h) \\ &= \big((G_\star^{i_3})'(h) \times G_\star^{i_2}(h)\big)' \times G_\star^{i_1}(h) \\ &= \big((G_\star^{i_3})^{(2)}(h) \times G_\star^{i_2}(h) + (G_\star^{i_3})'(h) \times (G_\star^{i_2})'(h)\big) \times G_\star^{i_1}(h) \in \mathbb{R}.\end{aligned}$$

One can derive similar expression for $1 \leq i_1, \ldots, i_k \leq d$. In line with Theorem A.5, we make the following Assumption on the vector field $\mathbf{G}_\star$.

Assumption 4. The vector field $\mathbf{G}_\star$ satisfies

$$\frac{(L_x \tau d)^{N+1}}{(N+1)!} \Gamma_{N+1}(\mathbf{G}_\star) \to 0$$

as $N \to \infty$.

We can write the ℓ_2 and ℓ_1 norms of $\alpha_{\star,N}$ as functions of the differential product of $\mathbf{G}_\star$.

Lemma A.6. *We have that*

$$\|\alpha_{\star,N}\|_2 \leq \Big(\sum_{k=1}^N d^k \Gamma_k(\mathbf{G}_\star)^2\Big)^{1/2}$$

and

$$\|\alpha_{\star,N}\|_1 \leq \sum_{k=1}^N d^k \Gamma_k(\mathbf{G}_\star).$$

Proof. Starting with the ℓ_2 norm, one has

$$\begin{aligned}\|\alpha_{\star,N}\|_2 &= \Big(\sum_{k=1}^N \sum_{1 \leq i_1, i_2, \ldots, i_k \leq d} G_\star^{i_1} \star \cdots \star G_\star^{i_k}(\mathbf{0})^2\Big)^{1/2} \\ &\leq \Big(\sum_{k=1}^N d^k \max_{1 \leq i_1, i_2, \ldots, i_k \leq d} |G_\star^{i_1} \star \cdots \star G_\star^{i_k}(\mathbf{0})|^2\Big)^{1/2} \\ &\leq \Big(\sum_{k=1}^N d^k \Gamma_k(\mathbf{G}_\star)^2\Big)^{1/2}.\end{aligned}$$

Moving on to the ℓ_1 norm, we similarly obtain

$$\begin{aligned}\|\alpha_{\star,N}\|_1 &= \Big(\sum_{k=1}^N \sum_{1 \leq i_1, i_2, \ldots, i_k \leq d} |G_\star^{i_1} \star \cdots \star G_\star^{i_k}(\mathbf{0})|\Big) \\ &\leq \Big(\sum_{k=1}^N d^k \max_{1 \leq i_1, i_2, \ldots, i_k \leq d} |G_\star^{i_1} \star \cdots \star G_\star^{i_k}(\mathbf{0})|\Big) \\ &\leq \sum_{k=1}^N d^k \Gamma_k(\mathbf{G}_\star).\end{aligned}$$

□

A.5. Signature of a Discretized Path

We recall the following result from Bleistein et al. (2023).

Theorem A.7. *Let $x : [0,\tau] \to \mathbb{R}^d$ be a path satisfying Assumption 1. Let $D = \{t_1, \dots, t_K\} \subset [0,\tau]$ be a grid of sampling points, and x^D the piecewise constant interpolation of the path x sampled on the grid D. For all $\alpha \in \mathbb{R}^q$, where $q := \frac{d^N-1}{d-1}$, we have*

$$|\alpha^\top \big(\mathbf{S}_N(x_{[0,t]}) - \mathbf{S}_N(x^D_{[0,t]})\big)| \le c_3(N)\, \|\alpha\|\, |D|,$$

where

$$c_3(N) = 2e\frac{(L_x t)^{N-1} - 1}{L_x t - 1} L_x.$$

A.6. The Cox Connection

Signature-based embeddings. Consider a continuous path of bounded variation $x : t \mapsto (x(t), t) \in \mathbb{R}^d$. First, remark that for every word of size one $I \in \{1, \dots, d\}$, the signature writes

$$\mathbf{S}^I(x_{[0,t]}) = \int_{0<u_1<t} dx^{(I)}(s) = x^{(I)}(t).$$

Furthermore, for any word $I = (d, \dots, d)$ of size k made only of the letter d, i.e., words that only include the time channel, we have

$$\mathbf{S}^I(x_{[0,t]}) = \int_{0<u_1<\dots<u_k<t} du_1 \dots du_k = \frac{1}{k!} t^k.$$

This shows that for $x = x^{i,D}$

$$\alpha^\top \mathbf{S}_N(x^{i,D}_{[0,t]}) = \alpha^\top_{I_1}(1, t, t^2, \dots, t^N) + \alpha^\top_{I_2}\mathbf{X}^i(t) + \sum_{I \in I_3} \alpha_I \mathbf{S}^I(x^{i,D}_{[0,t]})$$

where α_{I_1} is the subvector of α collecting all coefficients associated to the words $\{d\}, \{d,d\}, \dots, \{d, \dots, d\}$ containing only the letter d, α_{I_2} is the subvector collecting all coefficients associated to the $d-1$ words $\{1\}, \{2\}, \dots, \{d-1\}$ of size 1, and α_{I_3} collects the remaining coefficients.

NCDEs. For any $N \ge 1$, consider the augmented vector field

$$\tilde{\mathbf{G}}_\psi(z) = \begin{bmatrix} \mathbf{G}_\psi(z) & \mathbf{0}_{p\times(N-1)} \\ \mathbf{0}_{(N-1)\times d} & \mathbf{I}_{(N-1)\times(N-1)} \end{bmatrix} \in \mathbb{R}^{(N-1+p)\times(N-1+d)}, \quad z \in \mathbb{R}^p,$$

and an embedding of the time series $\mathbf{X}^i$ of the form $\tilde{x}^{i,D}(s) = (\mathbf{X}^i(t_k), s, s^2, \dots, s^N) \in \mathbb{R}^{d+N-1}$ for $s \in [t_k, t_{k+1}[$. The latent state of the NCDE model is now updated as

$$\begin{aligned} \tilde{z}^{i,D}_\theta(t_{k+1}) &= \tilde{z}^{i,D}_\theta(t_k) + \tilde{\mathbf{G}}_\psi(\tilde{z}^{i,D}_\theta(t_k))\Delta\tilde{\mathbf{X}}^i(t_{k+1}) \\ &= \tilde{z}^{i,D}_\theta(t_k) + \begin{bmatrix} \mathbf{G}_\psi(z^{i,D}_\theta)\Delta\mathbf{X}^i(t_{k+1}) \\ \Delta t_{k+1} \\ \vdots \\ \Delta t^N_{k+1} \end{bmatrix} = \begin{bmatrix} z^{i,D}_\theta(t_k) + \mathbf{G}_\psi(z^{i,D}_\theta)\Delta\mathbf{X}^i(t_{k+1}) \\ t_{k+1} \\ \vdots \\ t^N_{k+1} \end{bmatrix}. \end{aligned}$$

This proves that in the NCDE based model, the intensity can similarly be written as

$$\alpha^\top \tilde{z}^{i,D}_\theta(t) = \alpha^\top_{I_1} z^{i,D}_\theta(t) + \alpha^\top_{I_2}(1, t, t^2, \dots, t^N)$$

where $\alpha = (\alpha_{I_1}, \alpha_{I_2})$, and $\alpha_1 \in \mathbb{R}^p$ and $\alpha_2 \in \mathbb{R}^N$.

A.7. Self-concordance

We now state a self-concordance bound, which can be found along with its proof in Bach (2010).

Lemma A.8. *Let $g : \mathbb{R} \to \mathbb{R}$ be a convex, three times differentiable function such that*

$$|g^{(3)}(x)| \leq M g^{(2)}(x)$$

for all $x \in \mathbb{R}$ and for some $M \geq 0$. Then it follows that

$$\frac{g^{(2)}(0)}{M^2}\Phi(-Mt) \leq g(t) - g(0) - tg'(0) \leq \frac{g^{(2)}(0)}{M^2}\Phi(Mt)$$

for all $t \geq 0$, where

$$\Phi : t \mapsto \exp(t) - t - 1.$$

A.8. Decomposition of the difference in likelihoods

We first define the empirical KL-divergence between the true and parameterized intensity associated to the sample $\mathcal{D}_n$ as

$$\mathrm{KL}_n(\lambda_\star, \lambda_\theta^D) := \frac{1}{n}\sum_{i=1}^n \int_0^\tau \log \frac{\lambda_\star^i(s)}{\lambda_\theta^i(s)} \lambda_\star^i(s) Y^i(s) ds - \frac{1}{n}\sum_{i=1}^n \int_0^\tau \big(\lambda_\star^i(s) - \lambda_\theta^i(s)\big) Y^i(s) ds.$$

This definition is classical for intensities of counting processes (Aalen et al., 2008; Lemler, 2016). We now show that minimizing the empirical KL-divergence between the true and the parameterized intensity amounts to minimizing the empirical log likelihood, ignoring a noise term that will be canceled by setting the penalty accordingly.

Proposition A.9. *For every $\theta \in \Theta$, the difference in likelihoods $\ell_n^D(\theta) - \ell_n^\star$ decomposes as*

$$\mathrm{KL}_n(\lambda_\star, \lambda_\theta^D) - \frac{1}{n}\sum_{i=1}^n \int \log \frac{\lambda_\theta^{i,D}(s)}{\lambda_\star^i(s)} dM^i(s),$$

where $M^i : [0, \tau] \to \mathbb{R}$ is a local square integrable martingale.

This proposition is a consequence of the Doob-Meyer decomposition $N^i(t) = \int_0^t \lambda_\star^i(s) Y^i(s) ds + M^i(t)$ of the counting process (Aalen et al., 2008). We now furthermore define the total variation divergence as

$$\mathrm{TV}_n(\lambda_\star, \lambda_\theta^D) := \frac{1}{n}\sum_{i=1}^n \int_0^\tau |\lambda_\star^i(s) - \lambda_\theta^{i,D}(s)| Y^i(s) ds$$

and the quadratic log divergence $\mathrm{D}_n^2(\lambda_\star, \lambda_\theta^D)$ as

$$\frac{1}{n}\sum_{i=1}^n \int_0^\tau \big(\log \lambda_\theta^{i,D}(s) - \log \lambda_\star^i(s)\big)^2 \lambda_\star^i(s) Y^i(s) ds.$$

Proposition A.10. *There exist two constants $c_1, c_2 > 0$ such that*

$$c_1 \mathrm{TV}_n(\lambda_\star, \lambda_\theta^D)^2 \leq \mathrm{KL}_n(\lambda_\star, \lambda_\theta^D) \leq c_2 \mathrm{D}_n^2(\lambda_\star, \lambda_\theta^D).$$

More precisely, the constants c_1, c_2 are functions of Θ, L_x, τ and $L_{\mathbf{G}_\star}$ and are given explicitly in Appendix B.2.

This bound is obtained by combining a Pinsker-type inequality (Proposition B.1) and a self-concordance bound (Proposition B.2). It is informative in two ways. First, it shows that minimizing the negative empirical log-likelihood and hence the KL-divergence between the true and parameterized intensity will lead to a minimization of the total variation between the two intensities. Secondly, it shows that the KL-divergence is upper bounded by a term involving the difference of the logarithms of the intensities. We make use of this second bound to obtain a bias-variance decomposition.

B. Proofs

B.1. Proof of Proposition A.9

Proof. Thanks to the Doob-Meyer decomposition of Equation (4), the log-likelihood associated to individual i is

$$\begin{aligned}\ell_i^D(\theta) &= \int_0^\tau \lambda_\theta^{i,D}(s)Y^i(s)ds - \int_0^\tau \log \lambda_\theta^{i,D}(s)dN^i(s) \\ &= \int_0^\tau \Big(\lambda_\theta^{i,D}(s) - \log \lambda_\theta^{i,D}(s)\lambda_\star^i(s)\Big)Y^i(s)ds - \int_0^\tau \log \lambda_\theta^{i,D}(s)dM^i(s).\end{aligned}$$

Similarly, the log-likelihood associated to the true intensity $\lambda_\star^i$ writes

$$\begin{aligned}\ell_n^\star &= \int_0^\tau \lambda_\star^i(s)Y^i(s)ds - \int_0^\tau \log \lambda_\star^i(s)dN^i(s) \\ &= \int_0^\tau \Big(\lambda_\star^i(s) - \log \lambda_\star^i(s)\lambda_\star^i(s)\Big)Y^i(s)ds - \int_0^\tau \log \lambda_\star^i(s)dM^i(s).\end{aligned}$$

Hence, we get

$$\begin{aligned}&\ell_n^D(\theta) - \ell_n^\star \\ &= \frac{1}{n}\sum_{i=1}^n \int \Big[\lambda_\theta^{i,D}(s) - \lambda_\star^i(s) - \lambda_\star^i(s)\log\frac{\lambda_\theta^{i,D}(s)}{\lambda_\star^i(s)}\Big]Y^i(s)ds - \frac{1}{n}\sum_{i=1}^n \int \log\frac{\lambda_\theta^{i,D}(s)}{\lambda_\star^i(s)}dM^i(s) \\ &= \mathrm{KL}_n(\lambda_\star, \lambda_\theta^D) - \frac{1}{n}\sum_{i=1}^n \int \log\frac{\lambda_\theta^{i,D}(s)}{\lambda_\star^i(s)}dM^i(s).\end{aligned}$$

This concludes the proof.

□

B.2. Proof of Proposition A.10

To prove Proposition A.10, we essentially combine a Pinsker-type inequality and a self-concordance bound. We prove these two bounds separatly bellow. Combining them yields the double inequality of Proposition A.10. In the following, for all $t \in [0, \tau]$, we let

$$\Lambda_\theta^{i,D}(t) = \int_0^t \lambda_\theta^{i,D}(s)Y^i(s)ds$$

and

$$\Lambda_\star^i(t) = \int_0^t \lambda_\star^i(s)Y^i(s)ds$$

be the cumulative hazard functions.

Proposition B.1 (Pinsker's inequality.)**.** *Let $\lambda_\star$ be the true intensity defined in Equations (1) and (2), and λ_θ be an intensity parameterized by $\theta \in \Theta$. Under Assumptions 1, 2, and 3, we have that*

$$c_1 \mathrm{TV}_n(\lambda_\star, \lambda_\theta^D)^2 \leq \mathrm{KL}_n(\lambda_\star, \lambda_\theta^D),$$

with

$$c_1 := \frac{1}{\tau}\exp(-B_{\beta,2}B_{\mathbf{W}})\Big[\frac{4}{3}\exp\Big(\|\mathbf{G}_\star(0)\|_{\mathrm{op}} L_x\tau \exp\big(L_{\mathbf{G}_\star}L_x\tau\big)\Big) + \frac{2}{3}\exp\big(B_\alpha \exp(L_x\tau)\big)\Big]^{-1}$$

for signature-based embeddings and

$$c_1 := \frac{1}{\tau}\exp(-B_{\beta,2}B_{\mathbf{W}})\Big[\frac{4}{3}\exp\Big(\|\mathbf{G}_\star(0)\|_{\mathrm{op}} L_x\tau \exp\big(L_{\mathbf{G}_\star}L_x\tau\big)\Big) + \frac{2}{3}\exp\Big(\|\mathbf{G}_\psi(0)\|_{\mathrm{op}} L_x\tau \exp\big(L_{\mathbf{G}_\psi}L_x\tau\big)\Big)\Big]^{-1}$$

for NCDEs.

Proof. We have

$$\mathrm{TV}_n(\lambda_\star, \lambda_\theta^D) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau |\lambda_\star^i(s) - \lambda_\theta^{i,D}(s)| Y^i(s)ds = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left|\frac{\lambda_\star^i(s)}{\lambda_\theta^{i,D}(s)} - 1\right| \lambda_\theta^{i,D}(s) Y^i(s)ds$$
$$= \int_0^\tau \sqrt{\left(\frac{\lambda_\star^i(s)}{\lambda_\theta^{i,D}(s)} - 1\right)^2} \lambda_\theta^{i,D}(s) Y^i(s)ds,$$

where we have used that $|x| = \sqrt{x^2}$. Note that by definition, $\lambda_\theta^{i,D}(s) = \exp\left(z_\theta^{i,D}(s) + \beta^\top \mathbf{W}^i\right) > 0$ for all $s \in [0, \tau]$: we can thus safely divide by this term. Now, since for all $x \geq 0$

$$(x-1)^2 \leq \Big(\frac{4}{3} + \frac{2}{3}x\Big)\xi(x)$$

with

$$\xi(x) := x \log x - x + 1,$$

one obtains

$$\mathrm{TV}_n(\lambda_\star, \lambda_\theta^D) \leq \frac{1}{n}\sum_{i=1}^n \int_0^\tau \sqrt{\Big(\frac{4}{3} + \frac{2}{3}\frac{\lambda_\star^i(s)}{\lambda_\theta^{i,D}(s)}\Big)\xi\Big(\frac{\lambda_\star^i(s)}{\lambda_\theta^{i,D}(s)}\Big)} \lambda_\theta^{i,D}(s) Y^i(s)ds.$$

Using the Cauchy-Schwarz inequality yields

$$\begin{aligned}
\mathrm{TV}_n(\lambda_\star, \lambda_\theta) &\leq \frac{1}{n}\sum_{i=1}^n \left[\int_0^\tau \Big(\frac{4}{3} + \frac{2}{3}\frac{\lambda_\star^i(s)}{\lambda_\theta^{i,D}(s)}\Big)\lambda_\theta^{i,D}(s) Y^i(s)ds\right]^{1/2} \left[\int_0^\tau \xi\Big(\frac{\lambda_\star^i(s)}{\lambda_\theta^{i,D}(s)}\Big)\lambda_\theta^{i,D}(s) Y^i(s)ds\right]^{1/2} \\
&\leq \frac{1}{n}\sum_{i=1}^n \left[\frac{4}{3}\Lambda_\theta^{i,D}(\tau) + \frac{2}{3}\Lambda_\star^i(\tau)\right]^{1/2} \left[\int_0^\tau \Big(\lambda_\star^i(s)\log\frac{\lambda_\star^i(s)}{\lambda_\theta^{i,D}(s)} + \lambda_\theta^{i,D}(s) - \lambda_\star^i(s)\Big) Y^i(s)ds\right]^{1/2} \\
&\leq \max_{i=1,\dots,n}\left[\frac{4}{3}\Lambda_\theta^{i,D}(\tau) + \frac{2}{3}\Lambda_\star^i(\tau)\right]^{1/2} \sqrt{\mathrm{KL}_n(\lambda_\star, \lambda_\theta^D)}.
\end{aligned}$$

Taking the square on both sides yields

$$\mathrm{TV}_n(\lambda_\star, \lambda_\theta^D)^2 \leq \max_{i=1,\dots,n}\left[\frac{4}{3}\Lambda_\theta^{i,D}(\tau) + \frac{2}{3}\Lambda_\star^i(\tau)\right] \mathrm{KL}_n(\lambda_\star, \lambda_\theta^D).$$

Bounding the true cumulative hazard function. We have

$$\max_{i=1,\dots,n}\left[\frac{4}{3}\Lambda_\theta^{i,D}(\tau) + \frac{2}{3}\Lambda_\star^i(\tau)\right] \leq \frac{4}{3}\max_{i=1,\dots,n}\Lambda_\theta^{i,D}(\tau) + \frac{2}{3}\max_{i=1,\dots,n}\Lambda_\star^i(\tau).$$

Moreover, Lemma 2.1 yeilds that, for all $s \in [0, \tau]$,

$$\begin{aligned}
\log \lambda_\star^i(s) &\leq B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\mathrm{op}} L_x s \exp\left(L_{\mathbf{G}_\star} L_x s\right) \\
&\leq B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\mathrm{op}} L_x \tau \exp\left(L_{\mathbf{G}_\star} L_x \tau\right).
\end{aligned}$$

Hence for all $i = 1, \dots, n$, it holds that

$$\Lambda_\star^i(\tau) = \int_0^\tau \lambda_\star^i(s) Y^i(s)ds \leq \tau \sup_{s\in[0,\tau]} \lambda_\star^i(s) \leq \tau \exp\Big(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\mathrm{op}} L_x \tau \exp\left(L_{\mathbf{G}_\star} L_x \tau\right)\Big).$$

Since this last bound does not depend on i, this gives us that

$$\frac{2}{3}\max_{i=1,\dots,n}\Lambda_\star^i(\tau)\leq\frac{2\tau}{3}\exp\Big(B_{\beta,2}B_{\mathbf{W}}+\|\mathbf{G}_\star(0)\|_{\text{op}}L_x\tau\exp\big(L_{\mathbf{G}_\star}L_x\tau\big)\Big).$$

Similarly, one obtains that

$$\frac{4}{3}\max_{i=1,\dots,n}\Lambda_\theta^{i,D}(\tau)\leq\frac{4\tau}{3}\max_{i=1,\dots,n}\sup_{s\in[0,\tau]}\lambda_\theta^i(s).$$

Signature-based embeddings. For signature-based embeddings, we have

$$\sup_{s\in[0,\tau]}\lambda_\theta^i(s)\leq\exp(B_{\beta,2}B_{\mathbf{W}})\Big[\exp(B_\alpha\exp(L_x\tau))\Big].$$

NCDEs. For NCDEs, one obtains

$$\sup_{s\in[0,\tau]}\lambda_\theta^i(s)\leq\exp\big(B_{\beta,2}B_{\mathbf{W}}\big)\exp\Big[\|\mathbf{G}_\psi(0)\|_{\text{op}}L_x\tau\exp\big(L_{\mathbf{G}_\psi}L_x\tau\big)\Big].$$

Final Bound. Putting everything together, one finally has that

$$c_1\mathrm{TV}_n(\lambda_\star,\lambda_\theta^D)^2\leq\mathrm{KL}_n(\lambda_\star,\lambda_\theta),$$

where

$$c_1=\frac{1}{\tau}\exp(-B_{\beta,2}B_{\mathbf{W}})\Big[\frac{4}{3}\exp(\|\mathbf{G}_\star(0)\|_{\text{op}}L_x\tau\exp\big(L_{\mathbf{G}_\star}L_x\tau\big))+\frac{2}{3}\exp(B_\alpha\exp(L_x\tau))\Big]^{-1}$$

when using signatures and

$$c_1:=\frac{1}{\tau}\exp(-B_{\beta,2}B_{\mathbf{W}})\Big[\frac{4}{3}\exp(\|\mathbf{G}_\star(0)\|_{\text{op}}L_x\tau\exp\big(L_{\mathbf{G}_\star}L_x\tau\big))+\frac{2}{3}\exp\Big[\|\mathbf{G}_\psi(0)\|_{\text{op}}L_x\tau\exp\big(L_{\mathbf{G}_\psi}L_x\tau\big)\Big]\Big]^{-1}$$

when using NCDEs.

□

We also prove the following self-concordance bound, a close result can be found in Lemler (2016).

Proposition B.2 (A self-concordance bound.)**.** *Let $\lambda_\star$ be the true intensity defined in Equations (1) and (2), and λ_θ be a intensity parameterized by $\theta\in\Theta$. Under Assumptions 1, 2, and 3, it holds that*

$$\mathrm{KL}_n(\lambda_\star,\lambda_\theta)\leq c_2\mathrm{D}_n^2(\lambda_\star,\lambda_\theta^D),$$

where

$$c_2:=\frac{\exp M-M-1}{M},$$

and

$$M:=\exp(B_{\beta,2}B_{\mathbf{W}})\Big[\exp\Big(\|\mathbf{G}_\star(0)\|_{\text{op}}L_x\tau\exp\big(L_{\mathbf{G}_\star}L_x\tau\big)\Big)+\exp\Big(B_\alpha\exp(L_x\tau)\Big)\Big]$$

when using signatures and

$$M:=\exp(B_{\beta,2}B_{\mathbf{W}})\Big[\exp\Big(\|\mathbf{G}_\star(0)\|_{\text{op}}L_x\tau\exp\big(L_{\mathbf{G}_\star}L_x\tau\big)\Big)+\exp\Big(\|\mathbf{G}_\psi(0)\|_{\text{op}}L_x\tau\exp\big(L_{\mathbf{G}_\psi}L_x\tau\big)\Big)\Big]$$

when using NCDEs.

Proof. Define

$$g : t \mapsto \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big[\exp\big(t(\log\lambda_\theta^{i,D}(s) - \log\lambda_\star^i(s))\big) - t\log\frac{\lambda_\theta^{i,D}(s)}{\lambda_\star^i(s)} - 1\Big]\lambda_\star^i(s)Y^i(s)ds.$$

This function satisfies all assumptions needed in Lemma A.8. The function $f : t \mapsto \exp(\gamma t) - \gamma t - 1$ is convex for all $\gamma \in \mathbb{R}$. Convexity is preserved by integration against a positive function: indeed, if $f(t,s)$ is convex in t for all s and $h(s) \geq 0$ for all $s \in \mathcal{A}$, then

$$\int_{\mathcal{A}} f(t,s)h(s)ds$$

is convex in t (see Boyd & Vandenberghe, 2004, Page 79). The function is also clearly C^∞. Finally, remark that by differentiating the integral, one obtains

$$g'(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big[\big(\log\lambda_\theta^{i,D}(s) - \log\lambda_\star^i(s)\big)\exp\big(t(\log\lambda_\theta^{i,D}(s) - \log\lambda_\star^i(s))\big) - \log\frac{\lambda_\theta^{i,D}(s)}{\lambda_\star^i(s)}\Big]\lambda_\star^i(s)Y^i(s)ds,$$

$$g^{(2)}(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big(\log\lambda_\theta^{i,D}(s) - \log\lambda_\star^i(s)\big)^2\exp\big(t(\log\lambda_\theta^{i,D}(s) - \log\lambda_\star^i(s))\big)\lambda_\star^i(s)Y^i(s)ds,$$

$$g^{(3)}(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big(\log\lambda_\theta^{i,D}(s) - \log\lambda_\star^i(s)\big)^3\exp\big(t(\log\lambda_\theta^{i,D}(s) - \log\lambda_\star^i(s))\big)\lambda_\star^i(s)Y^i(s)ds.$$

Hence

$$\begin{aligned}|g^{(3)}(t)| &\leq \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\|\lambda_\theta^{i,D} - \lambda_\star^i\right\|_\infty (\log\lambda_\theta^{i,D}(s) - \log\lambda_\star^i(s))^2\exp\big(t(\log\lambda_\theta^{i,D}(s) - \log\lambda_\star^i(s))\big)\lambda_\star^i(s)Y^i(s)ds \\ &= Mg^{(2)}(t)\end{aligned}$$

for all $t \in \mathbb{R}$ with

$$M := \max_{i=1,\dots,n}\left\|\lambda_\theta^{i,D} - \lambda_\star^i\right\|_\infty.$$

Using now Lemma A.8, we get at $t = 1$

$$\frac{g^{(2)}(0)}{M^2}\Phi(-M) \leq g(1)\underbrace{-g(0) - g'(0)}_{=0} \leq \frac{g^{(2)}(0)}{M^2}\Phi(M).$$

We have

$$g^{(2)}(0) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big(\log\lambda_\theta^{i,D}(s) - \log\lambda_\star^i(s)\big)^2\lambda_\star^i(s)Y^i(s)ds$$

Finally, remark that

$$g(1) = \mathrm{KL}_n(\lambda_\star, \lambda_\theta),$$

and hence

$$\mathrm{KL}_n(\lambda_\star, \lambda_\theta) \leq \frac{\exp M - M - 1}{nM}\sum_{i=1}^{n}\int_0^\tau \big(\log\lambda_\theta^{i,D}(s) - \log\lambda_\star^i(s)\big)^2\lambda_\star^i(s)Y^i(s)ds.$$

Turning to the constant M, remark that

$$M := \max_{i=1,\dots,n}\left\|\lambda_\theta^i - \lambda_\star^i\right\|_\infty \leq \max_{i=1,\dots,n}\left\|\lambda_\theta^i\right\|_\infty + \max_{i=1,\dots,n}\left\|\lambda_\star^i\right\|_\infty. \tag{6}$$

Similarly to what is done in the proof of Proposition B.1, we have

$$\max_{i=1,\dots,n} \left\|\lambda_\star^i\right\|_\infty \leq \exp\left(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\star}L_x\tau\right)\right),$$

and

$$\max_{i=1,\dots,n} \left\|\lambda_\theta^{i,D}\right\|_\infty \leq \exp(B_{\beta,2}B_{\mathbf{W}})\left[\exp\left(B_\alpha \exp(L_x\tau)\right)\right]$$

when using signatures and

$$\max_{i=1,\dots,n} \left\|\lambda_\theta^{i,D}\right\|_\infty \leq \exp\left(B_{\beta,2}B_{\mathbf{W}}\right)\exp\left[\|\mathbf{G}_\psi(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\psi}L_x\tau\right)\right]$$

when using NCDEs. □

B.3. Formal Statement and Proof of the Risk Decomposition

Proposition B.3. *Let $\lambda_\star$ be the true intensity defined in Equations 1- 2 and λ_θ^D be the intensity parameterized by $\theta \in \Theta$. Under Assumptions 1, 2, 3 and 4, for signature-based embeddings and $\theta = (\alpha_{\star,N}, \beta_\star)$, it holds that*

$$\mathrm{D}_n^2(\lambda_\star, \lambda_\theta^D) \leq 2\tau \exp\left(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\star}L_x\tau\right)\right) \underbrace{\left[\frac{(dL_x)^{N+1}}{(N+1)!}\Lambda_{N+1}(\mathbf{G}_\star)\right]^2 \frac{\tau^{2N+3}}{2N+3}}_{\text{Approximation bias}}$$
$$+ 2\tau \exp\left(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\star}L_x\tau\right)\right) \underbrace{c_3(N)^2 \sum_{k=1}^{N} d^k \Gamma_k(\mathbf{G}_\star)^2 |D|^2}_{\text{Discretization bias}}$$

whereas for NCDEs, for $\theta = (\alpha, \psi, \beta_\star)$, it holds that

$$\mathrm{D}_n^2(\lambda_\star, \lambda_\theta^D) \leq 2\tau \exp(2L_{\mathbf{G}_\psi}L_x)\Bigg[\underbrace{\left(B_\alpha\left(1 + L_{\mathbf{G}_\psi}L_x\mathcal{K}\right)L_x|D|\right)^2}_{\text{Discretization bias}} + \underbrace{\left(\max_{v\in\Omega}\left\|\alpha^\top \mathbf{G}_\psi(v) - \mathbf{G}_\star(v)\right\|\right)^2}_{\text{Approximation bias}}\Bigg]$$
$$\times \exp\left(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\star}L_x\tau\right)\right).$$

Proof. We have the following.

NCDEs. We first consider the case where the individual time series are embedded using NCDEs. Considering a general $\theta \in \Theta$, we then have

$$\begin{aligned}
\mathrm{D}_n^2(\lambda_\star, \lambda_\theta^D) &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left(\alpha^\top z_\theta^{i,D}(s) + \beta^\top \mathbf{W}^i - z_\star^i(s) + \beta_\star^\top \mathbf{W}^i\right)^2 \lambda_\star^i(s) Y^i(s) ds \\
&= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left(\alpha^\top z_\theta^{i,D}(s) - z_\star^i(s) + (\beta - \beta_\star)^\top \mathbf{W}^i\right)^2 \lambda_\star^i(s) Y^i(s) ds \\
&\leq \frac{2}{n}\sum_{i=1}^n \int_0^\tau \left(\alpha^\top z_\theta^{i,D}(s) - z_\star^i(s)\right)^2 \lambda_\star^i(s) Y^i(s) ds + \frac{2}{n}\sum_{i=1}^n \int_0^\tau \left((\beta - \beta_\star)^\top \mathbf{W}^i\right)^2 \lambda_\star^i(s) Y^i(s) ds,
\end{aligned}$$

where the last inequality is obtained using $(a+b)^2 \leq 2a^2 + 2b^2$. Since this is true for all $\theta \in \Theta$, we first chose $\beta = \beta_\star$, hence cancelling the second term. Turning to the remaining term, we have

$$\alpha^\top z_\theta^{i,D}(s) - z_\star^i(s) = \alpha^\top z_\theta^{i,D}(s) - \alpha^\top z_\theta^i(s) + \alpha^\top z_\theta^i(s) - z_\star^i(s),$$

where $z_\theta^i(s)$ is the solution to the CDE driven by continuous unobserved path

$$dz_\theta^i(t) = \mathbf{G}_\psi(z_\theta^i(t))dx^i(t)$$

with initial condition $z^i(0) = 0 \in \mathbb{R}$. Using the continuity of the flow (Theorem A.2), one obtains

$$\begin{aligned}|\alpha^\top z_\theta^{i,D}(s) - \alpha^\top z_\theta^i(s)| &\leq \|\alpha\| \left\|z_\theta^{i,D}(s) - z_\theta^i(s)\right\| \\ &\leq B_\alpha \exp(L_{\mathbf{G}_\psi} L_x)\big(1 + L_{\mathbf{G}_\psi} L_x \mathcal{K}\big) \left\|x^i - x^{i,D}\right\|_\infty .\end{aligned}$$

Using the fact that

$$\left\|x^i - x^{i,D}\right\| \leq L_x|D|,$$

one finally obtains

$$|\alpha^\top z_\theta^{i,D}(s) - \alpha^\top z_\theta^i(s)| \leq B_\alpha \exp(L_{\mathbf{G}_\psi} L_x)\big(1 + L_{\mathbf{G}_\psi} L_x \mathcal{K}\big) L_x|D|.$$

Using again the continuity of the flow, one also has that

$$|\alpha^\top z_\theta^i(s) - z_\star^i(s)| \leq \exp(L_{\mathbf{G}_\psi} L_x) \max_{v\in\Omega} \left\|\alpha^\top \mathbf{G}_\psi(v) - \mathbf{G}_\star(v)\right\|$$

since $\alpha^\top z_\theta^i(s)$ is the solution to the CDE

$$du^i(t) = \alpha^\top \mathbf{G}_\psi(u^i(t))dx^i(t)$$

with initial condition $u^i(0) = 0$. Putting everything together, one obtains

$$(\alpha^\top z_\theta^{i,D}(s) - z_\star^i(s))^2 \leq 2\exp(2L_{\mathbf{G}_\psi} L_x)\left[\Big(B_\alpha\big(1 + L_{\mathbf{G}_\psi} L_x \mathcal{K}\big) L_x|D|\Big)^2 + \Big(\max_{v\in\Omega} \left\|\alpha^\top \mathbf{G}_\psi(v) - \mathbf{G}_\star(v)\right\| \Big)^2\right].$$

Plugging this result in the original inequality on the squared log-divergence yields

$$\begin{aligned}\mathrm{D}_n^2(\lambda_\star, \lambda_\theta^D) \leq& 2\tau \exp(2L_{\mathbf{G}_\psi} L_x)\left[\Big(B_\alpha\big(1 + L_{\mathbf{G}_\psi} L_x \mathcal{K}\big) L_x|D|\Big)^2 + \Big(\max_{v\in\Omega} \left\|\alpha^\top \mathbf{G}_\psi(v) - \mathbf{G}_\star(v)\right\| \Big)^2\right] \\ &\times \exp\left(B_{\beta,2} B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x \tau \exp\left(L_{\mathbf{G}_\star} L_x \tau\right)\right).\end{aligned}$$

Signature-based embeddings. We proceed in a similar fashion. We have for any $\theta \in \Theta$ that

$$\begin{aligned}\big(\log \lambda_\theta^{i,D}(s) - \log \lambda_\star^i(s)\big)^2 &= \Big(\alpha^\top \mathbf{S}_N(x_{[0,s]}^{i,D}) + \beta^\top \mathbf{W}^i - \alpha_\star^\top \mathbf{S}(x_{[0,s]}^i) - \beta_\star^\top \mathbf{W}^i\Big)^2 \\ &= \Big(\alpha^\top \mathbf{S}_N(x_{[0,s]}^{i,D}) - \alpha_\star^\top \mathbf{S}(x_{[0,s]}^i) + (\beta - \beta_\star)^\top \mathbf{W}^i\Big)^2.\end{aligned}$$

Now, we furthermore have

$$\alpha^\top \mathbf{S}_N(x_{[0,s]}^{i,D}) - \alpha_\star^\top \mathbf{S}(x_{[0,s]}^i) = \alpha^\top \mathbf{S}_N(x_{[0,s]}^{i,D}) - \alpha_{\star,N}^\top \mathbf{S}_N(x_{[0,s]}^i) + \alpha_{\star,N}^\top \mathbf{S}_N(x_{[0,s]}^i) - \alpha_\star^\top \mathbf{S}(x_{[0,s]}^i).$$

In particular, taking $\alpha = \alpha_{\star,N}$, we have

$$\alpha_{\star,N}^\top \mathbf{S}_N(x_{[0,s]}^{i,D}) - \alpha_\star^\top \mathbf{S}(x_{[0,s]}^i) = \alpha_{\star,N}^\top \Big(\mathbf{S}_N(x_{[0,s]}^{i,D}) - \mathbf{S}_N(x_{[0,s]}^i)\Big) + \alpha_{\star,N}^\top \mathbf{S}_N(x_{[0,s]}^i) - \alpha_\star^\top \mathbf{S}(x_{[0,s]}^i).$$

Using Proposition A.4, one obtains

$$|\alpha_{\star,N}^\top \mathbf{S}_N(x_{[0,s]}^i) - \alpha_\star^\top \mathbf{S}(x_{[0,s]}^i)| \leq \frac{(dL_x s)^{N+1}}{(N+1)!} \Lambda_{N+1}(\mathbf{G}_\star).$$

Additionally, using Theorem A.7, we have

$$\left|\alpha_{\star,N}^\top\left(\mathbf{S}_N(x^{i,D}_{[0,s]}) - \mathbf{S}_N(x^i_{[0,s]})\right)\right| \le \|\alpha_{\star,N}\|\, c_3(N)|D|.$$

We obtain for $\theta = (\alpha_{\star,N}, \beta_\star)$ that

$$\left(\log \lambda^{i,D}_\theta(s) - \log \lambda^i_\star(s)\right)^2 \le 2\Big[\frac{(dL_x s)^{N+1}}{(N+1)!}\Lambda_{N+1}(\mathbf{G}_\star)\Big]^2 + 2\,\|\alpha_{\star,N}\|^2\, c_3(N)^2|D|^2$$

using the fact that $(a+b)^2 \le 2a^2 + 2b^2$. Finally, integrating yields

$$\begin{aligned}
\mathrm{D}^2_n(\lambda_\star, \lambda^D_\theta) \le{}& \frac{1}{n}\sum_{i=1}^n \exp\left(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\star}L_x\tau\right)\right) 2\tau\Big[\frac{(dL_x)^{N+1}}{(N+1)!}\Lambda_{N+1}(\mathbf{G}_\star)\Big]^2 \int_0^\tau s^{2(N+1)}ds \\
&+ \frac{1}{n}\sum_{i=1}^n \exp\left(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\star}L_x\tau\right)\right) 2\tau\,\|\alpha_{\star,N}\|^2\, c_3(N)^2|D|^2 \\
&\le 2\tau \exp\left(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\star}L_x\tau\right)\right)\left[\frac{(dL_x)^{N+1}}{(N+1)!}\Lambda_{N+1}(\mathbf{G}_\star)\right]^2 \frac{\tau^{2N+3}}{2N+3} \\
&+ 2\tau \exp\left(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\star}L_x\tau\right)\right)\|\alpha_{\star,N}\|^2\, c_3(N)^2|D|^2.
\end{aligned}$$

Using Lemma A.6, we can furthermore simplify the bound to

$$\begin{aligned}
\mathrm{D}^2_n(\lambda_\star, \lambda^D_\theta) \le{}& 2\tau \exp\left(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\star}L_x\tau\right)\right)\left[\frac{(dL_x)^{N+1}}{(N+1)!}\Gamma_{N+1}(\mathbf{G}_\star)\right]^2 \frac{\tau^{2N+3}}{2N+3} \\
&+ 2\tau \exp\left(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\star}L_x\tau\right)\right) c_3(N)^2 \sum_{k=1}^N d^k\Gamma_k(\mathbf{G}_\star)^2|D|^2.
\end{aligned}$$

□

B.4. Proof of Theorem 3.1

Theorem B.4 (Formal statement of Theorem 3.1)**.** *Consider the signature-based embedding. Let $\hat\theta = (\hat\alpha, \hat\beta)$ be the solution of (3) with $\text{pen}(\theta) = \eta_1\|\alpha\|_1 + \eta_2\|\beta\|_1$. Under Assumptions 1, 2, 3 and 4, and writing $\beta_\star = (\beta_\star^{(1)}, \dots, \beta_\star^{(s)})$, we have for any $N \ge 1$ and any $x > 0$ that with probability greater than $1 - 4e^{-x}$*

$$\begin{aligned}
\ell^D_n(\hat\theta) - \ell^\star_n \le{}& 2\tau \exp\left(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\star}L_x\tau\right)\right) \underbrace{\left[\frac{(dL_x)^{N+1}}{(N+1)!}\Gamma_{N+1}(\mathbf{G}_\star)\right]^2 \frac{\tau^{2N+3}}{2N+3}}_{\text{Approximation bias}} \\
&+ 2\tau \exp\left(B_{\beta,2}B_{\mathbf{W}} + \|\mathbf{G}_\star(0)\|_{\text{op}} L_x\tau \exp\left(L_{\mathbf{G}_\star}L_x\tau\right)\right) \underbrace{c_3(N)^2 \sum_{k=1}^N d^k\Gamma_k(\mathbf{G}_\star)^2|D|^2}_{\text{Discretization bias}} \\
&+ 4\frac{(L_x\tau)^{k^\star}}{k^\star!}\sqrt{\frac{2\big(x + \log(Nd^N)\big)\lambda_\infty}{n}} \sum_{k=1}^N d^k\Gamma_k(\mathbf{G}_\star) + 4B_{\beta,0}\sup_{k=1,\dots,s}|\beta_\star^{(k)}|B_{\mathbf{W}}\sqrt{\frac{2\tau\lambda_\infty(x+\log s)}{n}}.
\end{aligned}$$

Proof. By optimality of $\hat\theta$, we have for all $\theta \in \Theta$ that

$$\ell^D_n(\hat\theta) + \text{pen}(\hat\theta) - \ell^\star_n \le \ell^D_n(\theta) + \text{pen}(\theta) - \ell^\star_n$$

and hence, using Proposition A.9, one obtains

$$\mathrm{KL}_n(\lambda_\star, \lambda_{\hat{\theta}}^D) \leq \mathrm{KL}_n(\lambda_\star, \lambda_\theta^D) + \mathrm{pen}(\theta) - \mathrm{pen}(\hat{\theta}) + \frac{1}{n}\sum_{i=1}^n \int \log \frac{\lambda_{\hat{\theta}}^{i,D}(s)}{\lambda_\theta^{i,D}(s)} dM^i(s).$$

Using Proposition A.10, the KL-divergence on the right hand side can be bounded by the squared log divergence, yielding

$$\mathrm{KL}_n(\lambda_\star, \lambda_{\hat{\theta}}^D) \leq c_2 \mathrm{D}_n^2(\lambda_\star, \lambda_\theta^D) + \mathrm{pen}(\theta) - \mathrm{pen}(\hat{\theta}) + \frac{1}{n}\sum_{i=1}^n \int \log \frac{\lambda_{\hat{\theta}}^{i,D}(s)}{\lambda_\theta^{i,D}(s)} dM^i(s). \tag{7}$$

The "bias term" $c_2 \mathrm{D}_n^2(\lambda_\star, \lambda_\theta^D)$ can be bounded thanks to Proposition B.3. We shall now derive a bound for the term

$$\mathrm{pen}(\theta) - \mathrm{pen}(\hat{\theta}) + \frac{1}{n}\sum_{i=1}^n \int \log \frac{\lambda_{\hat{\theta}}^{i,D}(s)}{\lambda_\theta^{i,D}(s)} dM^i(s).$$

Since Equation (7) holds true for all $\theta \in \Theta$, we can set $\theta = (\alpha_{\star,N}, \beta_\star)$. We now study the term

$$\frac{1}{n}\sum_{i=1}^n \int_0^\tau \log \frac{\lambda_{\hat{\theta}}^{i,D}(s)}{\lambda_{\theta_N^\star}^{i,D}(s)} dM^i(s) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left(\hat{\alpha}^\top \mathbf{S}_N(x_{[0,s]}^{i,D}) + \hat{\beta}^\top \mathbf{W}^i - \alpha_{\star,N}^\top \mathbf{S}(x_{[0,t]}^i) - \beta_\star^\top \mathbf{W}^i\right) dM^i(s) \tag{8}$$

$$= (\hat{\alpha} - \alpha_{\star,N})^\top \frac{1}{n}\sum_{i=1}^n \int_0^\tau \mathbf{S}_N(x_{[0,s]}^{i,D}) dM^i(s) + (\hat{\beta} - \beta_\star)^\top \frac{1}{n}\sum_{i=1}^n \mathbf{W}^i M^i(t). \tag{9}$$

which appears in Proposition A.9 when considering signature based embeddings. We make a repeated use of the following lemmas in our derivation of a bound for this term.

Lemma B.5. *Let $\mathbf{S}_{[k],j}(x_{[0,t]}^{i,D}) \in \mathbb{R}$ be the signature coefficient associated to the j-th word of the k-th signature layer of a time series $x_{[0,t]}^{i,D}$ evaluated at time $t \leq \tau$. Then we have*

$$\left\|\mathbf{S}_{[k],j}(x_{[0,\cdot]}^{i,D})\right\|_{\infty,[0,t]} = \max_{s\in[0,t]} |\mathbf{S}_{[k],j}(x_{[0,s]}^{i,D})| \leq \frac{(L_x t)^k}{k!} \leq \frac{(L_x \tau)^k}{k!}.$$

Proof. For all $s \in [0,t]$, we have

$$|\mathbf{S}_{[k],j}(x_{[0,s]}^{i,D})| \leq \left\|\mathbf{S}_{[k]}(x_{[0,s]}^{i,D})\right\| \leq \frac{\left\|x_{[0,t]}^{i,D}\right\|_{\text{1-var}}^k}{k!}$$

where $\mathbf{S}_{[k]}(x_{[0,s]}^{i,D})$ refers to the full signature layer of depth k, and the last inequality can be found in Fermanian (2021). Using Assumption 1, we have

$$\frac{\left\|x_{[0,t]}^{i,D}\right\|_{\text{1-var}}^k}{k!} \leq \frac{\left\|x_{[0,t]}^{i}\right\|_{\text{1-var}}^k}{k!} \leq \frac{(L_x t)^k}{k!} \leq \frac{(L_x \tau)^k}{k!}.$$

□

The following deviation inequality is a direct consequence from the one in Van de Geer (1995) and derives from inequalities for general martingales that can be found in Shorack & Wellner (2009) for instance.

Lemma B.6 (Deviation inequality for a martingale)**.** *Let Υ be a locally square integrable martingale .Then, for any $x > 0$ and $t \geq 0$, the following holds true for*

$$\mathbb{P}\Big(\left|\Upsilon(t)\right| \geq \sqrt{2v(t)x} + \frac{B(t)x}{3}, \langle\Upsilon(t)\rangle \leq v(t), \sup_{s\in[0,t]} |\Delta\Upsilon(s)| \leq B(t)\Big) \leq 2e^{-x}, \tag{10}$$

where $\langle\Upsilon(t)\rangle$ is the predictable variation of Υ and $\Delta\Upsilon(t)$ its jump at time t.

Going back to Equation (8) and decomposing on the signature layers, we can write

$$\Big|(\hat{\alpha}-\alpha_{\star,N})^\top \frac{1}{n}\sum_{i=1}^n \int_0^\tau \mathbf{S}_N(x^{i,D}_{[0,s]})dM^i(s)\Big| = \Big|\sum_{k=1}^N (\hat{\alpha}_{[k]}-\alpha_{\star,[k]})^\top \frac{1}{n}\sum_{i=1}^n \int_0^\tau \mathbf{S}_{[k]}(x^{i,D}_{[0,s]})dM^i(s)\Big|$$
$$\leq \sum_{k=1}^N \big\|\hat{\alpha}_{[k]}-\alpha_{\star,[k]}\big\|_1 \sup_{1\leq j\leq d^k}\Big|\frac{1}{n}\sum_{i=1}^n \int_0^\tau \mathbf{S}_{[k],j}(x^{i,D}_{[0,s]})dM^i(s)\Big|.$$

Because the martingales M^i are independent, and the signature coefficients bounded, the term

$$\chi_n(\tau) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \mathbf{S}_{[k],j}(x^{i,D}_{[0,s]})dM^i(s)$$

is itself a martingale. Moreover, since each M^i comes from a counting process via a Doob Meier decomposition, its jumps are bounded by 1 and, at a given time, there is (almost surely) at most one M^i that jumps. As a consequence, we get the following bound with jumps bounded by

$$\sup_{t\in[0,\tau]}\big|\Delta\chi_n(t)\big| \leq \frac{1}{n}\sup_{i=1,\dots,n}\big\|\mathbf{S}_{[k],j}(x^{i,D}_{[0,\cdot]})\big\|_{\infty,[0,\tau]} \leq \frac{(L_x\tau)^k}{nk!},$$

and quadratic variation given at time $t\in[0,\tau]$ by

$$\begin{aligned}\langle\chi_n(t)\rangle &= \Big\langle \frac{1}{n}\sum_{i=1}^n \int_0^t \mathbf{S}_{[k],j}(x^{i,D}_{[0,s]})dM^i(s)\Big\rangle = \frac{1}{n^2}\sum_{i=1}^n \int_0^t \mathbf{S}_{[k],j}(x^{i,D}_{[0,s]})^2\lambda^i_\star(s)Y^i(s)ds \\ &\leq \sup_{i=1,\dots,n}\|\mathbf{S}_{[k],j}(x^{i,D}_{[0,\cdot]})\|^2_{\infty,[0,\tau]}\frac{1}{n^2}\sum_{i=1}^n \Lambda^i_\star(t) \\ &\leq \frac{1}{n}\sup_{i=1,\dots,n}\|\mathbf{S}_{[k],j}(x^{i,D}_{[0,\cdot]})\|^2_{\infty,[0,\tau]}\sup_{i=1,\dots,n}\Lambda^i_\star(t) \\ &\leq \frac{L_x^{2k}\tau^{2k+1}\lambda_\infty}{n(k!)^2},\end{aligned} \tag{11}$$

where we have used Lemma B.5 and the fact that

$$\Lambda^i_\star(t) = \int_0^t \lambda^i_\star(s)Y^i(s)ds \leq t\sup_{s\in[0,t]}\lambda^i_\star(s) \leq \tau\sup_{s\in[0,\tau]}\lambda^i_\star(s) \leq \tau\lambda_\infty,$$

with

$$\lambda_\infty = \exp\big(B_{\beta,2}B_{\mathbf{W}}\big)\exp\big(\|\mathbf{G}_\star(0)\|_{\text{op}}L_x\tau\exp\big(L_{\mathbf{G}_\star}L_x\tau\big)\big)$$

according to Lemma 2.1. Lemma B.6 now warrants that for any $\varepsilon>0$ with a probability greater than $1-2e^{-\varepsilon}$

$$\Big|\frac{1}{n}\sum_{i=1}^n\int_0^\tau \mathbf{S}_{[k],j}(x^{i,D}_{[0,s]})dM^i(s)\Big| \leq \sqrt{\frac{2\varepsilon L_x^{2k}\tau^{2k+2}\lambda_\infty}{n(k!)^2}} + \frac{\varepsilon(L_x\tau)^k}{3nk!} \leq \frac{(L_x\tau)^{k^\star}}{k^\star!}\Big(\sqrt{\frac{2\varepsilon\tau^2\lambda_\infty}{n}}+\frac{\varepsilon}{3n}\Big)$$

where

$$k^\star = \text{argmax}_{k\geq 1}\frac{(L_x\tau)^k}{k!}.$$

A double union bound on the signature layers and the signature coefficients within each layer ensures that for any $\varepsilon>0$ with a probability greater than $1-2e^{-\varepsilon}$

$$\sup_{1\leq k\leq N}\sup_{1\leq j\leq d^k}\Big|\frac{1}{n}\sum_{i=1}^n\int_0^\tau \mathbf{S}_{[k],j}(x^{i,D}_{[0,s]})dM^i(s)\Big| \leq \frac{(L_x\tau)^{k^\star}}{k^\star!}\Big(\sqrt{\frac{2\big(\varepsilon+\log(Nd^N)\big)\tau^2\lambda_\infty}{n}}+\frac{\varepsilon+\log(Nd^N)}{3n}\Big).$$

As a consequence

$$\Big|(\hat{\alpha}-\alpha_{\star,N})^\top \frac{1}{n}\sum_{i=1}^n \int_0^\tau \mathbf{S}_N(x^{i,D}_{[0,s]})dM^i(s)\Big| \le \|\hat{\alpha}-\alpha_{\star,N}\|_1 \frac{(L_x\tau)^{k^\star}}{k^\star!}\Big(\sqrt{\frac{2\big(\varepsilon+\log(Nd^N)\big)\tau^2\lambda_\infty}{n}}+\frac{\varepsilon+\log(Nd^N)}{3n}\Big)$$

for any $\varepsilon > 0$ with a probability greater than $1-2e^{-\varepsilon}$.

We apply the same line of reasoning to

$$\Big|(\hat{\beta}-\beta_\star)^\top \frac{1}{n}\sum_{i=1}^n \mathbf{W}^i M^i(t)\Big| \le \|\hat{\beta}-\beta_\star\|_1 \sup_{1\le m\le s}\Big|\frac{1}{n}\sum_{i=1}^n W^i_m M^i(t)\Big|.$$

For each m, the term $\sum_{i=1}^n W^i_m M^i(\cdot)$ is a martingale with predictable variation less than

$$\sum_{i=1}^n (W^i_m)^2\Lambda^i_\star(t) \le B^2_{\mathbf{W}}\tau\lambda_\infty.$$

Its jumps are bounded by $B_{\mathbf{W}}$, and therefore Lemma B.6 applies. Via an union bound, we deduce that for any $\varepsilon > 0$ and with a probability greater that $1-2e^{-\varepsilon}$

$$\Big|(\hat{\beta}-\beta_\star)^\top \frac{1}{n}\sum_{i=1}^n \mathbf{W}^i M^i(t)\Big| \le \|\hat{\beta}-\beta_\star\|_1\sqrt{\frac{2B^2_{\mathbf{W}}\tau\lambda_\infty(\varepsilon+\log s)}{n}}+\frac{B_{\mathbf{W}}(\varepsilon+\log s)}{3n}.$$

Now defining the penalty

$$\begin{aligned}\text{pen}(\theta) &= \|\alpha\|_1\frac{(L_x\tau)^{k^\star}}{k^\star!}\Big(\sqrt{\frac{2\big(\varepsilon+\log(Nd^N)\big)\tau^2\lambda_\infty}{n}}+\frac{\varepsilon+\log(Nd^N)}{3n}\Big)\\ &\quad+\|\beta\|_1\Big(\sqrt{\frac{2B^2_{\mathbf{W}}\tau\lambda_\infty(\varepsilon+\log s)}{n}}+\frac{B_{\mathbf{W}}(\varepsilon+\log s)}{3n}\Big),\end{aligned}$$

we obtain

$$\begin{aligned}&\text{pen}(\theta_{\star,N})-\text{pen}(\hat{\theta})+\frac{1}{n}\sum_{i=1}^n\int\log\frac{\lambda^{i,D}_{\hat{\theta}}(s)}{\lambda^{i,D}_{\theta}(s)}dM^i(s)\\ &\le 2\|\alpha_{\star,N}\|_1\frac{(L_x\tau)^{k^\star}}{k^\star!}\Big(\sqrt{\frac{2\big(x+\log(Nd^N)\big)\tau^2\lambda_\infty}{n}}+\frac{x+\log(Nd^N)}{3n}\Big)\\ &\quad+2\|\beta_\star\|_1\Big(\sqrt{\frac{2B^2_{\mathbf{W}}\tau\lambda_\infty(x+\log s)}{n}}+\frac{B_{\mathbf{W}}(x+\log s)}{3n}\Big)\end{aligned}$$

with a probability greater that $1-4e^{-\varepsilon}$ for any $\varepsilon > 0$. For n large enough, we can write

$$\begin{aligned}&2\|\alpha_{\star,N}\|_1\frac{(L_x\tau)^{k^\star}}{k^\star!}\Big(\sqrt{\frac{2\big(\varepsilon+\log(Nd^N)\big)\tau^2\lambda_\infty}{n}}+\frac{\varepsilon+\log(Nd^N)}{3n}\Big)\\ &+2\|\beta_\star\|_1\Big(B_{\mathbf{W}}\sqrt{\frac{2\tau\lambda_\infty(\varepsilon+\log s)}{n}}+\frac{B_{\mathbf{W}}(\varepsilon+\log s)}{3n}\Big)\\ &\le 4\|\alpha_{\star,N}\|_1\frac{(L_x\tau)^{k^\star}}{k^\star!}\sqrt{\frac{2\big(\varepsilon+\log(Nd^N)\big)\tau^2\lambda_\infty}{n}}+4\|\beta_\star\|_1 B_{\mathbf{W}}\sqrt{\frac{2\tau\lambda_\infty(\varepsilon+\log s)}{n}}.\end{aligned}$$

Finally, using Lemma A.6 and Assumption 3, we can bound the ℓ_1 norms of $\alpha_{\star,N}$ and $\beta_{\star,N}$ and obtain that for large n, for any $\varepsilon > 0$ we have with probability greater than $1 - 4e^{-\varepsilon}$ that

$$
\begin{aligned}
&4\|\alpha_{\star,N}\|_1 \frac{(L_x\tau)^{k^\star}}{k^\star!}\sqrt{\frac{2\big(\varepsilon + \log(Nd^N)\big)\tau^2\lambda_\infty}{n}} + 4\|\beta_\star\|_1 B_{\mathbf{W}}\sqrt{\frac{2\tau\lambda_\infty(\varepsilon + \log s)}{n}} \\
&\quad \le 4\frac{(L_x\tau)^{k^\star}}{k^\star!}\sqrt{\frac{2\big(\varepsilon + \log(Nd^N)\big)\tau^2\lambda_\infty}{n}}\sum_{k=1}^{N} d^k\Gamma_k(\mathbf{G}_\star) + 4B_{\beta,0}\sup_{k=1,\dots,s}|\beta_\star^{(k)}|B_{\mathbf{W}}\sqrt{\frac{2\tau\lambda_\infty(\varepsilon + \log s)}{n}}.
\end{aligned}
$$

□

C. Algorithmic and Implementation Details

In this Section, we provide extra information about learning algorithms described in the main paper and their hyperparameters optimization by gridsearch.

C.1. Description of Competing Methods

C.1.1. CoxSig and CoxSig+

Implementation. We use `iisignature` (Reizenstein & Graham, 2020) to compute signatures. Alternatives for computing signatures include the `signatory` library (Kidger & Lyons, 2020).

Training. We minimize the penalized negative log-likelihood (defined in 3 in the main paper) using a vanilla proximal point algorithm (Boyd & Vandenberghe, 2004).

Hyperparameters. The initial learning rate of the proximal gradient algorithm is set to e^{-3} and the learning rate for each iteration is chosen by back tracking linesearch method (Boyd & Vandenberghe, 2004). The hyperparameters of penalization strength (η_1, η_2) and truncation depth N are chosen by 1-fold cross-validation of a mixed metric equal to the difference between the C-index and the Brier score. We select the best hyperparameters that minimize the average of this mixed metric on the validation set. We list the hyperparameters search space of this algorithm below.

- η_1: $\{1, e^{-1}, e^{-2}, e^{-3}, e^{-4}, e^{-5}\}$;
- η_2: $\{1, e^{-1}, e^{-2}, e^{-3}, e^{-4}, e^{-5}\}$;
- N: $\{2, 3\}$. Larger values were considered in the beginning of experiments but were removed from the cross-validation grid because they yielded bad performance and numerical instabilities.

C.1.2. NCDE

Implementation. We implement the fill-forward discrete update of NCDEs in `Pytorch`.

Structure. The neural vector field is a feed-forward network composed of two fully connected hidden layers whose hidden dimension is set to 128. We choose to represent the latent state in 4 dimensions—the number of nodes in the input layer is therefore set to 4. The dimension of the output layer is equal to the multiplication of the dimension of the hidden layer (128) and the dimension of the sample paths of a given data set. `tanh` is set to be the activation function for all the nodes in the network.

Training. The model was trained for 50 epochs using the Adam optimizer (Kingma & Ba, 2015) with a batch size of 32 and cross-validated learning rate set to e^{-4}.

C.1.3. Cox Model

Implementation and Training. We use a classical Cox model with elastic-net penalty as a baseline, which is given either the first measured value of the individual time series or the static features if they are available. The intensity of this model has then the form

$$\lambda_\theta^i(t) = \lambda_0(t) \exp(\beta^\top \mathbf{W}^i),$$

where $\mathbf{W}^i = \mathbf{X}^i(0)$ if no static features are available. We use the implementation provided in the Python package `scikit-survival` and called `CoxnetSurvivalAnalysis` (Pölsterl, 2020).

Hyperparameters. The ElasticNet mixing parameter γ is set to 0.1. The hyperparameter of penalization strength η is chosen by cross-validation as described above. We crossvalidate over the set $\{1, e^{-1}, e^{-2}, e^{-3}, e^{-4}, e^{-5}\}$ to select the best value.

C.1.4. Random Survival Forest

Implementation. We use the implementation of RSF (Ishwaran et al., 2008) provided in the Python package `scikit-survival` (Pölsterl, 2020).

Training. We train this model with static features $\mathbf{W}^i$ as the only input. Similarly to our implementation of the Cox model, we use the first value of the time series as static features if no other features are available.

Hyperparameters. We cross-validate two hyperparameters on the following grids.

- `max_features`: {`None`, `sqrt`};
- `min_samples_leaf`: $\{1, 5, 10\}$;

C.1.5. Dynamic Deep-Hit (Lee et al., 2019)

DDH is a dynamical survival analysis algorithm that frames dynamical survival analysis as a classification problem. It divides the considered time period $[0, \tau]$ into a set of contiguous time intervals. The network is then trained to predict a time interval of event for every subject, which is a multiclass classification task.

Network Architecture. Being adapted to competing events, Dynamic Deep-Hit combines a shared network with a cause-specific network. The *shared network* is a combination of a RNN-like network that processes the longitudinal data and an attention mechanism, which helps the network decide which part of the history of the measurements is important. The *cause-specific network* is a feed-forward network taking as an input the history of embedded measurements and learning a cause-specific representation. See Figure 6 for a graphical representation of the network's structure.

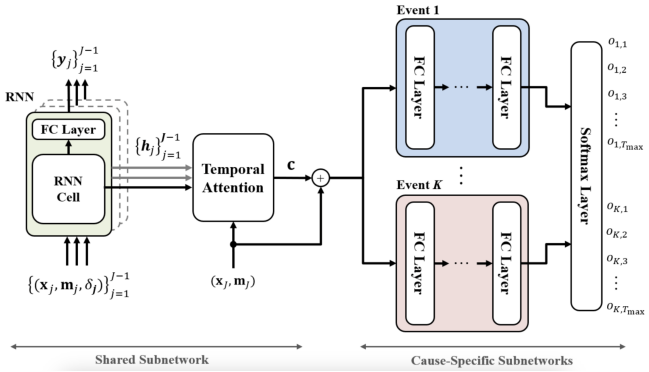

Figure 6: Network structure of Dynamic DeepHit. Figure is taken from Lee et al. (2019).

Loss Function. The loss function of DDH is a sum of three loss functions

$$\ell_{\text{Dynamic DeepHit}} = \ell_{\text{log-likelihood}} + \ell_{\text{ranking}} + \ell_{\text{prediction}}.$$

The first loss maximizes the conditional likelihood of dying in the interval $[t_k, t_{k+1}[$ given that the individual has survived up to time t_k. On a side note, we notice that the claim of Lee et al. (2019) that this loss corresponds to "the negative log-likelihood of the joint distribution of the first hitting time and corresponding event considering the right-censoring" of the data is hence inexact. This might explain the results observed in Figure 4: DDH's performance, in terms of Brier score, strongly degrades as δt increases because the model is only trained to predict one step ahead, instead of maximizing the full likelihood.

The second loss favors correct rankings among at risk individuals: an individual experiencing an event at time T^i should have a higher risk score at time $t < T^i$ than an individual j for which $T^j > T^i$.

The third loss is a prediction loss, which measures the difference between the value of the time-dependent features and a prediction of this value made by the shared network. The loss is minimized using Adam (Kingma & Ba, 2015).

Hyperparameters. In our setting, we use the network in its original structure. The learning rate is set to e^{-4} and the number of epochs to 300.

C.1.6. SURVLATENT ODE (MOON ET AL., 2022)

Network Architecture. SurvLatent ODE is a variational autoencoder architecture (Kingma & Welling, 2013). The encoder embeds the entire longitudinal features into an initial latent state, and the decoder uses this latent state to drive the latent trajectory and to estimate the distribution of event time. In this framework, the encoder is an ODE-RNN architecture (Rubanova et al., 2019), which handles the longitudinal features sequentially backward in time and outputs the posterior over the initial latent state. The decoder, which is adapted to competing events, consists of an ODE model and cause-specific decoder modules. The latent trajectory derived from the ODE model is shared across cause-specific decoder modules to estimate the cause-specific discrete hazard functions. See Figure 7 for a graphical representation of the network's structure.

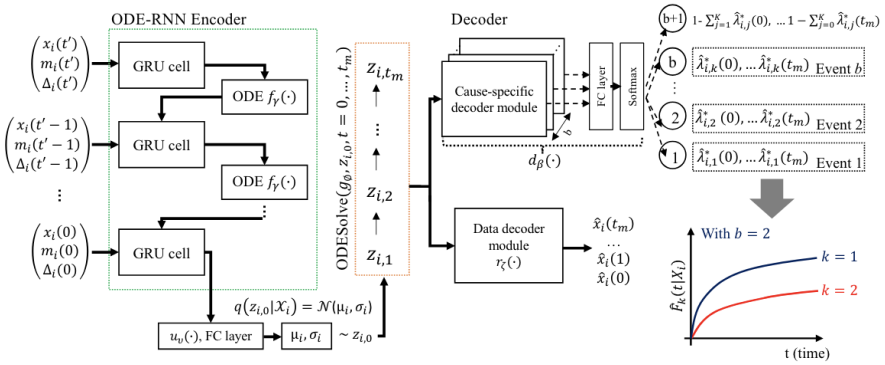

Figure 7: Network structure of SurvLatent ODE. Figure taken from Moon et al. (2022).

Loss Function. The loss function is a combination of the log-likelihood and the Kullback-Leibler divergence between the approximate and the true posterior over the initial latent state.

Hyperparameters. In our setting, we use the network in its original structure. The learning rate is set to e^{-2} and the number of epochs to 15, as in the original paper. The training of this framework cannot use subjects whose last longitudinal measurement time is equal to the event time, which is not the case for our proposed methods as well as other competing methods. In order to avoid this problem, we then stop observing the longitudinal measurement before the time-to-event for a period equal to 80 % of the event time of these subjects when training the model of this framework.

C.2. Computation of the Different Metrics

The following lemma details the computation of the conditional survival function.

Lemma C.1. *For any $i \in \{1, \dots, n\}$,*

$$r^i_\theta(t, \delta t) = \exp\Big(- \int_t^{t+\delta t} \lambda^i_\theta(u, x^{i,D}_{[0,u\wedge t]}) du \Big),$$

where $r^i_\theta(t, \delta t) = \mathbb{P}\big(T^i > t + \delta t \,|\, T^i > t, x^{i,D}_{[0,t]}\big)$ is the survival function of individual i, as estimated by the model with parameters θ, at time $t + \delta t$ for $\delta t > 0$ conditional on survival up to time t, and on observation of the longitudinal features up to time t, and the notation $\lambda^i_\theta(u, x^{i,D}_{[0,u\wedge t]})$ means that the intensity at time u is computed by using the longitudinal features up to time $u \wedge t = \min(u, t)$.

Proof. Since Bayes rule gives

$$r^i_\theta(t, \delta t) = \mathbb{P}\Big(T^i > t + \delta t \,|\, T^i > t, x^{i,D}_{[0,t]}, \mathbf{W}^i\Big) = \frac{\mathbb{P}\Big(T^i > t + \delta t \,|\, x^{i,D}_{[0,t]}, \mathbf{W}^i\Big)}{\mathbb{P}\Big(T^i > t \,|\, x^{i,D}_{[0,t]}, \mathbf{W}^i\Big)},$$

we can compute this score by using the fact that

$$\mathbb{P}\Big(T^i > t \,|\, x^{i,D}_{[0,t]}, \mathbf{W}^i\Big) = \exp(-\Lambda^{i,D}_\theta(t)),$$

where we recall that $\Lambda^{i,D}_\theta(t)$ is the cumulative hazard function

$$\Lambda^{i,D}_\theta(t) := \int_0^t \lambda^{i,D}_\theta(s) Y^i(s) ds.$$

We refer the reader unfamiliar with survival analysis to Aalen et al. (2008, Chapter 1, p. 6) for a proof of this expression of the survival function. This then yields

$$\begin{aligned} r^i_\theta(t, \delta t) &= \frac{\exp(-\int_0^{t+\delta t} \lambda^i_\theta(u, x^{i,D}_{[0,u\wedge t]}) du)}{\exp(-\int_0^{t} \lambda^i_\theta(u, x^{i,D}_{[0,u\wedge t]}) du)} \\ &= \exp(-\int_t^{t+\delta t} \lambda^i_\theta(u, x^{i,D}_{[0,u\wedge t]}) du). \end{aligned}$$

□

Beside the two metrics described in the main paper, we report our results in term of two more metrics namely the weighted Brier Score and the area under the receiver operating characteristic curve (AUC). The details of these metrics are given below.

Weighted Brier Score. The weighted version of the Brier score, which we write $\text{WBS}(t, \delta t)$, is defined as

$$\sum_{i=1}^n \mathbb{1}_{T^i \le t,\, \Delta^i = 1} \frac{r^i_\theta(t, \delta t))^2}{\hat{G}(T^i)} + \mathbb{1}_{T^i \ge t} \frac{(1 - r^i_\theta(t, \delta t))^2}{\hat{G}(t)},$$

where $\hat{G}(\cdot)$ is the probability of censoring weight, estimated by the Kaplan-Meier estimator.

AUC. We define the area under the receiver operating characteristic curve $\text{AUC}(t, \delta t)$ as

$$\frac{\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}_{r^i_\theta(t,\delta t) > r^j_\theta(t,\delta t)} \mathbb{1}_{T^i > t+\delta t,\, T^j \in [t, t+\delta t]} w_j}{(\sum_{i=1}^n \mathbb{1}_{T^i > t+\delta t})(\sum_{i=1}^n \mathbb{1}_{T^i \in [t, t+\delta t]} w_i)},$$

where w_i are inverse probability of censoring weights, estimated by the Kaplan-Meier estimator.

D. Details of Experiments and Datasets

The main characteristics of the datasets used in the paper are summarized in Table 3 and we provide more detailed information of these datasets in subsections below. For the experiments, each dataset is randomly divided into a training set (80%) and test set (20%). Hyperparameter optimization is performed as follows. We split the training set, using 4/5 for training and 1/5 for validation. We then re-fit on the whole training set with the best hyperparameters and report the results on the test set for 10 runs. Note that the performance is evaluated at numerous points (t, δ_t), where t is set to the 5th, 10th, and 20th percentile of the distribution of event times.

D.1. Hitting Time of a partially observed SDE

Time series. The paths $x_t = (x_t^{(1)}, \ldots, x_t^{(d-1)})$ are $(d-1)$-dimensional sample paths of a fractional Brownian motion with Hurst parameter $H = 0.6$, and $B^i(t)$ is a Brownian noise term. We set $d = 5$. The paths are sampled at 1000 times over the time interval $[0, 10]$. All simulations are done using the `stochastic` package[1]. The time series $\mathbf{X}^i$ are identical, up to observation time, to the ones used for simulations.

[1] Available at https://github.com/crflynn/stochastic

Name	n	d	Static Features	Censoring	Avg. Observation Times	Source
Hitting time	500	5	✗	Terminal (3.2%)	177	Simulation
Tumor Growth	500	2	✗	Terminal (8.4%)	250	Simeoni et al. (2004)
Predictive Maintenance	200	17	✗	Online (50%)	167	Saxena et al. (2008)
Churn	1043	14	✗	Terminal (38.4%)	25	Private dataset

Table 3: Description of the 4 datasets we consider. The integer d is the dimension of the time series including the time channel. *Terminal* censoring means that the individuals are censored at the end of the overall observation period $[0, \tau]$ if they have not experienced any event. It is opposed to *online* censoring that can happen at any time in $[0, \tau]$. The reported percentage indicates the censoring level i.e. the share of the population that does not experience the event.

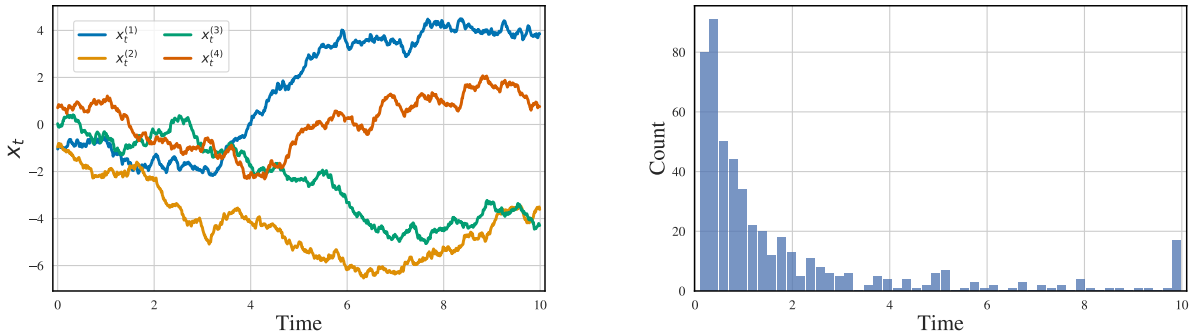

Figure 8: Full sample path of an individual (**left**) and distribution of the event times (**left**) for the partially observed SDE experiment. The surge in events at the terminal time indicates terminal censorship.

Event definition We consider the stochastic differential equation

$$dw_t = -\omega(w_t - \mu)dt + \sum_{i=1}^{d} dx_t^{(i)} + \sigma dB_t,$$

where w_t is trajectory of each individual with $(\sigma, \mu, \omega) \in \mathbb{R}^3$ are fixed parameters. In our experiment, the parameters are chosen to be $\sigma = 1$, $\mu = 0.1$ and $\omega = 0.1$. We then define the time-of-event as the time when trajectory cross the threshold $w_\star \in \mathbb{R}$ during the observation period $[t_0\ t_N]$, which is

$$T^\star = \min\{t_0 \leq t \leq t_N \,|\, w_t \geq w_\star\}.$$

In our experiments, we use the threshold value $w_\star = 2.5$. The target SDE is simulated using an Euler discretization. We train on $n = 500$ individuals.

Censorship We censor individuals whose trajectory does not cross the threshold during the observation period. This means that individuals are never censored during the observation period, but only at the end. The simulated censoring level is 3.2%.

Supplementary Figures. Figure 8 provides an example of the full sample path of an individual and the distribution of the event times of the whole population. We add additional results on the test set in Figures 9, 10, 11, 12 and 13.

D.2. Tumor Growth

Time series. Similarly to the partially observed SDE experiment described above, we set $d = 2$ which includes 1-dimensional sample path x_t of a fractional Brownian motion with Hurst parameter $H = 0.6$. The paths are sampled at 1000 times over the time interval $[0, 10]$. All simulations are done using the `stochastic` package. The time series $\mathbf{X}^i$ are identical, up to observation time, to the ones used for simulations.

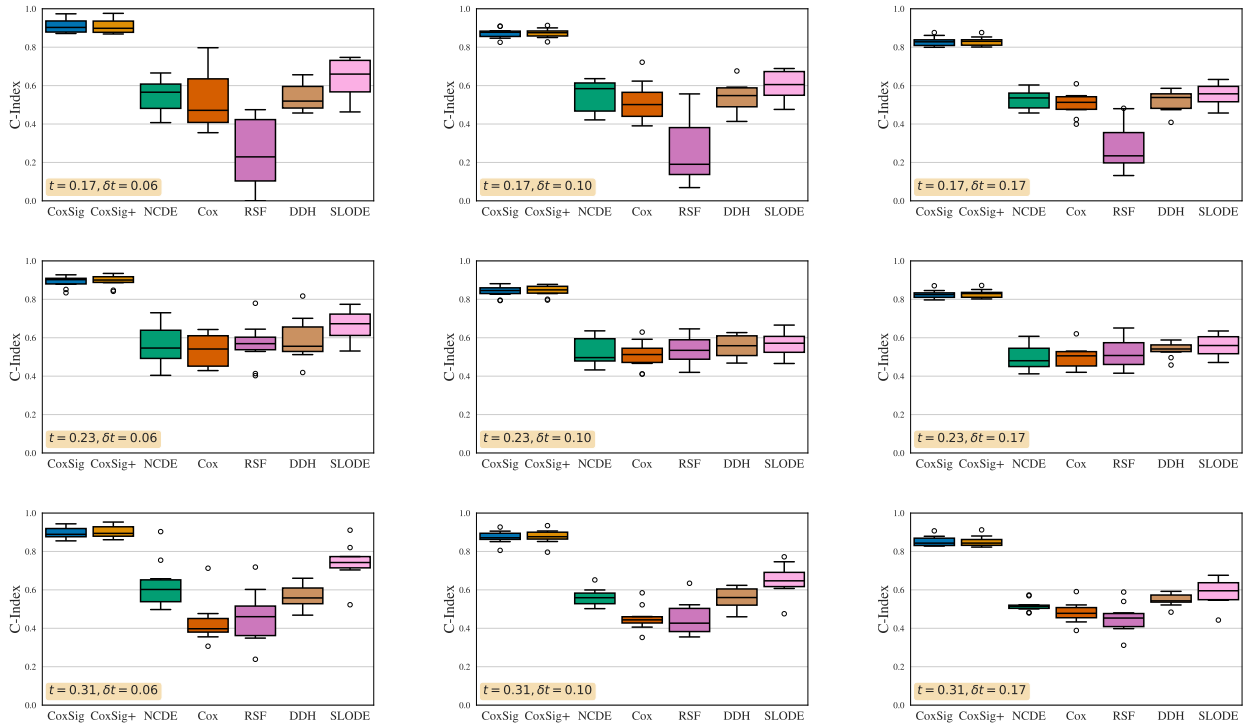

Figure 9: C-Index (*higher* is better) for **hitting time of a partially observed SDE** for numerous points $(t, \delta t)$.

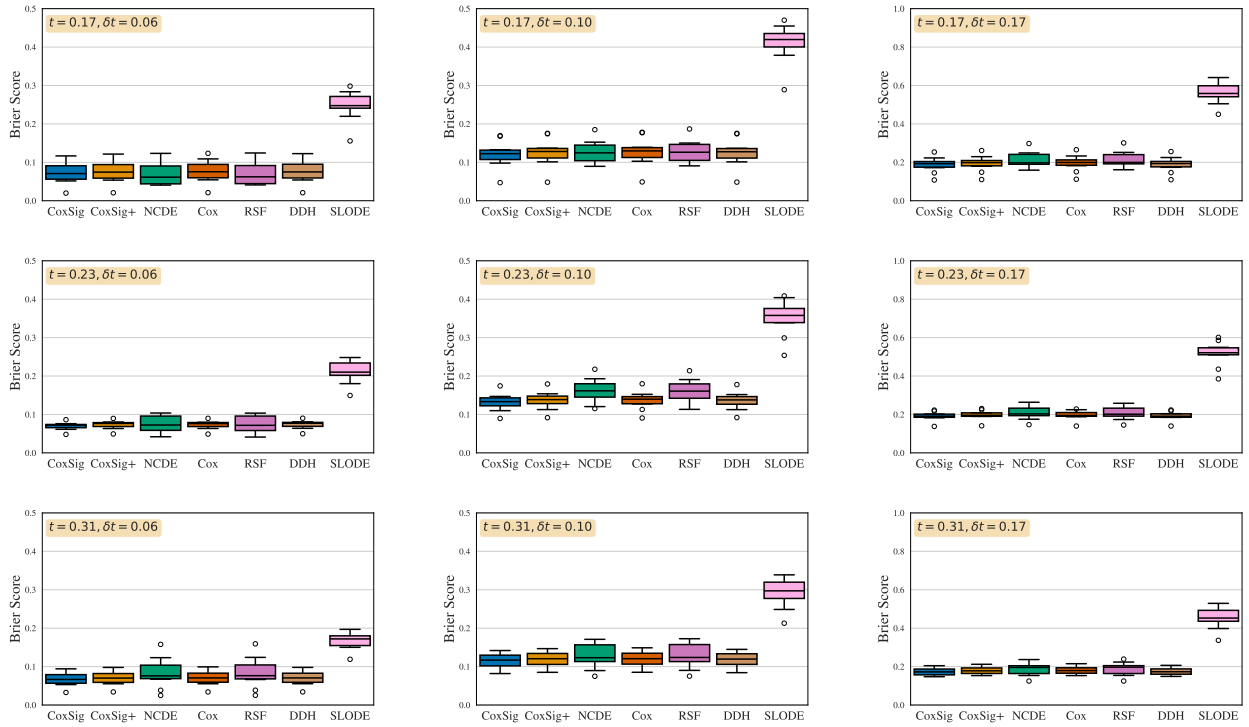

Figure 10: Brier score (*lower* is better) for **hitting time of a partially observed SDE** for numerous points $(t, \delta t)$.

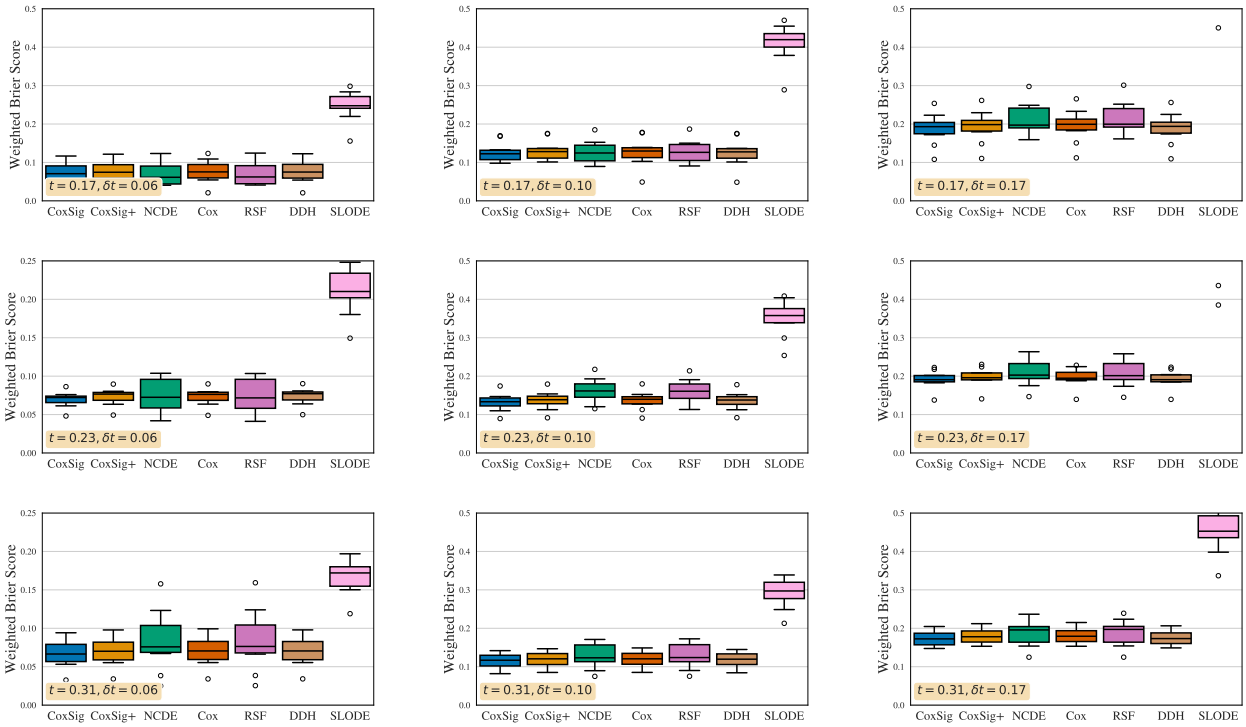

Figure 11: Weighted Brier score (*lower* is better) for **hitting time of a partially observed SDE** for numerous points $(t, \delta t)$.

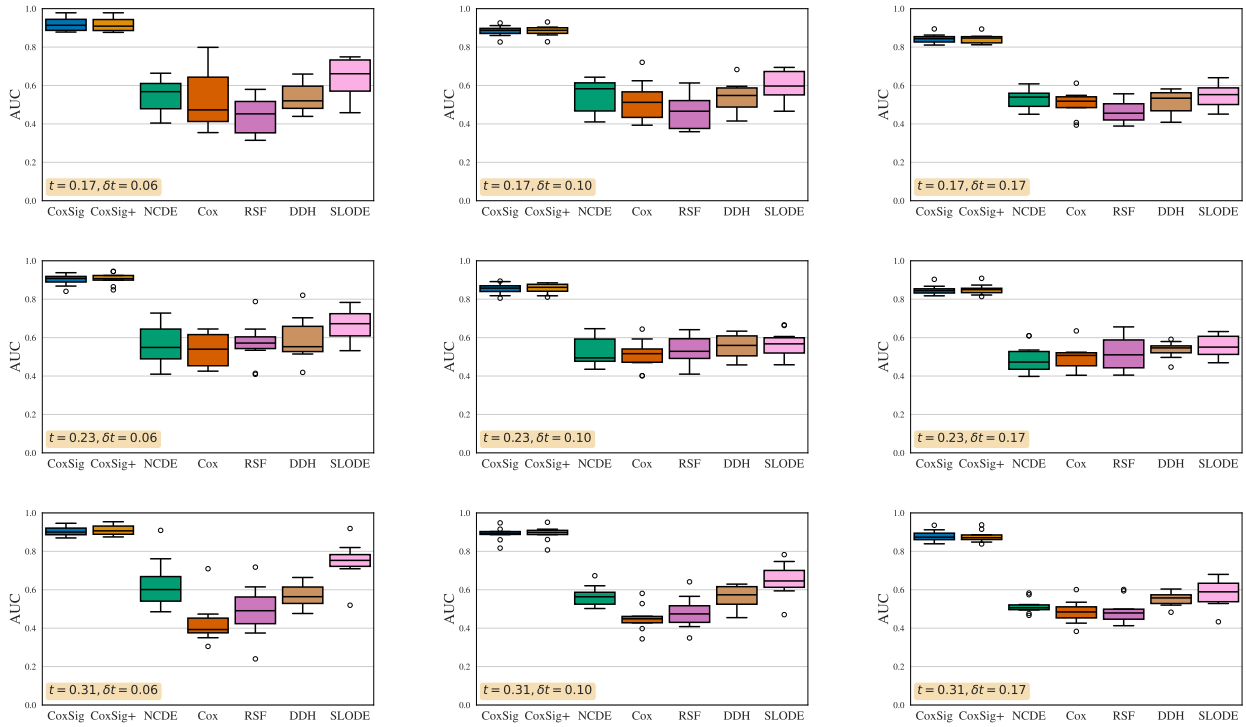

Figure 12: AUC (*higher* is better) for **hitting time of a partially observed SDE** for numerous points $(t, \delta t)$.

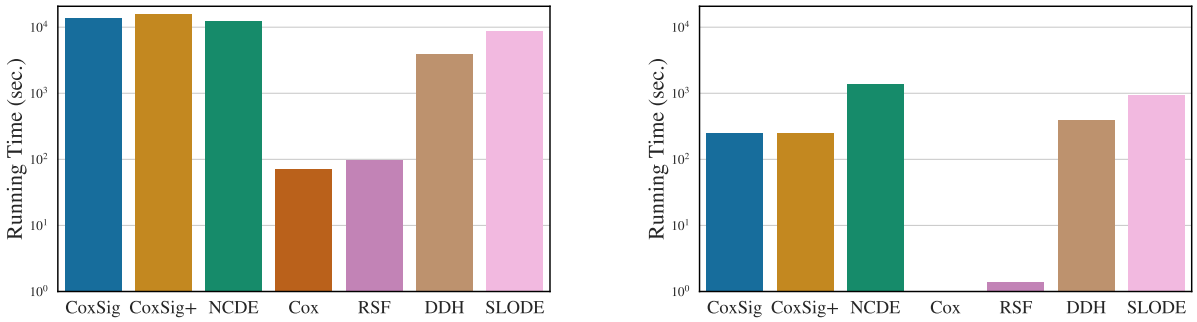

Figure 13: Running times on the partially observed SDE experiment (log-scale) averaged over 10 runs including cross-validation of the hyperparameters on CoxSig, CoxSig+, Cox and RSF (**left**) and over 1 run without cross-validation of the hyperparameters on CoxSig, CoxSig+, Cox and RSF (**right**).

Event definition. Following Simeoni et al. (2004), we consider the differential equations

$$\begin{aligned}
\frac{du_t^{(1)}}{dt} &= \frac{\lambda_0 u_t^{(1)}}{\left[1 + (\frac{\lambda_0}{\lambda_1} w_t)^{\Psi}\right]^{1/\Psi}} - \kappa_2 x_t u_t^{(1)} \\
\frac{du_t^{(2)}}{dt} &= \kappa_2 x_t u_t^{(1)} - \kappa_1 u_t^{(2)} \\
\frac{du_t^{(3)}}{dt} &= \kappa_1 (u_t^{(2)} - u_t^{(3)}) \\
\frac{du_t^{(4)}}{dt} &= \kappa_1 (u_t^{(3)} - u_t^{(4)}) \\
w_t &= u_t^{(1)} + u_t^{(2)} + u_t^{(3)} + u_t^{(4)},
\end{aligned}$$

where w_t is trajectory of each individual with initial status of $(u_0^{(1)}, u_0^{(2)}, u_0^{(3)}, u_0^{(4)}) = (0.8, 0, 0, 0)$ and $(\lambda_0, \lambda_1, \kappa_1, \kappa_2, \Psi) \in \mathbb{R}^5$ are fixed parameters. In our experiment, the parameters are chosen to be $\lambda_0 = 0.9$, $\lambda_1 = 0.7$, $\kappa_1 = 10$, $\kappa_2 = 0.15$ and $\Psi = 20$. We then define the time-of-event as the time when trajectory cross the threshold $w_\star \in \mathbb{R}$ during the observation period $[t_0 \; t_N]$, which is

$$T^\star = \min\{t_0 \leq t \leq t_N \,|\, w_t \geq w_\star\}.$$

In our experiments, we use the threshold value $w_\star = 1.7$. The target differential equations are simulated using an Euler discretization. We train on $n = 500$ individuals.

Censorship. Similarly to the partially observed SDE experiment, we consider terminal censorship: individuals that do not experience the event within the observation period are censored. The censoring level is 8.4%.

Supplementary Figures. Figure 14 provides an example of the full sample path of an individual and the distribution of the event times of the whole population. We add additional results on the test set in Figures 15, 16, 17 and 18.

D.3. Predictive Maintenance

Time series. This dataset describes the degradation of 200 aircraft gas turbine engines, where 22 measurements of sensors and 3 operational settings are recorded each operational cycle until its failure. After removing low-variance features, 16 longitudinal features are selected for training models. The average time length of these features is about 25 cycles. Note that we apply standardization for selected features before training.

Event definition. The times of event are given as-is in the dataset. We refer to Saxena et al. (2008) for a precise description of the data generation.

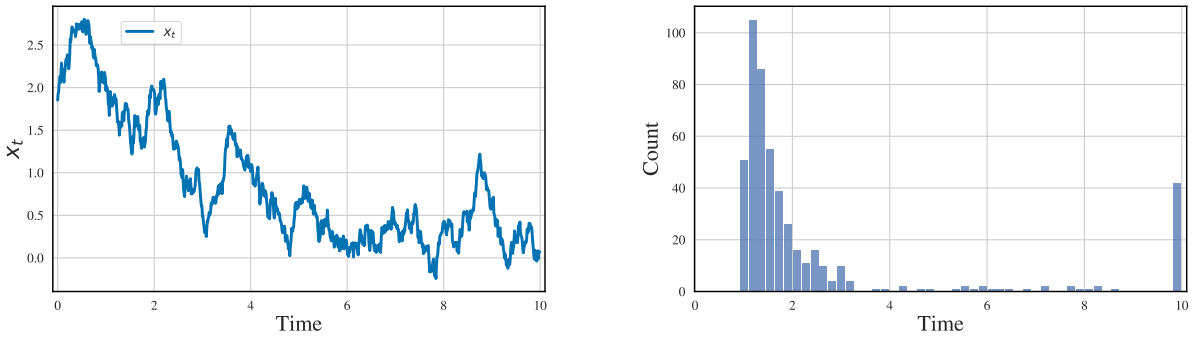

Figure 14: Full sample path of an individual (**left**) and distribution of the event times (**left**) for the tumor growth experiment. The surge in events at the terminal time indicates terminal censorship.

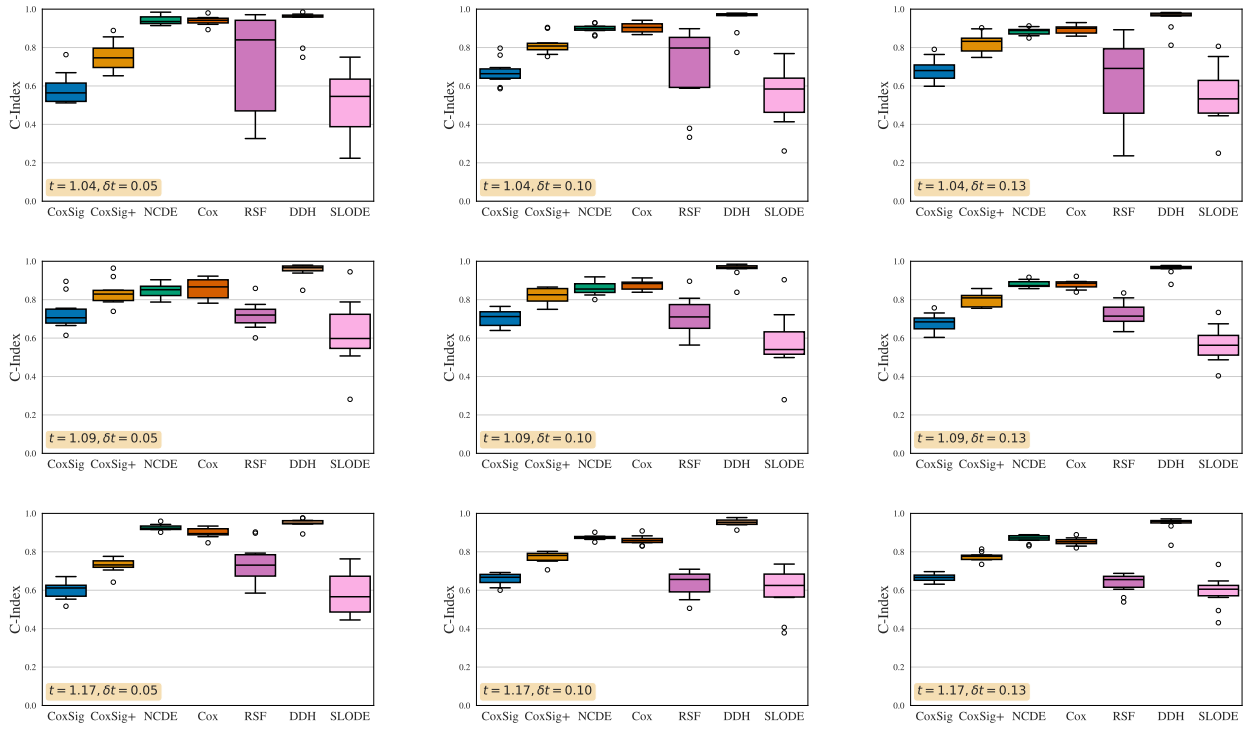

Figure 15: C-Index (*higher* is better) for **Tumor Growth** for numerous points $(t, \delta t)$.

Censorship. Censorship is given as-in in the dataset. The censoring level of this dataset is 50%, which is a high censorship rate in survival analysis. We refer again to Saxena et al. (2008) for more details.

Supplementary Figures. Figure 19 provides an example of several randomly picked sample paths of an individual and the distribution of the event times of the whole population. We add additional results in Figures 20, 21, 22 and 23.

D.4. Churn Prediction

For this dataset, the amount of details that we can release is limited both because of the sensitive nature of the data and of the anonymity requirements of the reviewing process.

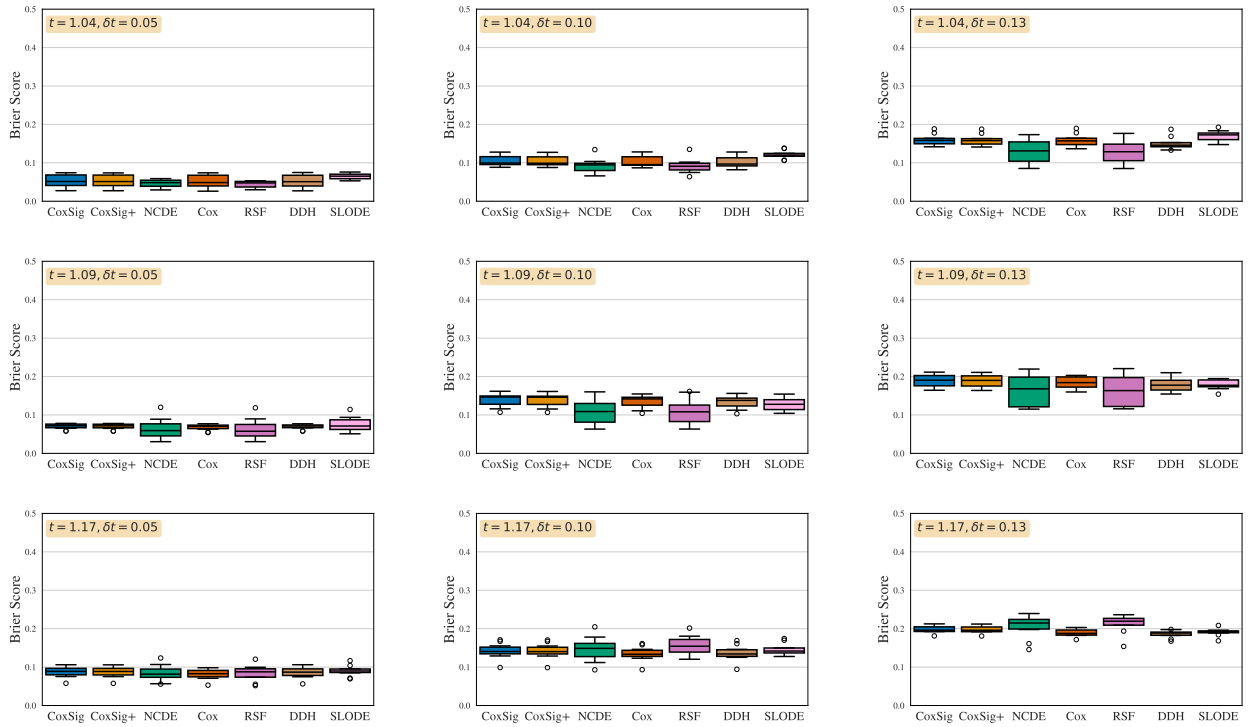

Figure 16: Brier score (*lower* is better) for **Tumor Growth** for numerous points $(t, \delta t)$.

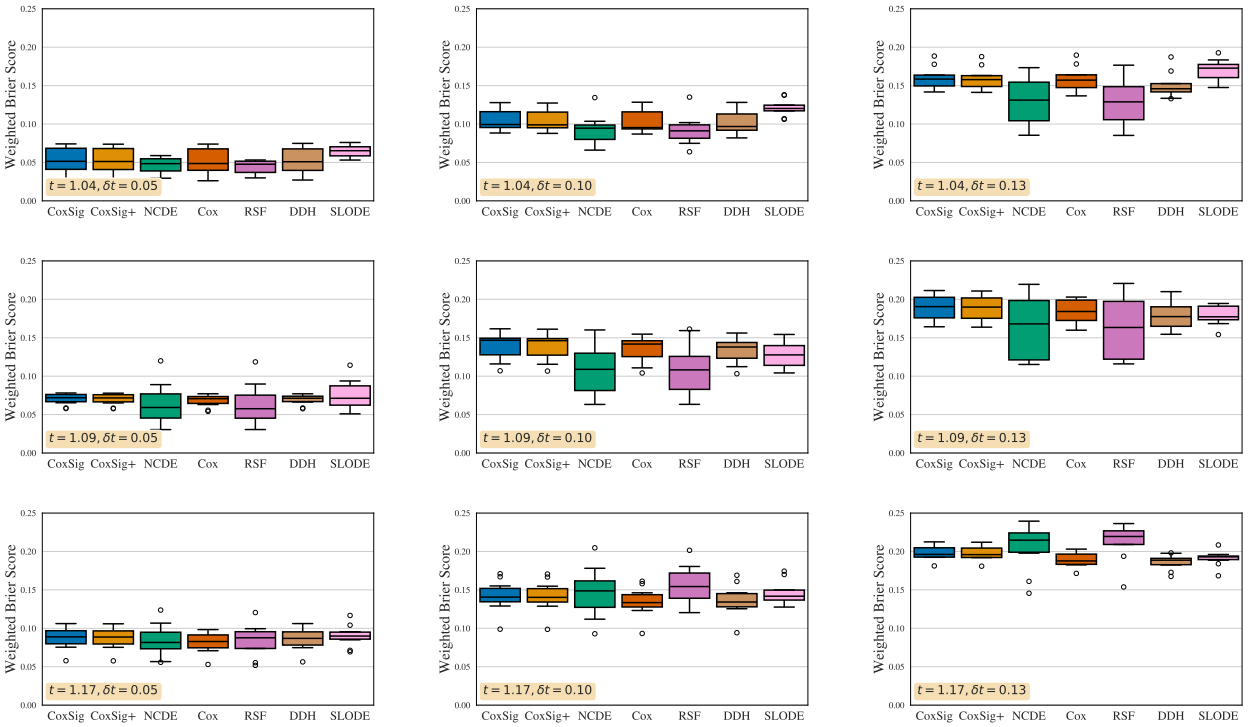

Figure 17: Weighted Brier score (*lower* is better) for **Tumor Growth** for numerous points $(t, \delta t)$.

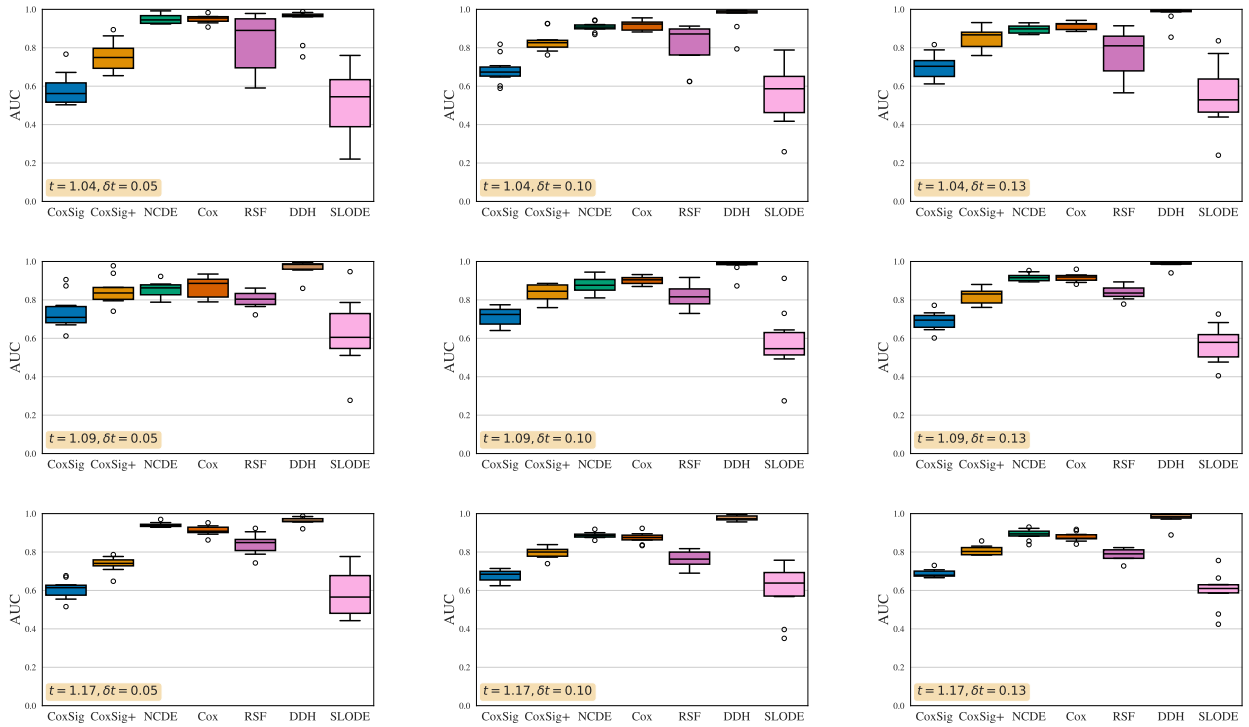

Figure 18: AUC (*higher* is better) for **Tumor Growth** for numerous points $(t, \delta t)$.

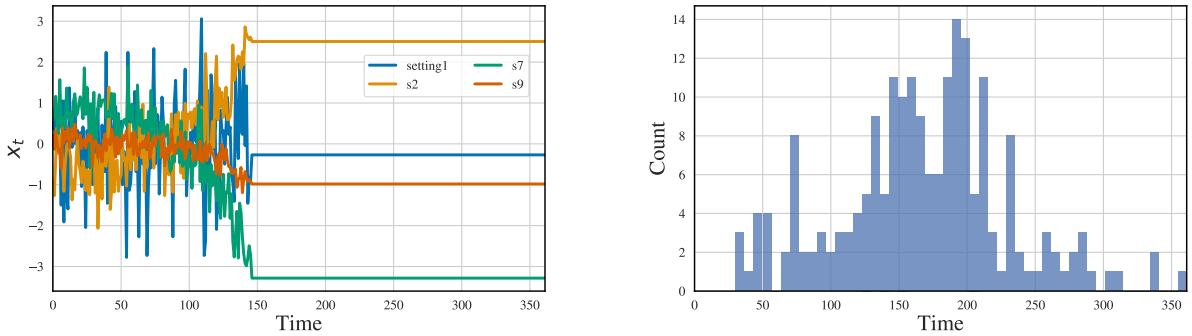

Figure 19: Partial sample path of an individual (**left**) and distribution of the event times (**left**) for the predictive maintenance experiment. On the left, the time series is filled with the last observed value from the time of the event on.

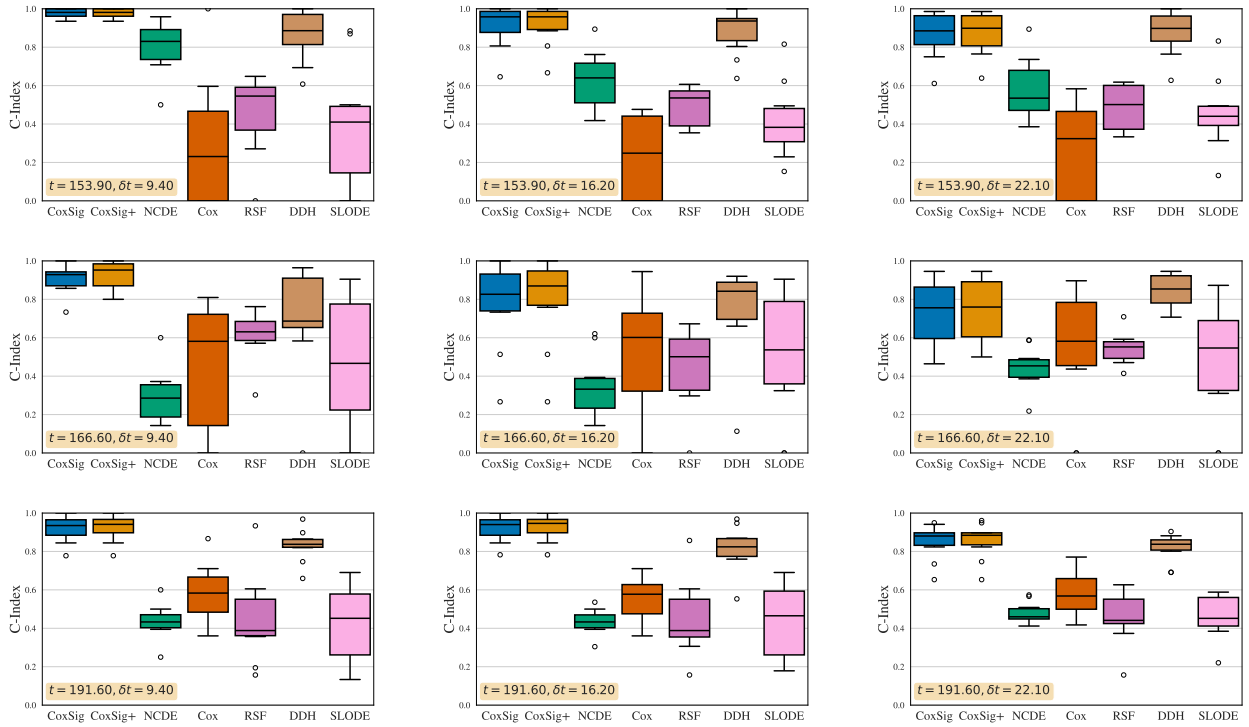

Figure 20: C-Index (*higher* is better) for **predictive maintenance** for numerous points $(t, \delta t)$.

Time series. All longitudinal features have been computed on a temporal window of one week, the raw data corresponding to all product orders placed on the platform from 06-12-2021 to 12-11-2023. For clients who have no order during the week, we fill zero value for all longitudinal measurements this week. After removing features with more than 90 % of missingness, 14 longitudinal features of 1043 clients are selected for the training step. Note that we apply standardization for selected features before training.

Event definition. We consider that a customer has churned if she has no passed any order in the last 4 weeks. If the customer starts ordering again after a churn, we register her as a new customer.

Censorship. Censorship is terminal based on the data collection period (give dates here). Hence any customer that has not churned by 12-11-2023 is censored. In this dataset, 38.4% of the clients are terminally censored.

Supplementary Figures. Figure 24 provides an example of four sample paths of four randomly chosen individuals. We add additional results in Figures 25, 26, 27 and 28.

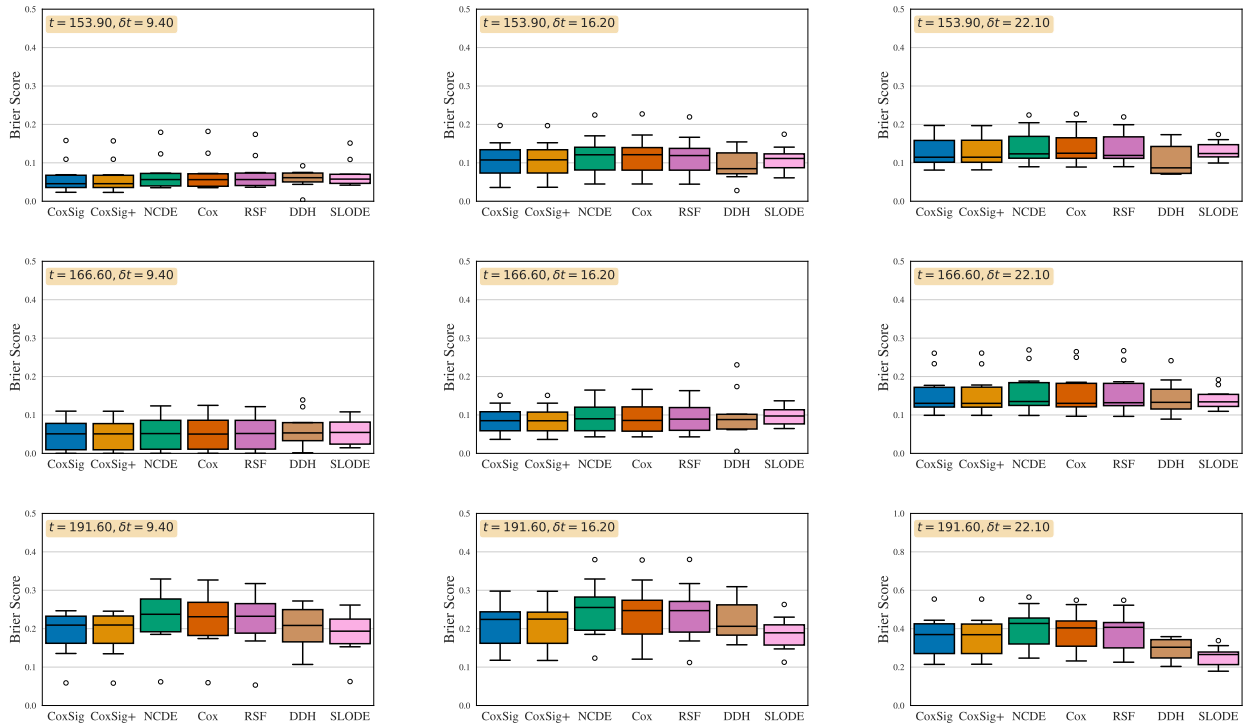

Figure 21: Brier Score (*lower* is better) for **predictive maintenance** for numerous points $(t, \delta t)$.

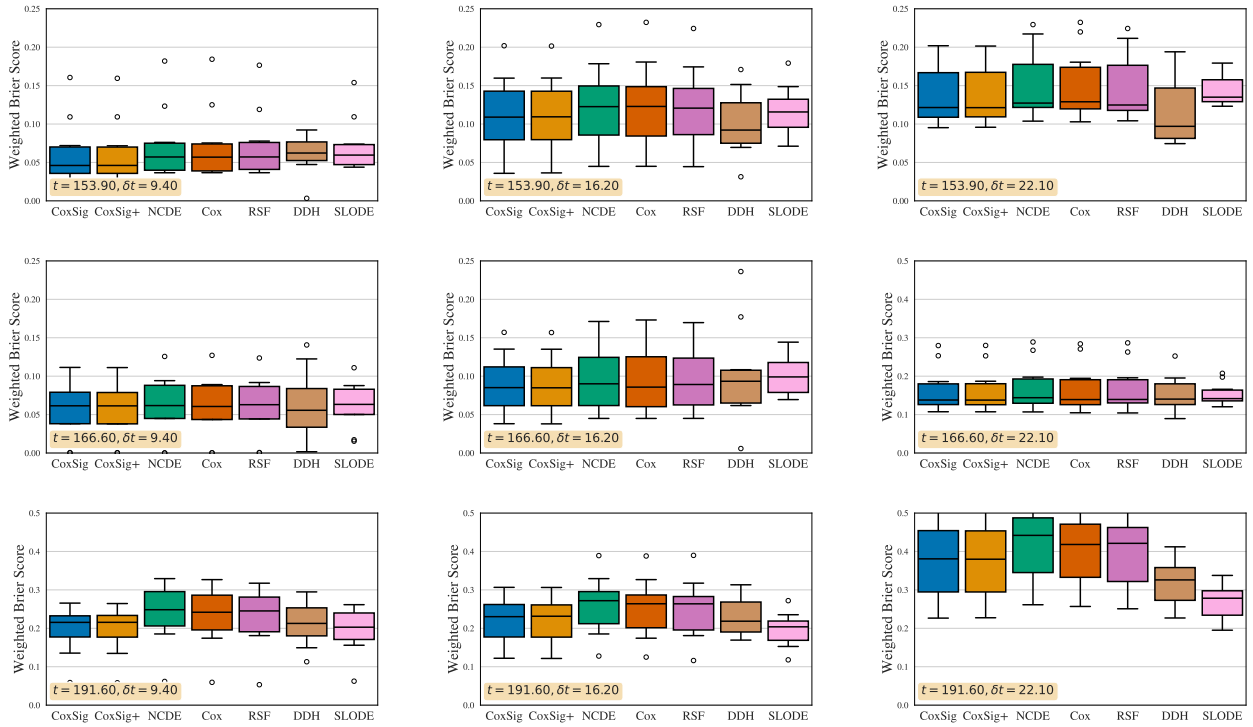

Figure 22: Weighted Brier Score (*lower* is better) for **predictive maintenance** for numerous points $(t, \delta t)$.

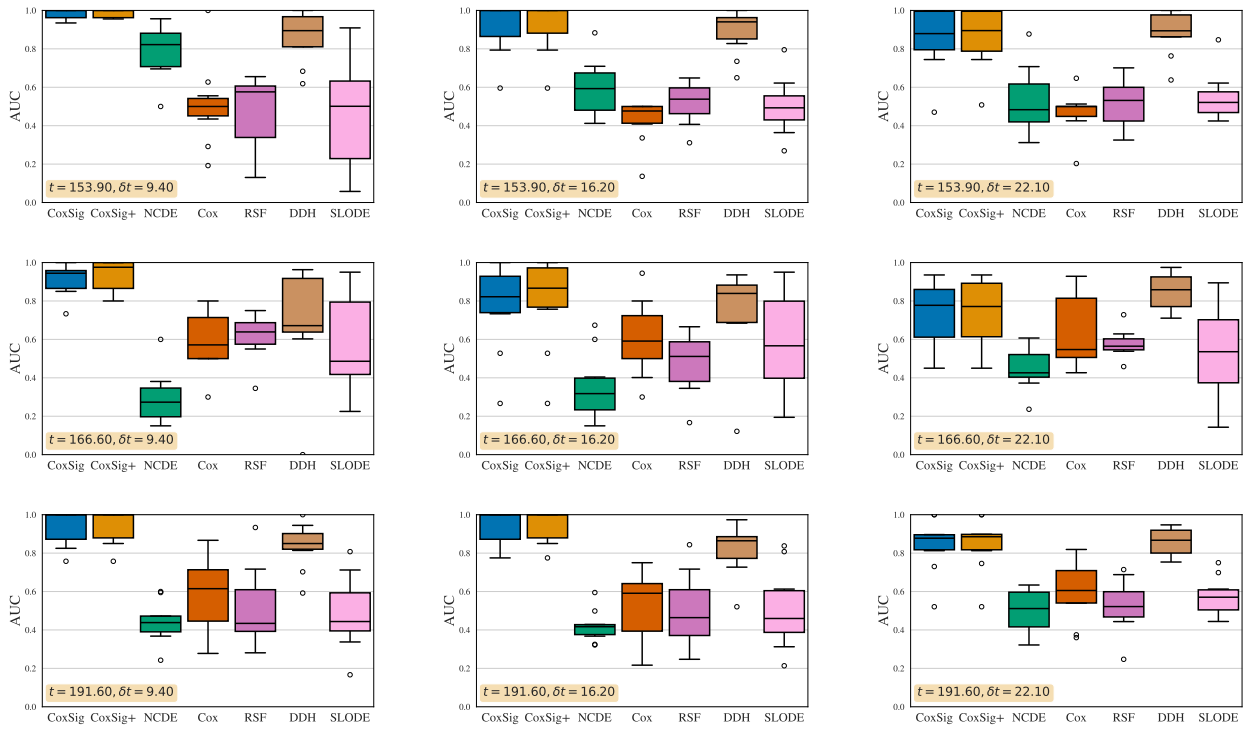

Figure 23: AUC (*higher* is better) for **predictive maintenance** for numerous points $(t, \delta t)$.

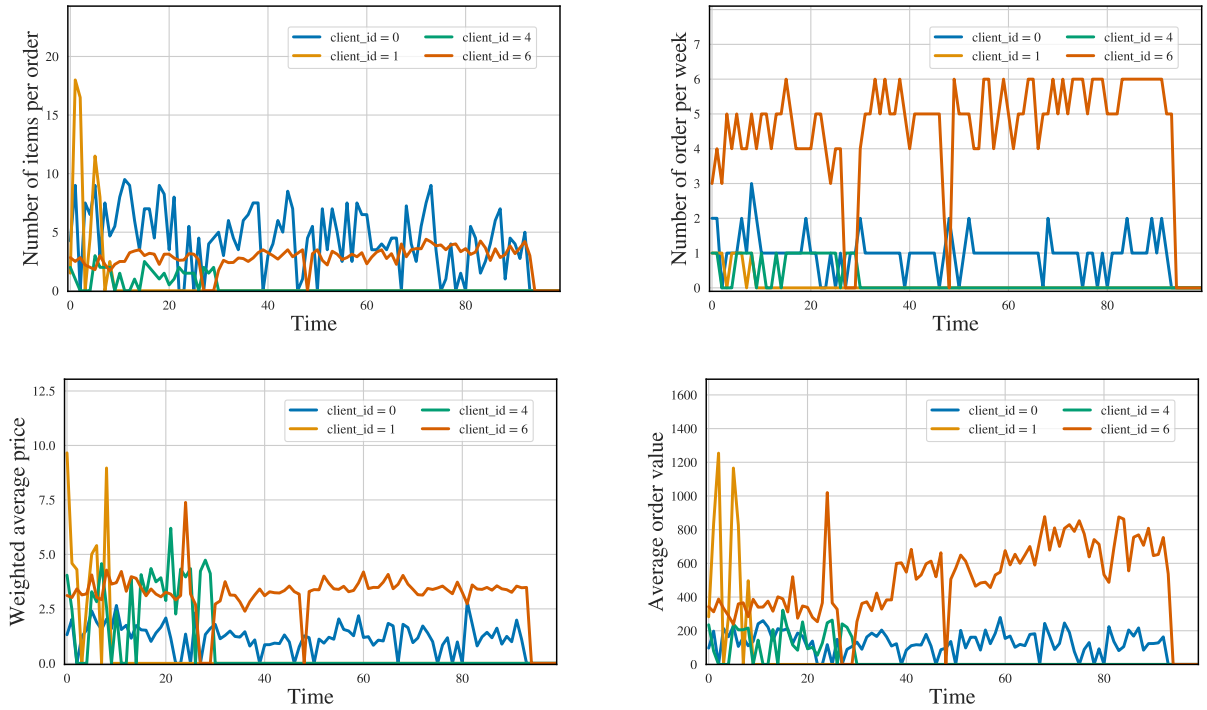

Figure 24: Values of 4 different time-dependent features for 4 randomly chosen individuals from the **churn prediction** dataset. Individual time-to-event and distribution of the event times cannot be displayed to protect consumer and business privacy. A precise description of the different time-dependent features will be provided upon publication.

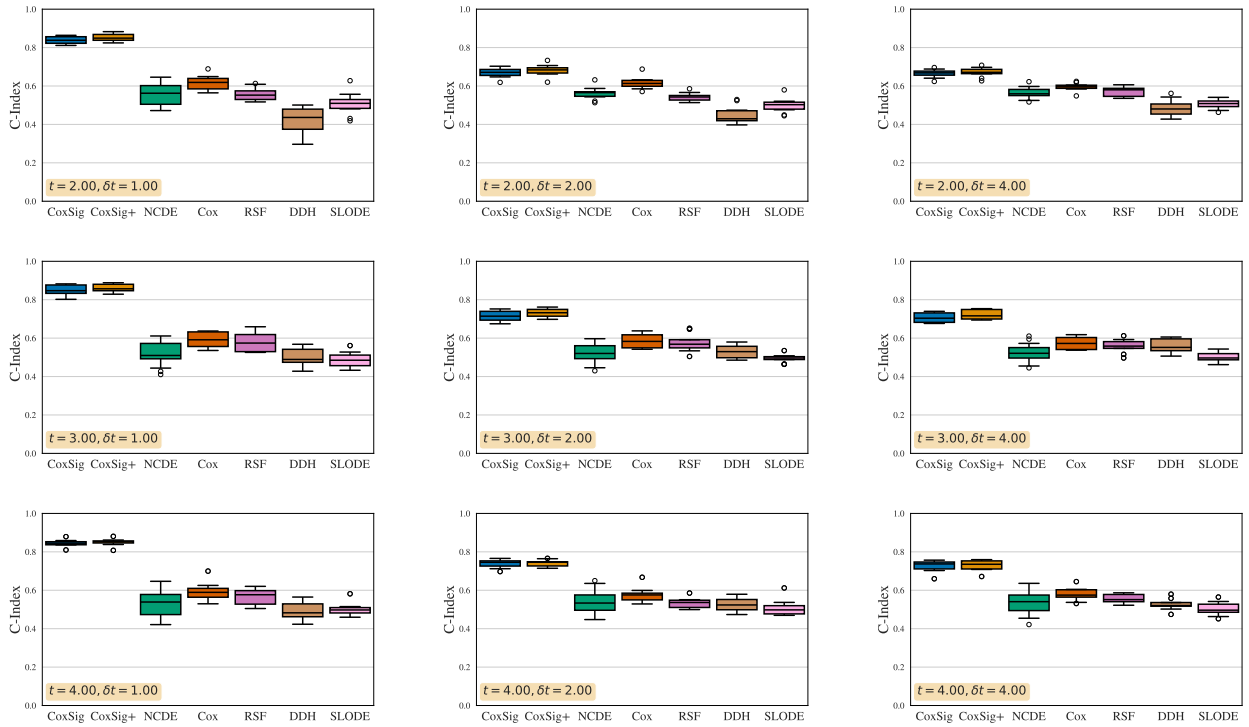

Figure 25: C-Index (*higher* is better) for **churn prediction** for numerous points $(t, \delta t)$.

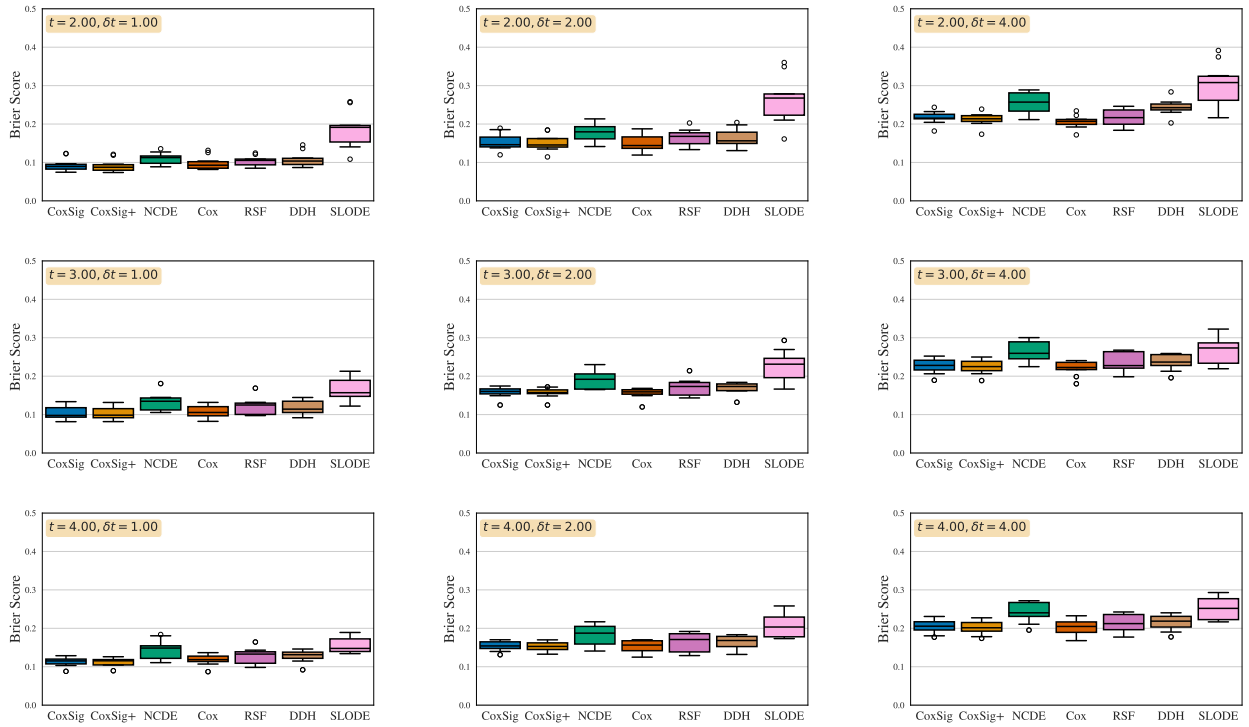

Figure 26: Brier score (*lower* is better) for **churn prediction** for numerous points $(t, \delta t)$.

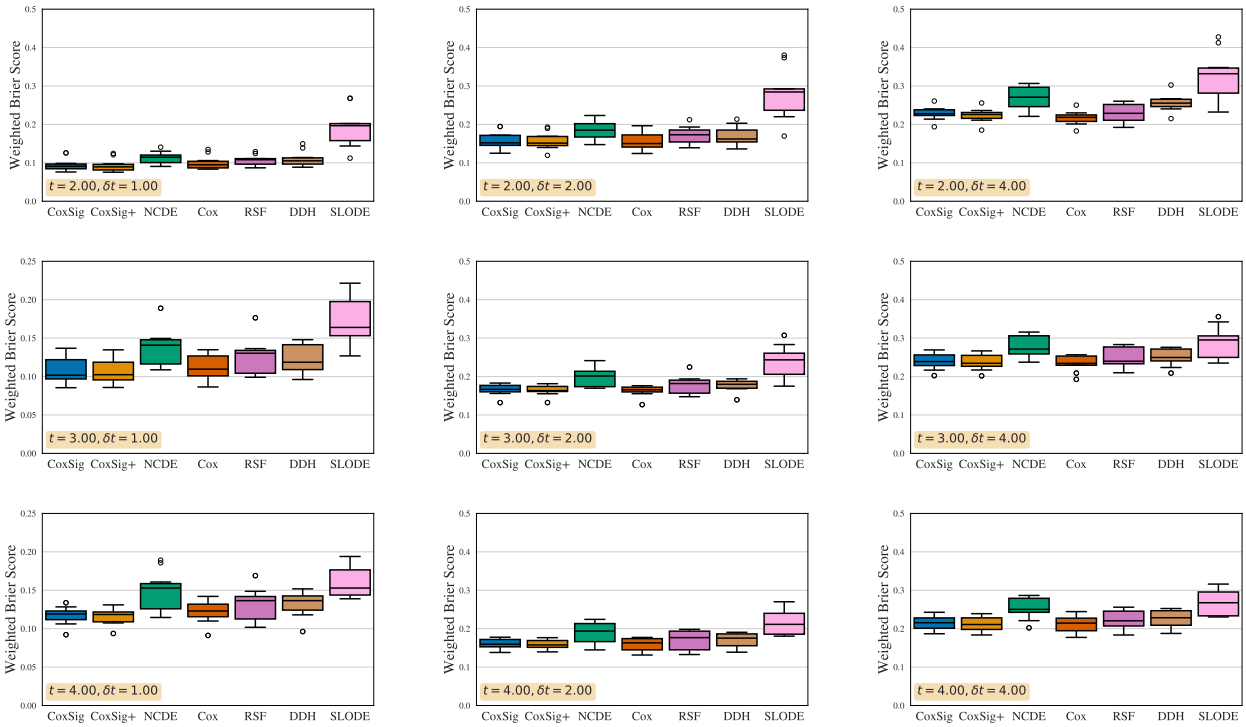

Figure 27: Weighted Brier score (*lower* is better) for **churn prediction** for numerous points $(t, \delta t)$.

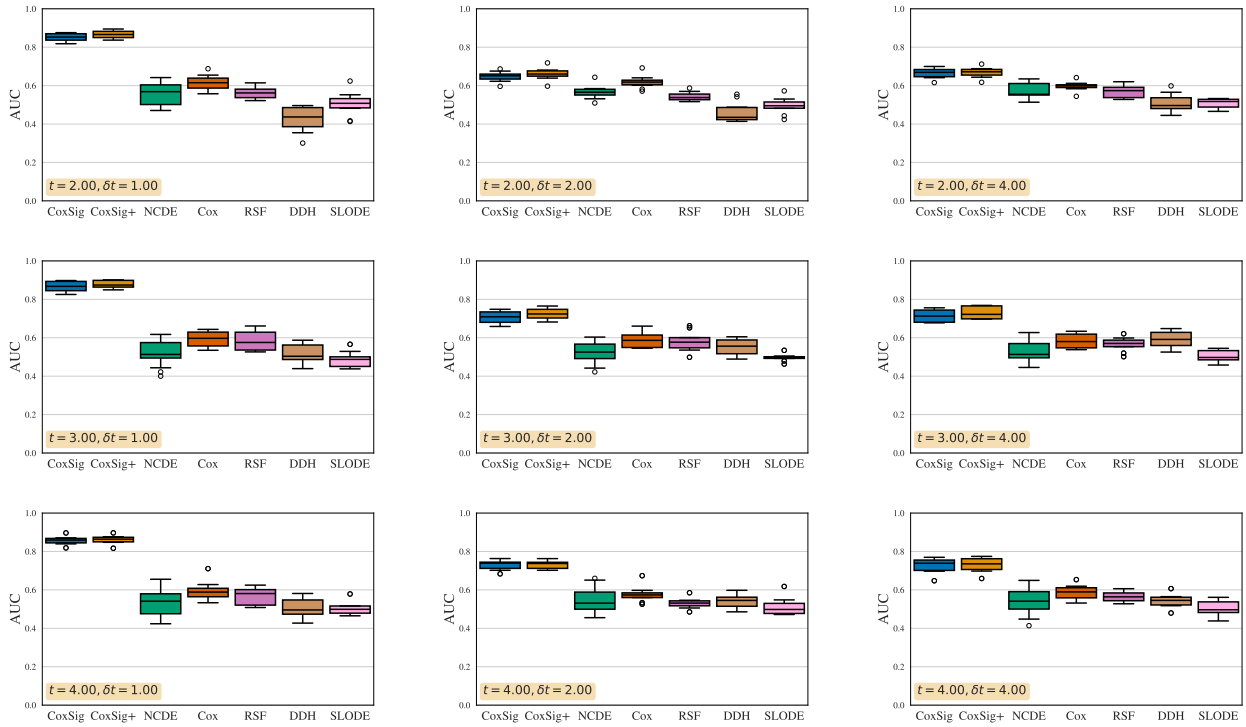

Figure 28: AUC (*higher* is better) for **churn prediction** for numerous points $(t, \delta t)$.