Dynamic Time Lag Regression: Predicting What and When
Mandar Chandorkar, Cyril Furtlehner, Bala Poduval, Enrico Camporeale, Michèle Sebag

To cite this version:

Mandar Chandorkar, Cyril Furtlehner, Bala Poduval, Enrico Camporeale, Michèle Sebag. Dynamic Time Lag Regression: Predicting What and When. ICLR 2020 - 8th International Conference on Learning Representations, Apr 2020, Addis Abeba, Ethiopia. hal-02422148

HAL Id: hal-02422148
https://inria.hal.science/hal-02422148
Submitted on 20 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
DYNAMIC TIME LAG REGRESSION: PREDICTING WHAT & WHEN

Mandar Chandorkar
Centrum Wiskunde en Informatica
Amsterdam 1098XG
Cyril Furtlehner
INRIA-Saclay
Bala Poduval
University of New Hampshire
Durham, NH 03824
Enrico Camporeale
CIRES, University of Colorado
Boulder, CO
Michèle Sebag
CNRS – Univ. Paris-Saclay

ABSTRACT

This paper tackles a new regression problem, called Dynamic Time-Lag Regression (DTLR), where a cause signal drives an effect signal with an unknown time delay. The motivating application, pertaining to space weather modelling, aims to predict the near-Earth solar wind speed based on estimates of the Sun’s coronal magnetic field. DTLR differs from mainstream regression and from sequence-to-sequence learning in two respects: firstly, no ground truth (e.g., pairs of associated sub-sequences) is available; secondly, the cause signal contains much information irrelevant to the effect signal (the solar magnetic field governs the solar wind propagation in the heliosphere, of which the Earth’s magnetosphere is but a minuscule region).

A Bayesian approach is presented to tackle the specifics of the DTLR problem, with theoretical justifications based on linear stability analysis. A proof of concept on synthetic problems is presented. Finally, the empirical results on the solar wind modelling task improve on the state of the art in solar wind forecasting.

1 INTRODUCTION

A significant body of work in machine learning concerns the modeling of spatio-temporal phenomena (Shi and Yeung, 2018; Rangapuram et al., 2018), including the causal analysis of time series (Peters et al., 2017), with applications ranging from markets (Pennacchioli et al., 2014) to bioinformatics (Brouard et al., 2016) to climate (Nooteboom et al., 2018).

This paper focuses on the problem of modeling the temporal dependency between two spatio-temporal phenomena, where the latter one is caused by the former one (Granger, 1969; Runge, 2018) with a non-stationary time delay.

The motivating application domain is that of space weather. The sun, a perennial source of charged energetic particles, is at the origin of geomagnetic phenomena within the sun-earth system. Specifically, the sun ejects charged particles into the surrounding space in all directions and some of these particle clouds, a.k.a. solar wind, reach the Earth’s vicinity. High speed solar wind is a major threat for the modern world, causing severe damages to e.g., satellites, telecommunication infrastructures, under sea pipelines, among others.

A key prediction task thus is to forecast the speed of the solar wind in the vicinity of the Earth (Munteanu et al., 2013; Haaland et al., 2010; Reiss et al., 2019), sufficiently early to emit an alarm and be able to prevent the damage to the best possible extent. Formally the goal is to model the dependency between heliospheric observations (available at light speed), referred to as cause series, and the solar wind speed series recorded at the Lagrangian point $L_1$ (a point on the Sun-Earth line 1.5 million kilometers away from the Earth), referred to as effect series. The key difficulty is that the adverse impact of space weather is estimated to cost 200 to 400 million USD per year, but can sporadically lead to much larger losses.
time lag between an input and its effect, the solar wind recorded at \( L_1 \), varies from circa 2 to 5 days depending on, among many factors, the initial direction of emitted particles and their energy. Would the lag be constant, the solar wind prediction problem would boil down to a mainstream regression problem. The challenge here is to predict, from the solar image \( x(t) \) at time \( t \) the value \( y(t + \Delta t) \) of the solar wind speed reaching the earth at time \( t + \Delta t \) where both the value \( y(t + \Delta t) \) and the time lag \( \Delta t \) depend on \( x(t) \).

**Related work.** Indeed, the modeling of dependencies among time series has been intensively tackled (see e.g., Zhou and Sornette (2006); Runge (2018)). When considering varying time lag, many approaches rely on dynamic time warping (DTW) (Sakoe and Chiba, 1978). For instance, DTW is used in Gaskell et al. (2015), taking a Bayesian approach to achieve the temporal alignment of both series under some restricting assumptions (considering slowly varying time lags and linear relationships between the cause and effect time series). More generally, the use of DTW in time series analysis relies on simplifying assumptions on the cause and effect series (same dimensionality and structure) and builds upon available cost matrices for the temporal alignment.

Also related is sequence-to-sequence learning (Sutskever et al., 2014), primarily aimed to machine translation. While Seq2Seq modelling relaxes some former assumptions (such as the fixed or comparable sizes of the source and target series), it still relies on the known segmentation of the source series into disjoint units (the sentences), each one being mapped into a large fixed-size vector using an LSTM; and this vector is exploited by another LSTM to extract the output sequence. Attention-based mechanisms (Graves (2013), Bahdanau et al. (2015) alleviate the need to encode the full source sentence into a fixed-size vector, by learning the alignment and allowing the model to search for the parts of the source sentence relevant to predict a target part. More advanced attention mechanisms (Kim et al. 2017; Vaswani et al. 2017) refine the way the source information is leveraged to produce a target part. But to our best knowledge, the end-to-end learning of the sequence-to-sequence modelling relies on the segmentation of the source and target series, and the definition of associated pairs of segments (e.g. the sentences).

Our claim is that the regression problem of predicting both what the effect is and when the effect is observed, called Dynamic TimeLag Regression (DTLR), constitutes a new ML problem: With respect to the modeling of dependencies among time series, it involves stochastic dependencies of arbitrary complexity; the relationship between the cause and the effect series can be non-linear (the what model). Furthermore, the time lag phenomenon (the when model) can be non smooth (as opposed to e.g. Zhou and Sornette (2006)).

With respect to sequence-to-sequence translation, a main difference is that the end-to-end training of the model cannot rely on pairs of associated units (the sentences), adversely affecting the alignment learning.

Lastly, and most importantly, in the considered DTLR problem, even if the cause series has high information content, only a small portion of it is relevant to the prediction of the effect series. On one hand, the cause series might be high dimensional (images) whereas the effect series is scalar; on the other hand, the cause series governs the solar wind speed in the whole heliosphere and not just in near-Earth space. In addition to avoiding typically one or two orders of magnitude expansion of an already large input signal dimension, inserting the time-lag inference explicitly in the model can also potentially improve its interpretability.

**Organization of the paper.** The Bayesian approach proposed to tackle the specifics of the DTLR regression problem is described in section 2, the associated learning equations are discussed, followed by a stability analysis and a proof of consistency (section 3). The algorithm is detailed in section 4. The experimental setting used to validate the approach is presented in section 5; the empirical validation on toy problems and on the real-world problem are discussed in section 6.

# 2 Probabilistic Dynamically Delayed Regression

## 2.1 Position of the Problem

Given two time series, the cause series \( \mathbf{x}(t) \) \( \in \mathcal{X} \subset \mathbb{R}^D \) and the observed effect series \( y(t) \), the sought model consists of a mapping \( f(\cdot) \) which maps each input pattern \( x(t) \) to an output \( y(\phi(t)) \), and a mapping \( g(\cdot) \) which determines the time delay \( \phi(t) - t \) between the input and output patterns:
\[ y(\phi(t)) = f[x(t)] \]
\[ \phi(t) = t + g[x(t)] \] (1)
(2)

with \( f : \mathcal{X} \to \mathbb{R} \), and \( g : \mathcal{X} \to \mathbb{R}^+ \),

where \( t \in \mathbb{R}^+ \) represents the continuous temporal domain. The input signal \( x(t) \) is possibly high dimensional and contains the hidden cause to the effect \( y(t) \in \mathbb{R}; y(t) \) is assumed to be scalar in the remainder of the paper. Function \( g : \mathcal{X} \to \mathbb{R}^+ \) represents the time delay between inputs and outputs. Vectors are written using bold fonts.

As said, Eqs 1-2 define a regression problem that differs from standard regression in two ways: Firstly, vectors are written using bold fonts. Secondly, \( g[x(t)] \) is non-stationary as it depends on \( x(t) \). Secondly, \( g[x(t)] \) is unknown, i.e. it is not recorded explicitly in the training data.

**Assumption.** For the sake of the model identifiability and computational stability, the time warping function \( \phi(t) = t + g[x(t)] \) is assumed to be sufficiently regular w.r.t. \( t \). Formally, \( \phi(.) \) is assumed to be continuous.

### 2.2 Probabilistic Dynamic Time-Lag regression

For practical reasons, cause and effect series are sampled at constant rate, and thereafter noted \( x_t \) and \( y_t \), with \( t \in \mathcal{T} \). Accordingly, mapping \( g \) maps \( x_t \) onto a finite set \( \mathcal{T} \) of possible time lags, with \( \mathcal{T} = \{\Delta t_{\text{min}}, \ldots, \Delta t_{\text{max}}\} \) an integer interval defined from domain knowledge. The unavoidable error due to the discretization of the continuous time lag is mitigated along a probabilistic model, associating to each cause \( x_t \), a set of predictors \( \hat{y}_t(x_t) \) and a probability distribution \( \hat{p}(x_t) \) on \( \mathcal{T} \) estimating the probability of delay of the effects of \( x_t \). Overall, the DTLR solution is sought as a probability distribution conditioned on cause \( x_t \), mixture of Gaussians centered on the predictors \( \hat{y}_t(x_t) \), where the mixture weights are defined from \( \hat{p}(x_t) \). More formally, letting \( y_t \) denote the vector of random variables \( \{y_{t+i}, i \in \mathcal{T}\} \):

\[
P[y_t|x_t = x] = \sum_{\tau_i \in \{0,1\}, i \in \mathcal{T}} \hat{p}(\tau_1, \ldots, \tau_{|\mathcal{T}|}|x) \mathcal{N}(\hat{y}(x), \Sigma(\tau))
\]

with \( \Sigma = \text{Diag}(\sigma_i(\tau)^2) \) the diagonal matrix of variance parameters attached to each time-lag \( i \in \mathcal{T} \).

Two simplifications are made for the sake of the analysis. Firstly, the stochastic time lag is modelled as the vector \( (\tau_i) \), \( i \in \mathcal{T} \) of binary latent variables, where \( \tau_i \) indicates whether \( x_t \) drives \( y_{t+i} \) \( (\tau_i = 1) \) or not \( (\tau_i = 0) \). The assumption that every cause has a single effect is modelled by imposing:

\[
\sum_{i \in \mathcal{T}} \tau_i = 1.
\]

From constraint (4), probability distribution \( \hat{p}(x_t) \) is thus sought as the vector \( \hat{p}_i(x_t) \) for \( i \in \mathcal{T} \), summing to 1, such that \( \hat{p}_i(x_t) \) stands for the probability of the effect of \( x_t = x \) to occur with delay \( i \).

The second simplifying assumption is that the variance \( \sigma_i^2(\tau) \) of predictor \( \hat{y}_i \) does not depend on \( x_t \), by setting:

\[
\sigma_i(\tau)^2 = \left( 1 + \sum_j \alpha_{ij} \tau_j \right) \sigma^2,
\]

with \( \sigma^2 \) a default variance and \( \alpha_{ij} \geq 0 \) a matrix of non-negative real parameters. This particular formulation supports the tractable analysis of the posterior probability of \( \tau_i \) in (supplementary

---

2 For some authors, Zhou and Sornette [2006], the monotonicity of \( \phi(.) \) is additionally required and enforced using constraints:

\[ \phi(t_1) \leq \phi(t_2), \forall t_1 \leq t_2 \]

This assumption is not retained as one may achieve a similar effect by using regularization based smoothness penalties.

3 Note that pre-processing can be used if needed to map non-Gaussian data onto Gaussian data.
material). The fact that \( x \) can influence \( y_i \) through predictor \( \hat{y}_i(x) \) even when \( \tau_i = 0 \) reflects an indirect influence due to the auto-correlation of the \( y \) series. This influence comes with a higher variance, enforced by making \( \alpha_{ij} \) a decreasing function of \( |i-j| \). More generally, a large value of \( \alpha_{ii} \) compared to \( \alpha_{ij} \) for \( i \neq j \) corresponds to a small auto-correlation time of the effect series.

2.3 Learning criterion

The joint distribution is classically learned by maximizing the log likelihood of the data, which can here be expressed in closed form. Let us denote respectively the dataset and parameters as \( \{ (x, y) \}_{data} \) and \( \theta = (\hat{y}, \hat{p}, \sigma, \alpha) \). From Eq. (3), the conditional probability \( q_i(x, y) \) reads:

\[
q_i(x, y) = \frac{1}{Z(x, y|\theta)} \hat{p}_i(x) \exp\left( -\frac{1}{2\sigma^2} \sum_{j \in T} \alpha_{ji} (y_j - \hat{y}_j(x))^2 + \frac{1}{2} \sum_{j \in T} \log(1 + \alpha_{ji}) \right)
\]

with normalization constant

\[
Z(x, y|\theta) = \sum_{i \in T} \hat{p}_i(x) \exp\left( -\frac{1}{2\sigma^2} \sum_{j \in T} \alpha_{ji} (y_j - \hat{y}_j(x))^2 + \frac{1}{2} \sum_{j \in T} \log(1 + \alpha_{ji}) \right).
\]

The log-likelihood then reads (intermediate calculations in supplementary material, appendix A):

\[
L\{\{ (x, y) \}_{data}|\theta\} = -|T| \log(\sigma) - \mathbb{E}_{data}\left[ \sum_{i \in T} \frac{1}{2\sigma^2} (y_i - \hat{y}_i(x))^2 - \log(Z(x, y|\theta)) \right]
\]

where \( \mathbb{E}_{data} \) denotes averaging over the dataset. For notational simplicity, the time index \( t \) is omitted in the following and the empirical averaging on the data is noted \( \mathbb{E}_{data} \). The hyper-parameters \( \sigma \) and matrix \( \alpha \) of the model are obtained by optimizing \( L \):

\[
\sigma^2 = \mathbb{E}_{data}\left[ (y_i - \hat{y}_i(x))^2 q_j(x, y) \right] / \mathbb{E}_{data}\left[ q_j(x, y) \right].
\]

In addition the optimal \( \hat{y} \) and \( \hat{p} \) reads:

\[
\hat{y}_i(x) = \mathbb{E}_{data} \left[ y_i \left( 1 + \sum_{j \in T} \alpha_{ij} q_j(x, y) \right) \right] / \mathbb{E}_{data} \left[ 1 + \sum_{j \in T} \alpha_{ij} q_j(x, y) \right],
\]

\[
\hat{p}_i(x) = \mathbb{E}_{data} \left[ q_i(x, y) \right].
\]

These are self-consistent equations, since \( q_i(x, y) \) depends on the parameters \( \sigma^2 \) and \( \alpha_{ij} \), \( \hat{y} \) and \( \hat{p} \).

The proposed algorithm detailed in section 4 implements the saddle point method defined from Eqs (7) and (8) alternatively, hyper-parameters \( \sigma \) and \( \alpha_{ij} \) are updated from Eq. (7) based on the current \( \hat{y}_i \) and \( \hat{p}_i \), and predictors \( \hat{y}_i \) and mixture weights \( \hat{p}_i \) are updated according to Eqs (8) and (9) respectively.

3 Theoretical analysis

The proposed DTLR approach is shown to be consistent and analyzed in the simple case where \( \alpha \) is a diagonal matrix (\( \alpha_{ij} = \alpha \delta_{ij} \)).

3.1 Loss function and related optimal predictor

Let us assume that the hyper-parameters of the model have been identified together with predictors \( \hat{y}_i(x) \) and weights \( \hat{p}_i(x) \). These are leveraged to achieve the prediction of the effect series. For any given input \( x \), the sought eventual predictor is expressed as \( (\hat{y}(x), \hat{I}(x)) \) where \( \hat{I}(x) \) is the predicted time lag and \( \hat{y}(x) \) the predicted value. The associated \( L_2 \) loss is:

\[
L_2(\hat{y}, \hat{I}) = \mathbb{E}_{data}\left[ (y_I(x) - \hat{y}(x))^2 \right].
\]

Then it comes:
Proposition 3.1. With same notations as in Eq. [2], with $\alpha_{ij} = \alpha \delta_{ij}$, $\alpha > 0$, the optimal composite predictor $(y^*, I^*)$ is given by
\[
y^*(x) = \hat{y}_I(x) = \arg \max_i \left( \hat{p}_i(x) \right),
\]
with $I^*(x) = \arg \max_i \left( \hat{p}_i(x) \right)$.

Proof. In supplementary material, Appendix C.

### 3.2 Linear Stability Analysis

The saddle point (Eqs [7] [8] [9] [10]) admits among others a degenerate solution, corresponding to $\hat{y}_i(x) = 1/|T|$, $\alpha_{ij} = 0$ for all pairs $(i, j)$, with $\sigma^2 = \sigma^2_0$. Informally the model converges toward this degenerate trivial solution when there is not enough information to build specialized predictors $\hat{y}_i$.

Let us denote $\Delta y_i^2(x) = (y_i - \hat{y}_i(x))^2$ the square error made by predictor $\hat{y}_i$ for $x$, and
\[
\sigma_0^2 = \frac{1}{|T|} \mathbb{E}_{\text{data}} \left( \sum_{i \in T} \Delta y_i^2(x) \right)
\]
the average of MSE over the set of the predictors $\hat{y}_i$, $i \in T$.

Let us investigate the conditions under which the degenerate solution may appear, by computing the Hessian of the log-likelihood and its eigenvalues. Under the simplifying assumption $\alpha_{ij} = \alpha \delta_{ij}$, the model involves $2|T|$ functional parameters $\hat{y}$ and $\hat{p}$ and two hyper-parameters $\alpha$ and $r = \sigma^2 / \sigma_0^2$. After the computation of the Hessian (in supplementary material, Appendix B) the system involves three key statistical quantities, two global ones:
\[
C_1[\mathbf{q}] = \frac{1}{\sigma_0^2} \mathbb{E}_{\text{data}} \left( \sum_{i \in T} q_i(x, y) \Delta y_i^2(x) \right),
\]
\[
C_2[\mathbf{q}] = \frac{1}{\sigma_0^2} \mathbb{E}_{\text{data}} \left[ \sum_{i \in T} q_i(x, y) \left( \Delta y_i^2(x) - \sum_{j \in T} q_j(x, y) \Delta y_j^2(x) \right)^2 \right],
\]
and a local $|T|$-vector of components
\[
u_i(x, \mathbf{q}) = \frac{1}{\sigma_0^2} \mathbb{E}_{\text{data}} \left[ q_i(x, y) \left( \Delta y_i^2(x) - \sum_{j \in T} q_j(x, y) \Delta y_j^2(x) \right) \right].
\]

Up to a constant, $C_1$ represents the covariance between the latent variables $\{\tau_i\}$ and the normalized predictor errors. $C_1$ smaller than one indicates a positive correlation between the latent variables and small errors; the smaller the better. For the degenerate solution, i.e. $\mathbf{q} = \mathbf{q}_0$ uniform, $C_1[\mathbf{q}_0] = 1$ and $C_2[\mathbf{q}_0]$ represents the default variability among the prediction errors. $u_i(x, \mathbf{q})$ informally measures the quality of predictor $\hat{y}_i$ relatively to the other ones at $x$. More precisely, a negative value of $u_i(x, \mathbf{q})$ indicates that $\hat{y}_i$ is doing better than average in the neighborhood of $x$.

At a saddle point the parameters are given by:
\[
\sigma_0^2 \frac{\sigma^2}{\sigma_0^2} = \frac{|T|}{|T| - 1} \cdot C_1[\mathbf{q}] \quad \text{and} \quad \alpha = \frac{|T|}{|T| - 1} \cdot 1 - C_1[\mathbf{q}].
\]

The predictors $\hat{y}$ are decoupled from the rest whenever they are centered, which we assume. So the analysis can focus on the other parameters.

If $\hat{p}$ is fixed a saddle point is stable iff
\[
C_2[\mathbf{q}] < 2C_1^2[\mathbf{q}] + O \left( \frac{1}{|T|} \right).
\]

In particular, the degenerate solution is unstable if
\[
C_2[\mathbf{q}_0] > 2 \left( 1 - \frac{1}{|T|} \right).
\]

Note that for $\Delta y_i^2(x)$ iid centered with variance $\sigma^2_0$ and relative kurtosis $\kappa$ (conditionally to $x$) one has $C_2 = (2 + \kappa)(1 - 1/|T|)$. Therefore, whenever $\Delta y_i^2(x)$ fluctuates and the relative kurtosis is non-negative, the degenerate solution is unstable and will thus be avoided.
If \( \hat{p} \) is allowed to evolve (after Eq. (9)) the degenerate trivial solution becomes unstable as soon as \( C_2[q_0] \) is non-zero, due to the fact that the gradient points in the opposite direction to \( u(x) \) (with \( \dot{p}(x) \propto -C_2[q_0]u(x) \)), thus rewarding the predictors with lowest errors by increasing their weights.

The system is then driven toward other solutions, among which the localized solutions of the form:

\[
\hat{p}_i(x) = \delta_{i,I(x)},
\]

with an input dependent index \( I(x) \in T \). As shown (in supplementary material, Appendix C) the maximum likelihood localized solution also minimizes the loss function (Eq. 10). The stability of such localized solutions and the existence of other (non-localized) solutions is left for further work.

## 4 The DTLR Algorithm

The DTLR algorithm learns both regression models \( \hat{y}(x) \) and \( \hat{p}(x) \) from series \( x_t \) and \( y_t \), using alternate optimization of the model parameters and the model hyper-parameters \( \alpha \) and \( \sigma^2 \), after Eqs (7, 8, 9). The model search space is that of neural nets, parameterized by their weight vector \( \theta \). The inner optimization loop updates \( \theta \) using mini-batch based stochastic gradient descent. At the end of each epoch, after all minibatches have been considered, the outer optimization loop computes hyper-parameters \( \alpha \) and \( \sigma^2 \) on the whole data.

**Initialization** of \( \alpha \) and \( \sigma \)

\[
it \leftarrow 0;
\]

while \( it < \text{max it} \) do

while epoch do

\[
\begin{align*}
\theta & \leftarrow \text{Optimize}(\mathcal{L}(\theta, \alpha, \sigma^2)); \\
\sigma^2 & \leftarrow \sigma_0^2 \frac{|T| - C_1[q]}{|T| - 1}; \\
\alpha & \leftarrow \frac{|T| - 1}{|T| - C_1[q]};
\end{align*}
\]

end

Result: Model parameters \( \theta = \{\hat{y}, \hat{p}\} \), hyper-parameters \( \alpha, \sigma^2 \)

**Algorithm 1**: DTLR algorithm

The algorithm code is available in supplementary material and will be made public after the reviewing period. The initialization of hyper-parameters \( \alpha \) and \( \sigma \) is settled using preliminary experiments (same setting for all considered problems: \( \alpha \sim U(0.75, 2); \sigma^2 \sim U(10^{-3}, 5) \)).

The neural architecture implements predictors \( \hat{y}(x) \) and weights \( \hat{p}(x) \) on the top of a same feature extractor from input \( x \). In the experiments, the architecture of the feature extractor is a 2-hidden layer fully connected network. On the top of the feature extractor are the single layer \( \hat{y} \) and \( \hat{p} \) models, each with \( |T| \) output neurons, with \( |T| \) the size of the chosen domain for the time lag.

## 5 Experimental Setting

The goal of experimental validation is threefold. A first issue regards the accuracy of DTLR, measured from the mean absolute error (MAE), root mean square error (RMSE) and Pearson correlation of the learned DTLR model \( (y^*(x_t), I^*(x_t)) \). DTLR is compared to the natural baseline defined as the regressor with constant time lag, \( \hat{y}_\Delta(x_t) \), with \( \Delta \) being the average of all possible time lags in \( T \). The predictions of DTLR and the baseline are compared with the ground truth value of the effect series. However the predicted time lag \( I^*(x_t) \) can only be assessed if the ground truth time-lag relationship is known. In order to do so, three synthetic problems of increasing difficulty are defined below.

Secondly, the stability and non-degeneracy of the learned model are assessed from the statistical quantities \( \sigma_0 \) and \( C_1 \) (section 3.2), compared to the degenerate solution \( \hat{p}_i(x) = 1/|T| \). For \( C_1 < 1 \), the model accurately specializes the found predictors \( \hat{p}_i \).

Lastly, and most importantly, DTLR is assessed on the solar wind prediction problem, and compared to the best state of the art in space weather.
Synthetic Problems. Four synthetic problems of increasing difficulty are generated using Stochastic Langevin Dynamics. In all problems, the cause signal \( x_t \in \mathbb{R}^{10} \) and the effect signal \( y_t \) are generated as follows (with \( \eta = 0.02, s^2 = 0.7 \)):\[
\begin{align*}
    x_{t+1} &= (1 - \eta)x_t + \mathcal{N}(0, s^2) \\
    v_t &= k||x_t||^2 + c \\
    y_t + g(x_t) &= f(v_t),
\end{align*}
\]with time-lag mapping \( g(x_t) \) ranges in a time interval with width 20 (except for problem I where \(|T| = 15\)). The complexity of the synthetic problems is governed by the amplitude and time-lag functions \( f \) and \( g \) (more in appendix, Table 2).

<table>
<thead>
<tr>
<th>Problem</th>
<th>( f(v_t) )</th>
<th>( g(x_t) )</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>( v_t )</td>
<td>5</td>
<td>( k=10, c=0 )</td>
</tr>
<tr>
<td>II</td>
<td>( v_t )</td>
<td>( 100/v_m )</td>
<td>( k=1, c=10 )</td>
</tr>
<tr>
<td>III</td>
<td>( \sqrt{v_t^2 + 2ad} )</td>
<td>( (\sqrt{v_m^2 + 2ad} - v) / a )</td>
<td>( k=5, a=5, d=1000, c=100 )</td>
</tr>
<tr>
<td>IV</td>
<td>( v_t )</td>
<td>( g(x_t) = \exp(v_t)/1(1+\exp(v_t/20)) )</td>
<td>( k=10, c=40 )</td>
</tr>
</tbody>
</table>

Solar Wind Speed Prediction. The challenge of predicting solar wind speed from heliospheric data is due to the non-stationary propagation time of the solar plasma through the interplanetary medium. For the sake of a fair comparison with the best state of the art, the same experimental setting is used. The cause series \( x_t \) includes the solar magnetic (flux tube expansion, FTE) and the coronal magnetic field strength estimates produced by the current sheet source surface (Zhao and Hoeksema, 1995) model, exploiting the hourly magnetogram data recorded by the Global Oscillation Network Group from 2008 to 2016. The effect series, the hourly solar wind data is available from the OMNI data base from the Space Physics Data Facility. After domain knowledge, the time-lag ranges from 2 to 5 days, segmented in six-hour segments (thus \(|T| = 12\)). For the \( i \)-th segment, the “ground truth” solar wind \( y_i \) is set to its median value over the 6 hours.

DTRL is validated using a nine fold cross-validation (Table 3 in appendix), where each fold is a continuous period corresponding to a solar rotation.

6 Empirical validation

Table 1: DTLR performance: accuracy (MAE and RMSE, the lower the better; Pearson, the higher the better) and stability \( \sigma_0 \) and \( C_1 \) (the lower the better). For each indicator, is reported the DTLR value (9-fold CV), the baseline value and the time-lag error.

<table>
<thead>
<tr>
<th>Problem</th>
<th>M.A.E</th>
<th>R.M.S.E</th>
<th>Pearson Corr.</th>
<th>( \sigma_0 )</th>
<th>( C_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>8.82</td>
<td>21.79</td>
<td>0.021</td>
<td>12.35</td>
<td>28.79</td>
</tr>
<tr>
<td>II</td>
<td>10.15</td>
<td>27.40</td>
<td>0.4</td>
<td>13.70</td>
<td>35.11</td>
</tr>
<tr>
<td>III</td>
<td>3.17</td>
<td>11.01</td>
<td>0.17</td>
<td>4.63</td>
<td>14.99</td>
</tr>
<tr>
<td>IV</td>
<td>3.88</td>
<td>12.28</td>
<td>0.34</td>
<td>5.33</td>
<td>15.89</td>
</tr>
<tr>
<td>Solar Wind</td>
<td>56.35</td>
<td>60.45</td>
<td>74.20</td>
<td>84.53/ →</td>
<td>0.6/ 0.41/ →</td>
</tr>
</tbody>
</table>

6.1 Synthetic problems

On the easy Problem I, DTLR predicts the correct time lag for 97.93% of the samples. The higher value of \( \sigma_0 \) in problems I and II compared to the other problems is explained from the higher variance in the effect series \( y(t) \).

On Problem II, DTLR accurately learns the inverse relationship between \( x_t, g(x_t) \) and \( y_t \) on average. The time lag is overestimated in the regions with low time lag (with high velocity), which is blamed
on the low sample density in this region, due to the data generation process. Interestingly, Problems III and IV are better handled by DTLR, despite a more complex dynamic time lag relationship. In both latter cases however, the model tends to under-estimate the time lag in the high time lag region and conversely to over-estimate it in the low time lag region.

6.2 THE SOLAR WIND PROBLEM

DTLR finds an operational solar wind model (Table 1), though the significantly higher difficulty of the solar wind problem is witnessed by the $C_1$ value close to the degenerate value 1. The detailed comparison with the state of the art [Reiss et al. (2019)] (Fig. 1 Left) shows that DTLR improves on the current best state of the art (on all variants including ensemble approaches, and noting that median models are notoriously hard to beat). (Fig. 1 Right) shows the good correlation between the predicted solar wind and the measured solar wind.

<table>
<thead>
<tr>
<th>Model</th>
<th>M.A.E</th>
<th>R.M.S.E</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS</td>
<td>74.99</td>
<td>85.27</td>
</tr>
<tr>
<td>DCHB</td>
<td>83.83</td>
<td>103.43</td>
</tr>
<tr>
<td>WSA</td>
<td>68.54</td>
<td>82.62</td>
</tr>
<tr>
<td>Ensemble Median (WS)</td>
<td>71.52</td>
<td>83.36</td>
</tr>
<tr>
<td>Ensemble Median (DCHB)</td>
<td>78.27</td>
<td>100.04</td>
</tr>
<tr>
<td>Ensemble Median (WSA)</td>
<td>62.24</td>
<td>74.86</td>
</tr>
<tr>
<td>Persistence (4 days)</td>
<td>130.48</td>
<td>161.99</td>
</tr>
<tr>
<td>Persistence (27 days)</td>
<td>66.54</td>
<td>78.86</td>
</tr>
<tr>
<td>Fixed Lag Baseline</td>
<td>67.33</td>
<td>80.39</td>
</tr>
<tr>
<td>DTLR</td>
<td>60.19</td>
<td>72.64</td>
</tr>
</tbody>
</table>

(a) Comparative assessment on the Solar Wind problem compared to the state of the art [Reiss et al. (2019) Table 1]

Figure 1: DTLR on the solar wind problem. Left: comparative quantitative assessment w.r.t. the state of the art. Right: qualitative assessment of the prediction.

7 DISCUSSION AND PERSPECTIVES

The contribution of the paper is twofold. A new ML setting, Dynamic Time Lag Regression has been defined, aimed at the modelling of varying time-lag dependency between time series. The introduction of this new setting is motivated by an important scientific and practical problem from the domain of space weather, an open problem for over two decades.

Secondly, a Bayesian formalization has been proposed to tackle the DTLR problem, relying on a saddle point optimization process. A closed form analysis of the training procedure stability under simplifying assumptions has been conducted, yielding a practical alternate optimization formulation, implemented in the DTLR algorithm. This algorithm has been successfully validated on synthetic and real-world problems, although some bias toward the mean has been detected in some cases.

On the methodological side, this work opens a short term perspective (handling the bias) and a longer term perspective, extending the proposed nested inference procedure and integrating the model selection step within the inference architecture. The challenge is to provide the algorithm with the means of assessing online the stability and/or the degeneracy of the learning trajectory.

Regarding the motivating solar wind prediction application, a next step consists of enriching the data sources and the description of the cause series $x_t$, typically by directly using the solar images. Another perspective is to consider other applications of the general DTLR setting, e.g. considering fine-grained modelling of diffusion phenomena.

The predicted values, every 6 hours, are interpolated for comparison with the hourly measured solar wind.
REFERENCES


Diego Pennacchioli, Michele Coscia, Salvatore Rinzivillo, Fosca Giannotti, and Dino Pedreschi. The retail market as a complex system. EPJ Data Sci., 3(1):33, 2014.


APPENDIX A  LOG LIKELIHOOD OF THE LATENT MODEL (3)

A.1 DIRECT COMPUTATION

Due to the single effect constraint (4) the mixture model (3) can be expressed simply as

\[ P(y|x) = \left( \sum_{i \in T} \hat{p}_i(x) \prod_{j \in T} \sqrt{\frac{1 + \alpha_{ji}}{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_j - \hat{y}_j(x))^2} \right) \]

\[ = \left( \sum_{i \in T} \hat{p}_i(x) \prod_{j \in T} \sqrt{\frac{1 + \alpha_{ji}}{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_j - \hat{y}_j(x))^2} \right) \exp\left(-\frac{1}{2\sigma^2} \sum_{j \in T} (y_j - \hat{y}_j(x))^2\right) \]

Let \( \theta \stackrel{\text{def}}{=} (\hat{y}, \hat{p}, \sigma, \alpha) \) denote the parameters of the model and consider the probability that predictor \( \hat{y}_i \) is the good one conditionally to a pair of observation \((x, y)\):

\[ q_i(x, y) = P(\tau_i = 1 | x, y) \]

\[ = \frac{1}{Z(x, y | \theta)} \hat{p}_i(x) \exp\left(-\frac{1}{2\sigma^2} \sum_{j \in T} \alpha_{ji} (y_j - \hat{y}_j(x))^2 + \frac{1}{2} \sum_{j \in T} \log(1 + \alpha_{ji})\right) \]

with

\[ Z(x, y | \theta) = \sum_{i \in T} \hat{p}_i(x) \exp\left(-\frac{1}{2\sigma^2} \sum_{j \in T} \alpha_{ji} (y_j - \hat{y}_j(x))^2 + \frac{1}{2} \sum_{j \in T} \log(1 + \alpha_{ji})\right). \]

This gives immediately

\[ \mathcal{L}[(x, y) | \text{data}, \theta] = -|T| \log(\sigma) - \mathbb{E}_{\text{data}} \left[ \sum_{i \in T} \frac{1}{2\sigma^2} (y_i - \hat{y}_i(x))^2 - \log(Z(x, y | \theta)) \right] \]

A.2 LARGE DEVIATION ARGUMENT

Even though the log likelihood can be obtained by direct summation, for sake of generality we show how this can result from a large deviation principle. Assume that the number of learning samples tends to infinity, and so that in a small volume \( d\nu = dx dy \) around a given joint configuration \((x, y)\), the number of data \( N_{x,y} \) becomes large. Restricting the likelihood to this subset of the data yields the following:

\[ \mathcal{L}_{x,y} = \prod_{m=1}^{N_{x,y}} \sum_{\{\tau^{(m)}\}} \hat{p}(\tau^{(m)} | x) \prod_{i \in T} \frac{1}{\sqrt{2\pi} \sigma_i} \exp\left(-\frac{1}{2} \sum_{i \in T} \frac{(y_i - \hat{y}_i(x))^2}{\sigma_i^2} \right). \]

Upon introducing the relative frequencies:

\[ q_i(x, y) = \frac{1}{N_{x,y}} \sum_{m=1}^{N_{x,y}} \tau_i^{(m)} \]  

satisfying \( \sum_{i \in T} q_i(x, y) = 1 \),

the sum over the \( \tau_i^{(m)} \) is replaced by a sum over these new variables, with the summand obeying a large deviation principle

\[ \mathcal{L}_{x,y} \approx \sum_q \exp\left(-N_{x,y} \mathcal{F}_{x,y}[q]\right) \]

where the rate function reads

\[ \mathcal{F}_{x,y}[q] = |T| \log(\sigma) + \sum_{i \in T} \left[ (y_i - \hat{y}_i(x))^2 + \frac{1}{2\sigma^2} \sum_{j \in T} \alpha_{ij} q_j \right] - \frac{1}{2} \sum_{j \in T} \log(1 + \alpha_{ji}) + q_i \log \frac{q_i}{\hat{p}_i} \]

Taking the saddle point for \( q_i \) yield as a function of \( (x, y) \) expression (7). Inserting this into \( \mathcal{F} \) and taking the average over the data set yields the log likelihood (5) with opposite sign:

\[ \mathcal{L}[(x, y) | \text{data}, \theta] = -\mathbb{E}_{\text{data}} \left[ \mathcal{F}_{x,y}[q(x, y)] \right]. \]
A.3 Saddle point equations

Now we turn to the self-consistent equations relating the parameters \( \theta \) of the model at a saddle point of the log likelihood function. First, the optimization of the predictors \( \hat{y} \) yields:

\[
\frac{\partial L}{\partial \hat{y}_i(x)} = \frac{1}{\sigma^2} E_{data} \left[ (y_i - \hat{y}_i(x)) \left( 1 + \sum_{j \in T} \alpha_{ij} q_j(x, y) \right) \right].
\]

Then the optimization of \( \hat{p} \) gives:

\[
\frac{\partial L}{\partial \hat{p}_i(x)} = E_{data} \left[ q_i(x, y) \hat{p}_i(x) - \lambda(x) \right] = \frac{1}{\hat{p}_i(x)} E_{data} \left[ q_i(x, y) \right] - \lambda(x)
\]

with \( \lambda(x) \) a Lagrange multiplier to insure that \( \sum_i \hat{p}_i(x) = 1 \) This gives

\[
\hat{p}_i(x) = \frac{1}{\lambda(x)} E_{data} \left[ q_i(x, y) \right] \]

Hence

\[
\sum_{i \in T} \hat{p}_i(x) = \frac{1}{\lambda(x)} = 1 \quad \forall x
\]

in order to fulfill the normalization constraint, yielding finally expression (9).

Finally the optimization of \( \alpha \) reads:

\[
\frac{\partial L}{\partial \alpha_{ij}} = \frac{1}{2(1 + \alpha_{ij})} E_{data} \left[ q_j(x, y) \right] - \frac{1}{2\sigma^2} E_{data} \left[ (y_i - \hat{y}_i(x))^2 q_j(x, y) \right].
\]

**Appendix B Proof of Proposition 3.1**

Given \( I(x) \) a candidate index function we associate the point-like measure

\[
p_i(x) = \delta_{i,I(x)}.
\]

Written in terms of \( p \) the loss function reads

\[
L_2(\hat{y}, p) = E_{x,y} \left[ \sum_{i \in T} p_i(x) (y_i - \hat{y}(x))^2 \right].
\]

Under (3) (with \( \alpha_{ij} = \alpha \delta_{ij} \)) the loss is equal to

\[
L_2(\hat{y}, p) = \mathbb{E}_x \left[ \sum_{i \in T} p_i(x) \left( (\hat{y}_i(x) - \hat{y}(x))^2 - \hat{p}_i(x) \frac{\alpha \sigma^2}{1 + \alpha} \right) \right] + \sigma^2
\]

The minimization w.r.t. \( \hat{y} \) yields

\[
\hat{y}(x) = \sum_{i \in T} p_i(x) \hat{y}_i(x). \tag{16}
\]

In turn, as a function of \( p \), the loss being a convex combination, its minimization yields

\[
p_i(x) = \delta_{i,I(x)}, \tag{17}
\]

\[
I(x) = \arg \min_{i \in T} \left( (\hat{y}_i(x) - \hat{y}(x))^2 - \hat{p}_i(x) \frac{\alpha \sigma^2}{1 + \alpha} \right). \tag{18}
\]

Combining these equations (16,17,18) we get

\[
I(x) = \arg \max_{i \in T} \hat{p}_i(x),
\]

which concludes the proof.
APPENDIX C STABILITY ANALYSIS

The analysis is restricted for simplicity to the case $\alpha_{ij} = \alpha \delta_{ij}$. The log likelihood as a function of $r = \sigma^2 / \sigma_0^2$ and $\beta = \alpha / r$ after inserting the optimal $q = q(x, y)$ reads in that case

$$
L(r, \beta) = -\frac{|T|}{2} \log(r) - \frac{|T|}{2r} \log(1 + r \beta) + \mathbb{E}_{data} \left[ \log(Z) - \lambda(x) \sum_{i \in T} \hat{p}_i(x) \right]
$$

with

$$
Z = \sum_i \hat{p}_i(x) \exp \left( -\frac{\beta}{2\sigma_0^2} \Delta y_i^2(x) \right),
$$

and where $\lambda(x)$ is a Lagrange multiplier which has been added to impose the normalization of $\hat{p}$. The gradient reads

$$
\frac{\partial L}{\partial r} = \frac{1}{2r^2} \left( |T| (1 - r) + \frac{\beta r^2}{1 + \beta r} \right),
$$

$$
\frac{\partial L}{\partial \beta} = \frac{r}{2(1 + r \beta)} - \frac{1}{2} C_1[q],
$$

$$
\frac{\partial L}{\partial \hat{y}_i(x)} = \frac{1}{\sigma^2} \mathbb{E}_{data} \left[ \hat{y}_i(x) - \hat{y}_i(x) \right] \left( 1 + \alpha q_i(x, y) \right),
$$

$$
\frac{\partial L}{\partial \hat{p}_i(x)} = \frac{\mathbb{E}_{data} \left[ q_i(x, y) \right]}{\hat{p}_i(x)} - \lambda(x),
$$

with

$$
C_1[q] = \frac{1}{\sigma_0^2} \mathbb{E}_{data} \left( \sum_{i \in T} q_i(x, y) \Delta y_i^2(x) \right),
$$

This leads to the following relation at the saddle point:

$$
r = \frac{|T| - C_1[q]}{|T| - 1},
$$

$$
\alpha = \frac{|T| - 1 - C_1[q]}{|T| - 1 C_1[q]},
$$

$$
\hat{y}_i(x) = \frac{\mathbb{E}_{data} \left[ y_i (1 + \alpha q_i(x, y)) \right]}{\mathbb{E}_{data} \left[ 1 + \alpha q_i(x, y) \right]},
$$

$$
\hat{p}_i(x) = \mathbb{E}_{data} \left[ q_i(x, y) \right].
$$

Let us now compute the Hessian. It is easy to see that the block corresponding to the predictors $\hat{y}$ decouples from the rest as soon as these predictors are centered.

Denoting

$$
C_2[q] = \frac{1}{\sigma_0^4} \mathbb{E}_{data} \left[ \sum_{i \in T} q_i(x, y) \left( \Delta y_i^2(x) - \sum_{j=1}^n q_j(x, y) \Delta y_j^2(x) \right)^2 \right],
$$

13
we have
\[\frac{\partial^2 L}{\partial r^2} = \frac{1}{2} \left( -|\mathcal{T}| + 2 \frac{|\mathcal{T}|}{|\mathcal{T}|} \left( C_1[q] - 1 \right) - \beta^2 C_1^2[q] \right)\]
\[\frac{\partial^2 L}{\partial \beta \partial \beta} = \frac{1}{2} C_2^2[q]\]
\[\frac{\partial^2 L}{\partial \beta^2} = \frac{1}{4} \left( C_2[q] - 2 C_1^2[q] \right)\]
\[\frac{\partial^2 L}{\partial p_i(x) \partial p_j(x)} = -\frac{E_{\text{data}} \left[ q_i(x, y) q_j(x, y) | x \right]}{\hat{p}_i(x) \hat{p}_j(x)}\]
\[\frac{\partial^2 L}{\partial r \partial \hat{p}_i(x)} = 0\]
\[\frac{\partial^2 L}{\partial \beta \partial \hat{p}_i(x)} = -\frac{u_i[x, q]}{2 \hat{p}_i(x)},\]
where
\[u_i[x, q] \overset{\text{def}}{=} \frac{1}{\sigma_0^2} E_{\text{data}} \left[ q_i(x, y) (\Delta y_i^2(x) - \sum_{j \in \mathcal{T}} q_j(x, y) \Delta y_j^2(x)) | x \right].\]

There are two blocks in this Hessian, the one corresponding to \(r\) and \(\beta\) and the one corresponding to derivatives with respect to \(\hat{p}_i\). The stability of the first one depends on the sign of \(C_2[q] - 2 C_1^2[q]\) for \(|\mathcal{T}|\) large while the second block is always stable as being an average of the exterior product of the vector \((q_1(x, y) / \hat{p}_1(x), \ldots, q_{|\mathcal{T}|}(x, y) / \hat{p}_{|\mathcal{T}|}(x))\) by itself. At the degenerate point \(\alpha = 0, r = 1, \hat{p}_i = 1 / |\mathcal{T}|\) the Hessian simplifies as follows. Denote
\[d\eta = dr e_1 + d\beta e_2 + \int dx \sum_{i \in \mathcal{T}} dp_i(x) e_{i+2}(x)\]
a given vector of perturbations, decomposed onto a set of unit tangent vectors, \( \{ e_1 \text{ and } e_2 \} \) being respectively associated to \(r\) and \(\beta\), while \(e_i(x)\) associated to \(\hat{p}_i(x)\) for all \(i \in \mathcal{T}\) and \(x \in \mathcal{X}\). Denote
\[u = \sum_{i \in \mathcal{T}} \int dx u_i[x] e_i(x)\]
\[v(x) = \sum_{i \in \mathcal{T}} e_i(x)\]
with
\[C_2 = \frac{1}{|\mathcal{T}| \sigma_0^2} E_{\text{data}} \left[ \sum_{i \in \mathcal{T}} (\Delta y_i^2(x) - \frac{1}{|\mathcal{T}|} \sum_{j \in \mathcal{T}} \Delta y_j^2(x))^2 \right].\]
\[u_i[x] = \frac{1}{\sigma_0^2} E_{\text{data}} [\Delta y_i^2(x) - \sigma_0^2 | x].\]

With these notations the Hessian reads:
\[H = \frac{1}{2} \left( -|\mathcal{T}| e_1 e_1^t + e_1 e_2^t + e_2 e_1^t + \left( \frac{C_2}{2} - 1 \right) e_2 e_2^t - u e_2 - e_2 u^t - \int dx v(x) v^t(x) \right).\]

In fact we are interested in the eigenvalues of \(H\) in the subspace of deformations which conserve the norm of \(\vec{p}\), i.e. orthogonal to \(v(x)\), thereby given by
\[\eta = \eta_1 e_1 + \eta_2 e_2 + \eta_3 u.\]

In this subspace the Hessian reads
\[H = \frac{1}{2} \begin{bmatrix} -|\mathcal{T}| & 1 & 0 \\ 1 & \frac{C_2}{2} - 1 & -M|\mathcal{T}| C_2 \\ 0 & -M|\mathcal{T}| C_2 & 0 \end{bmatrix},\]
where \( M \) is the number of data points, resulting from the fact that
\[
\sum_{i \in T} \int dx u_i^2[x] = \frac{M}{\sigma_0^2} \mathbb{E}_{\text{data}} \left[ \sum_{i \in T} (\Delta y_i^2(x) - \sigma_0^2)^2 \right],
\]
\[
= MC_2,
\]
because \( \mathbb{E}_{\text{data}}(\cdot|x) \) as a function of \( x \) is actually a point-wise function on the data. If \( |u|^2 > 0 \) or if \( |u| = 0 \) and \( 1 + |T|(C_2/2 - 1) > 0 \) there is at least one positive eigenvalue. Let \( \Lambda \) be such an eigenvalue. After eliminating \( dr \) and \( d\beta \) from the eigenvalue equations in \( d\eta \), the deformation along this mode verifies
\[
d\eta \propto \Lambda e_1 + \Lambda(|T| + \Lambda)e_2 - M|T|(|T| + \Lambda)C_2u,
\]
which corresponds to increasing \( r \) and \( \alpha \) while decreasing for each \( x \) the \( \hat{p}_i \), having the highest mean relative error \( u_i[x] \).

Concerning solutions for which
\[
\hat{p}_i(x) = \delta_{iI(x)}
\]
is concentrated on some index \( \hat{I}(x) \), the analysis is more complex. In that case \( C_2[p] = 0 \) and \( C_1[p] > 0 \). The \((r, \beta)\) sector has 2 negative eigenvalues, while the \( p \) block is \((-\) a covariance matrix, so it has as well negative eigenvalues. The coupling between these two blocks could however in principle generate in some cases some instabilities.

Still, the log likelihood of such solutions reads
\[
\mathcal{L} = -\frac{|T|}{2} \log(\sigma^2) + \frac{1}{2} \log(1 + \alpha) - \frac{1}{2\sigma^2} \mathbb{E}_{\text{data}} \left[ \sum_{i \in T} \Delta y_i^2(x) \right] - \frac{\alpha}{2\sigma^2} \mathbb{E}_{\text{data}} \left[ \Delta y_{\hat{I}(x)}^2(x) \right]
\]
so we get the following optimal solution
\[
\sigma^2 = \frac{1}{|T|} \mathbb{E}_{\text{data}} \left[ \sum_{i \in T} \Delta y_i^2(x) \right],
\]
\[
\frac{1}{1 + \alpha} = \frac{\mathbb{E}_{\text{data}} \left[ \Delta y_{\hat{I}(x)}^2(x) \right]}{\sigma^2},
\]
\[
I(x) = \arg\min_{i \in T} \mathbb{E}_{\text{data}} \left[ \Delta y_i^2(x) | x \right].
\]
APPENDIX D  EXPERIMENTS: ADDITIONAL DETAILS

Here we provide some additional details and context to the experimental validation of the DTLR methodology described in section 5. Table 2 provides some information about the datasets used in the synthetic and solar wind prediction problems. Sections D.1.1 and D.1.2 give additional plots for evaluating the experimental results.

For the solar wind prediction task, the solar wind data was mapped into standardized Gaussian space using a quantile-quantile and inverse probit mapping. Nine fold cross-validation was performed using splits as specified in table 3. To compare the DTLR results with the state of the art solar wind forecasting, we used results from Reiss et al. (2019, Table 1). Since Reiss et al. (2019) compared the various forecasting methods on only one solar rotation (first row of table 3), comparing these results with DTLR can be considered as a preliminary examination. Nevertheless, the results presented in table 1a show encouraging signs for the competitiveness and usefulness of the DTLR method.

Table 2: Synthetic and Real-World Problems

<table>
<thead>
<tr>
<th>Problem</th>
<th># train</th>
<th># test</th>
<th>(d)</th>
<th>(T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>10,000</td>
<td>2,000</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>II</td>
<td>10,000</td>
<td>2,000</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>III</td>
<td>10,000</td>
<td>2,000</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>IV</td>
<td>10,000</td>
<td>2,000</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Solar Wind</td>
<td>77,367</td>
<td>2,205</td>
<td>374</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 3: Cross validation splits used to evaluate DTLR on the solar wind forecasting task

<table>
<thead>
<tr>
<th>Split Id</th>
<th>Carrington Rotation</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2077</td>
<td>2008/11/20 07:00:04</td>
<td>2008/12/17 14:38:34</td>
</tr>
<tr>
<td>3</td>
<td>2104</td>
<td>2010/11/26 17:32:44</td>
<td>2010/12/24 01:15:56</td>
</tr>
<tr>
<td>4</td>
<td>2117</td>
<td>2011/11/16 07:04:41</td>
<td>2011/12/13 14:39:28</td>
</tr>
<tr>
<td>5</td>
<td>2130</td>
<td>2012/11/04 20:39:43</td>
<td>2012/12/02 04:06:23</td>
</tr>
<tr>
<td>6</td>
<td>2143</td>
<td>2013/10/25 10:17:52</td>
<td>2013/11/21 17:36:35</td>
</tr>
<tr>
<td>7</td>
<td>2157</td>
<td>2014/11/11 07:09:56</td>
<td>2014/12/08 14:41:02</td>
</tr>
<tr>
<td>9</td>
<td>2184</td>
<td>2016/11/16 17:41:04</td>
<td>2016/12/14 01:16:43</td>
</tr>
</tbody>
</table>

*In the solar wind problem, the training and test data sizes correspond to one cross-validation split*
D.1 Supplementary Plots

D.1.1 Synthetic Problems

(a) **Problem II**, Goodness of fit, Output $y(x)$

(b) **Problem II**, Goodness of fit, Time lag $\tau(t)$

(c) **Problem II**, Error of time lag prediction

(d) **Problem II**, Output vs Time Lag Relationship

Figure 2: **Problem II**, Results
Figure 3: Problem III, Results
Figure 4: Problem IV, Results
D.1.2 Solar Wind Prediction

(a) Hourly forecasts for period 2008-11-20 07:00 to 2008-12-17 14:00
(b) Hourly forecasts for period 2016-11-16 17:00 to 2016-12-14 01:00

Figure 5: Solar Wind Prediction: reconstructed time series predictions

Appendix E Neural Network Architecture Details

Table 4: Network Architecture Details

<table>
<thead>
<tr>
<th>Problem</th>
<th># Hidden layers</th>
<th>Layer sizes</th>
<th>Activations</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2</td>
<td>[40, 40]</td>
<td>[ReLU, Sigmoid]</td>
</tr>
<tr>
<td>II</td>
<td>2</td>
<td>[40, 40]</td>
<td>[ReLU, Sigmoid]</td>
</tr>
<tr>
<td>III</td>
<td>2</td>
<td>[40, 40]</td>
<td>[ReLU, Sigmoid]</td>
</tr>
<tr>
<td>IV</td>
<td>2</td>
<td>[60, 40]</td>
<td>[ReLU, Sigmoid]</td>
</tr>
<tr>
<td>Solar Wind</td>
<td>2</td>
<td>[50, 50]</td>
<td>[ReLU, Sigmoid]</td>
</tr>
</tbody>
</table>

Figure 6: Architecture of the neural network specified by the number of units \((n^n, n^1_h, n^2_h, 2|T|)\) in each layer.