

# Statistical Models for Unequally Spaced Time Series

Emre Erdogan <sup>\*</sup>    Sheng Ma <sup>†</sup>    Alina Beygelzimer <sup>‡</sup>    Irina Rish <sup>§</sup>

October 5, 2004

## Abstract

Irregularly observed time series and their analysis are fundamental for any application in which data are collected in a distributed or/and asynchronous manor. In this paper, we provide theory-sound, practical approaches to analyze such time series. We propose two models and their parameter estimation algorithms for both stationary and non-stationary irregular time series. Our models can be viewed as extensions of the well-known Auto-Regression (AR) model. We then develop a resampling strategy that uses the proposed models to reduce irregular time series to regular time series. This enables us to take advantage of the vast number of approaches developed for analyzing regular time series. Our experiments with real and synthetic data demonstrate that our approach performs well in computing the basic statistics and doing prediction.

## 1 Introduction

Unevenly sampled time series are common in many real-life applications when measurements are constrained by practical conditions. The irregularity of observations can have several fundamental reasons. First, any event-driven collection process (in which observations are recorded only when some event occurs) is inherently irregular. Second, in such applications as sensor networks (or any distributed monitoring infrastructure), data collection is distributed, and collection agents cannot easily synchronize with one another. In addition, their sampling intervals and policies may be different. Finally, in many applications measurements cannot be made regularly or have to be interrupted due to some events (either foreseen or not).

Time series analysis has a long history. The vast majority of methods, however, can only handle

regular time series and do not easily extend to unevenly sampled data. Continuous time series models can be directly applied for the problem (e.g., [5]), but they tend to be complicated (mostly due to the difficulty of estimating and evaluating them from discretely sampled data) and do not provide a satisfying solution in practice.

In data analysis practice, irregularity is a recognized data characteristic, and practitioners dealt with it heuristically. It is a common practice to ignore the times and treat the data as if it were regular. This can clearly introduce a significant bias leading to incorrect predictions. Consider, for example, the return of a slowly, but consistently changing stock, recorded first very frequently and then significantly less frequently. If we ignore the times, it would appear as if the stock became more rapidly changing, thus riskier, while in fact the statistical properties of the stock did not change. We will discuss other heuristics in the following section.

Many basic questions that are well understood for regular time series, are not dealt with for unequally spaced time series. The goal of this paper is to provide such a theoretical foundation. At the very least, we would like to be able to compute the basic statistics of a given time series (e.g., its mean, variance, autocorrelation), and predict its future values.

We propose two models for dealing with irregularly sampled time series. The first model assumes stationarity and can be viewed as a natural extension of the classical AR(1) model for regular time series. We discuss how to efficiently estimate model's parameters using the maximum likelihood method, and show how to use this model for prediction. The second model relaxes the stationarity assumption by allowing a more general dependence on the current time, time difference and the state of the process at a given time. Both models can be used for converting irregular time series to regular time series by resampling. Such a reduction allows to take advantage and automatically transfer existing models and methods from a well understood domain of regular time series

---

<sup>\*</sup>Columbia University, New York, NY, ee168@columbia.edu

<sup>†</sup>IBM T. J. Watson Research Center, Hawthorne, NY, shengma@us.ibm.com

<sup>‡</sup>IBM T. J. Watson Research Center, Hawthorne, NY, beygel@us.ibm.com

<sup>§</sup>IBM T. J. Watson Research Center, Hawthorne, NY, rish@us.ibm.com

to a relatively unexplored domain of irregular time series. We emphasize that both models are extremely simple and can be efficiently fit to the data.

This paper has the following structure. Section 2 discusses the basic concepts and related work, and Section 3 overviews our approach. Section 4 introduces a statistical model and discusses prediction and estimation of the basic statistics under the stationarity assumption. Section 5 presents a model for non-stationary irregularly sampled time series. In this section, we first introduce a simple model to explain the intuition behind our model and then present a generalization of the simple model. Section 6 further elaborates the resampling approach. Section 7 provides our experimental results with real data.

## 2 Background and Related Work

**2.1 Basic Notation and Definitions** We first define the basic concepts and introduce some of the notation that we will need in the paper.

Bold parameters denote vectors. For example  $\theta$  is a scalar while  $\boldsymbol{\theta}$  is a vector. The transpose of a vector  $\boldsymbol{\theta}$  is denoted by  $\boldsymbol{\theta}^T$ . We use the following convention when representing vectors: Elements in a column vector are separated by semicolons, elements in a row vector – by commas. The notation  $Z \sim \mathcal{N}(0, 1)$  is used to denote that  $Z$  is distributed as a normal random variable with mean 0 and variance 1.

A *time series*  $X(t)$  is an ordered sequence of observations of a variable  $X$  sampled at different points  $t$  over time. Time series data arise in monitoring any system. The observed temporal variations often have some internal structure (e.g., autocorrelation, trend) that can provide an insight into the dynamics of the system under investigation. By analyzing and modeling time series one obviously hopes to understand this dynamics and make predictions about the future. Economic forecasting, stock market analysis, yield projections, process control, workload projections are just a few applications of time series analysis.

Continuous time series are obtained when observations are recorded continuously over some time interval. Recording the observations continuously, however, may be difficult and/or expensive. Therefore, in practice most continuous time series are sampled at discrete time points. Let process  $X(t)$  be sampled at points  $t_0, t_1, \dots, t_n$  satisfying  $0 \leq t_0 < t_1 < \dots < t_n$ .<sup>1</sup> If the time points are equally spaced (i.e.,  $t_{i+1} - t_i = \Delta$  for all  $i = 0, \dots, n - 1$ , where

$\Delta > 0$  is some constant), we call the time series *regularly sampled time series*. Otherwise, the sequence of pairs  $\{X(t_i), t_i\}$  is called an *irregularly sampled time series*.

**Definition 1.** (AUTOCOVARANCE [4])

If  $\{X(t), t \in T\}$  is a process such that  $\mathbf{var}(X(t)) < \infty$  for each  $t \in T$ , then the autocovariance function  $\mathbf{cov}_X(X(t), X(s))$  of  $X(t)$  is given by

$$\mathbf{E}[(X(t) - \mathbf{E}[X(t)])(X(s) - \mathbf{E}[X(s)])]$$

for  $t, s \in T$ .

**Definition 2.** (STATIONARITY [4]) A time series  $\{X(t)\}$  is said to be stationary if

- (i)  $\mathbf{E}[|X(t)|^2] < \infty$ ,
- (ii)  $\mathbf{E}[X(t)] = c$ , for all  $t \in T$ ,
- (iii)  $\mathbf{cov}_X(X(t), X(s)) = \mathbf{cov}_X(X(t+h), X(s+h))$  for all  $t, s, h \in T$ .

In other words, a stationary process is a process whose statistical properties do not vary with time.

**Definition 3.** (AR(1) PROCESS [4]) A regularly sampled process  $\{X(t), t = 0, 1, 2, \dots\}$  is said to be an AR(1) process if  $\{X(t)\}$  is stationary and if for every  $t$

$$X(t) = \theta X(t-1) + \sigma \epsilon_t,$$

where  $\{\epsilon_t\}$  is a series of random variables with expectation  $\mathbf{E}(\epsilon_t) = 0$ , variance  $\mathbf{var}(\epsilon_t) = 1$ , and  $\mathbf{cov}(\epsilon_t, \epsilon_s) = 0$  for every  $t \neq s$ . (The process  $\{\epsilon_t\}$  is also called “white noise”.)

In this paper, we will assume that  $\epsilon_t \sim \mathcal{N}(0, 1)$  for all  $t$ . By recursive substitution, we can write  $X(t+h)$  for any positive integer  $h$  in terms of  $X(t)$  as

$$(2.1) \quad X(t+h) = \theta^h X(t) + \sigma \sum_{j=0}^{h-1} \theta^j \epsilon_{t+1+j}.$$

**2.2 Related work** A vast amount of techniques were developed for analyzing regularly sampled time series. Unfortunately, most of these techniques do not take into account sampling times, and cannot be easily generalized to irregularly sampled time series. Irregular time series are difficult to handle and only a few successful efforts have been made so far.

As a simple heuristic, we can ignore the times and treat the values as regularly sampled. Obviously, if there is enough structure and irregularity in sampling

<sup>1</sup>We will always assume that the sampling times are ordered.

times, we lose a lot of information about the dynamics of the system.

One of the earliest ideas was to resample the irregular time series into regular time series by interpolation, and then use techniques developed for regular time series. A survey of such interpolation techniques can be found in [1]. While this is a reasonable heuristic for dealing with missing values, the interpolation process typically results in a significant bias (e.g., smoothens the data) changing the dynamics of the process. The following quote from Krolik [8], summarizes another problem with this approach: “There is virtually no literature on which method does best with which kind of problem”. Although the quote is more than a decade old, it still seems to be true.

Many techniques have been proposed to handle time series with missing data, which in the limit can be viewed as irregularly sampled [10]. These models try to either fill in the missing values or estimate the parameters directly. However, these models can not be applied if the data is truly unequally spaced with no underlying sampling interval.

A number of authors suggested to use continuous time diffusion processes for the problem. Jones [6] proposed a state-space representation of the process, and showed that for Gaussian inputs and errors, the likelihood of data can be calculated exactly using Kalman [7] filters. A nonlinear (non-convex) optimization can then be used to obtain maximum likelihood estimates of the parameters. Brockwell [3] improved on this model and suggested a continuous time ARMA process driven by the Lévy process. His models, however, assume stationary, and parameter estimation is done via non-convex optimization using Kalman filtering, making these models not very practical. A similar approach is also proposed in [2] in the context of continuous-time finance modeling. This model assumes that the underlying stochastic process is stationary and that the diffusion equation used to describe the process has a specific structure, so it is not practical in areas other than finance.

**2.3 Summary of our Contributions** Our contributions in this paper can be summarised as follows:

- We propose two statistical models for handling irregular time series. Our first model is a simple extension of the classical AR(1) model and assumes that the data is stationary. Our second model does not assume stationarity and allows the state of the process at a time, say  $t + \Delta$ , non-linearly depends on  $t$ ,  $\Delta$  and the state of the process at time  $t$ .

- We show that the parameters of the both model can be estimated efficiently, and derive closed form solutions for estimating the parameters and using the fitted model for prediction.
- We propose two strategies based on the proposed models. The first strategy is to compute the basic statistics (e.g., auto-correlation function) and prediction directly from a model. This approach does not easily extend to non-linear time series and multiple irregular time series. The second strategy avoids these problems by using the model to convert irregular time series to regular time series. The reduction reduces the problem to a problem that has already been thoroughly analyzed and for which many approaches are available.

### 3 Overview of our Approach

Suppose that our irregularly sampled time series  $Y(t)$  does not have a “seasonal component” and can be decomposed as

$$(3.2) \quad Y(t) = a(t) + X(t),$$

where  $a(t)$  is a slowly changing deterministic function called the “trend component” and  $X(t)$  is the “random noise component”.

In general, one can observe only the values  $Y(t)$ . Therefore, our first goal is to estimate the deterministic part  $a(t)$ , and extract the random noise component  $X(t) = Y(t) - a(t)$ . Our second goal is to find a satisfactory probabilistic model for the process  $X(t)$ , analyze its properties, and use it together with  $a(t)$  to predict  $Y(t)$ .

Let  $\{y(t_i), t_i\}, i = 0, 1, \dots, n$  be a sample of  $Y(t)$ . We assume that  $a(t)$  is a polynomial of degree  $p$  in  $t$ , i.e.,

$$(3.3) \quad a_p(t) = \rho_0 + \rho_1 t + \rho_2 t^2 + \dots + \rho_p t^p$$

where  $p$  is a non-negative integer and  $\boldsymbol{\rho} = [\rho_0; \rho_1; \dots; \rho_p]$  is the vector of coefficients. A more general structure can also be used. The vector  $\boldsymbol{\rho}$  can be estimated using the least squares method. More precisely, we choose a vector  $\boldsymbol{\rho}$  that minimizes  $\sum_{i=0}^n (y(t_i) - a(t_i))^2$ , set  $x(t_i) = y(t_i) - a(t_i)$  and treat  $\{x(t_i), t_i\}$  as a sample of  $X(t)$ . Since fitting the parametric polynomial  $a_p(t)$  to data using the least squares method is a straightforward task, we will turn to developing a statistical model for analyzing  $X(t)$ .

We propose two parametric statistical models to analyze  $X(t)$ . The first model is a direct extension of the classical AR(1) model given in Definition 3 and assumes that  $X(t)$  is stationary. To estimate the

parameters we use the least squares method and the maximum likelihood approach.

The second model does not assume that  $X(t)$  is stationary. We model  $X(t + \Delta)$  as a function of  $t$ ,  $\Delta$ , and  $X(t)$ . In particular, we let both the mean and the variance of  $X(t + \Delta)$  depend on  $t$ ,  $\Delta$  and  $X(t)$ . To estimate the parameters of this model in a closed form we use the maximum likelihood approach. To better present the intuition behind this model, we first give a simple model and then generalize it.

#### 4 A statistical model for stationary $X(t)$

Suppose that  $X(t)$  obtained from  $Y(t)$  after removing the trend component  $a_p(t)$  (Equation 3.2) is a stationary process. We define an irregularly sampled stationary AR(1) (IS-AR(1)) process as follows.

**Definition 4.** (IS-AR(1) PROCESS) *A time series  $\{X(t_i), t_i\}$  is an irregularly sampled stationary AR(1) (IS-AR(1)) time series if  $\{X(t)\}$  is stationary and if for every  $t$  and  $\Delta \geq 0$ ,*

$$(4.4) \quad X(t + \Delta) = \theta^\Delta X(t) + \sigma_\Delta \epsilon_{t+\Delta},$$

where  $\epsilon_t \sim \mathcal{N}(0, 1)$  and  $\mathbf{cov}(\epsilon_t, \epsilon_s) = 0$  for every  $t \neq s$  and  $\sigma_\Delta^2 = \sigma^2 \left( \frac{1 - \theta^{2\Delta}}{1 - \theta^2} \right)$  for some  $\sigma > 0$ .

Note that if we compare the process given by equation (4.4) with the one given in Definition 2.1 and observing  $\sigma \sum_{j=0}^{h-1} \theta^j \epsilon_{t+1+j} \sim N(0, \sigma^2 \left( \frac{1 - \theta^{2h}}{1 - \theta^2} \right))$ , we can see that these processes are the same if  $\Delta = h$ . We can recover the original AR(1) process if each  $t_i$  of the IS-AR(1) process is a positive integer and  $t_{i+1} = t_i + 1$  for all  $i$ . Thus IS-AR(1) is an extension of AR(1) to irregularly sampled time series.

**4.1 Parameter estimation** In this section, we show how to estimate parameters  $\theta$  and  $\sigma$ , given a set of observations  $\{x(t_0), x(t_1), \dots, x(t_n)\}$  of  $X(t)$ .

Define  $\Delta_i = t_{i+1} - t_i$  for all  $i = 0, \dots, n-1$ . We may assume without loss of generality that  $\Delta_i > 0$  for all  $i$ . Furthermore, we can assume that all  $\Delta_i \geq 1$ , otherwise we rescale each  $\Delta_i$  by  $\min_i \{\Delta_i\}$ .

Since  $\mathbf{E}[\epsilon_t] = 0$  and  $\mathbf{cov}(\epsilon_t, \epsilon_s) = 0$  for all  $t \neq s$ , we can estimate  $\theta$  by the least squares method. In particular, we need to find  $\hat{\theta} \in (-1, 1)$  minimizing  $\sum_{i=0}^{n-1} (X(t_{i+1}) - \theta^{\Delta_i} X(t_i))^2$ . Since  $\Delta_i \geq 1$  for all  $i$ , this sum is a convex function of  $\theta$  that can be minimized very efficiently using convex optimization.

To estimate  $\sigma$ , we set  $z_i = x(t_{i+1}) - (\hat{\theta})^{\Delta_i} x(t_i)$ . By Definition 4, we have  $z_i \sim N(0, \sigma_{\Delta_i}^2)$  where  $\sigma_{\Delta_i}^2 = \sigma^2 \left( \frac{1 - (\hat{\theta})^{2\Delta_i}}{1 - (\hat{\theta})^2} \right)$ . We therefore estimate  $\sigma$  by maximizing the Gaussian likelihood of the residuals

$z(t_0), \dots, z(t_{n-1})$  at times  $t_0, t_1, \dots, t_{n-1}$ . The maximum likelihood estimator of  $\sigma$  is given by

$$(4.5) \quad \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} \frac{(x(t_{i+1}) - (\hat{\theta})^{\Delta_i} x(t_i))^2}{\rho_i}},$$

where  $\rho_i = \left( \frac{1 - \hat{\theta}^{2\Delta_i}}{1 - \hat{\theta}^2} \right)$  for all  $i$ . The proof of this statement is very similar to the proof of Proposition 5.1, and is omitted.

**4.2 Prediction using the IS-AR(1) model** In this section, first we establish conditions for  $X(t_0)$  under which  $X(t)$  is a stationary process. Then assuming the stationarity of  $X(t)$ , we derive equations for the one-step prediction and the auto-covariance function. Without of generality we will assume  $t_0 = 0$ .

Using Equation 4.4, independence of  $\epsilon_t$  and  $X(0)$ , and the fact that  $\mathbf{E}[\epsilon_t] = 0$  for all  $t$ , we can express  $\mathbf{E}[X(t)]$ ,  $\mathbf{var}[X(t)]$ , and  $\mathbf{cov}[X(t), X(t + \Delta)]$  in terms of  $X(0)$  as follows.

$$(4.6) \quad \mathbf{E}[X(t)] = \theta^t \mathbf{E}[X(0)]$$

$$(4.7) \quad \mathbf{var}[X(t)] = \theta^{2t} \mathbf{var}[X(0)] + \sigma^2 \frac{1 - \theta^{2t}}{1 - \theta^2}$$

$$(4.8) \quad \mathbf{cov}[X(t), X(t + \Delta)] = \theta^\Delta \mathbf{var}[X(0)]$$

**PROPOSITION 4.1.** *Assume that  $\mathbf{E}[X(t)^2] < \infty$  and  $\mathbf{E}[X(0)] = 0$ . Then  $X(t)$  in Definition 4 is a stationary process if  $\mathbf{var}[X(0)] = \frac{\sigma^2}{1 - \theta^2}$  and the auto-covariance function of  $X(t)$  is  $\mathbf{cov}_X(\Delta) = \mathbf{cov}[X(t), X(t + \Delta)] = \theta^\Delta \frac{\sigma^2}{1 - \theta^2}$ .*

*Proof.* For  $\mathbf{var}[X(0)] = \frac{\sigma^2}{1 - \theta^2}$ , Equation 4.7 gives

$$\mathbf{var}[X(t)] = \theta^{2t} \frac{\sigma^2}{1 - \theta^2} + \sigma^2 \frac{1 - \theta^{2t}}{1 - \theta^2} = \frac{\sigma^2}{1 - \theta^2},$$

yielding

$$\mathbf{cov}[X(t), X(t + \Delta)] = \theta^\Delta \mathbf{var}[X(t)] = \theta^\Delta \frac{\sigma^2}{1 - \theta^2}.$$

Since  $\mathbf{cov}[X(t), X(t + \Delta)]$  does not depend on  $t$ ,  $X(t)$  is stationary. ■

A one-step predictor of  $X(t + \Delta)$  given  $X(t)$  for any  $\Delta > 0$  is given by the conditional expectation of  $X(t + \Delta)$  (using Equation 4.6):

$$\hat{X}(t + \Delta) = \mathbf{E}[X(t + \Delta) | X(t)] = \theta^\Delta X(t).$$

**4.3 Analyzing  $Y(t)$  with a Stationary Component  $X(t)$ :** Algorithm 1 can be used for estimating the auto-covariance function of irregularly sampled time series  $Y(t)$  and for predicting  $Y(t + \Delta)$  given  $Y(t)$ .

---

**Algorithm 1** Auto-covariance function and one-step prediction

---

**Given:**  $\{Y(t_i), t_i\}$  with  $t_i < t_{i+1}$  for  $i = 0, \dots, n-1$ , time interval  $\delta > 0$ .

**Output:** auto-covariance function  $\mathbf{cov}_Y(\Delta)$ ; prediction  $\hat{Y}(t_n + \delta)$ .

- (1) Set  $\Delta_i = t_{i+1} - t_i$  for  $i = 0, 1, \dots, n-1$ .
- (2) Fit a polynomial function  $a_p(t)$  to  $Y(t)$ : Find the coefficients of  $a_p(t)$  minimizing  $\sum_{i=0}^n (Y(t_i) - a_p(t_i))^2$  for  $p = 0, \dots, P$  where  $P$  is some upper bound on the degree of the polynomial. Choose  $a_p$  with the best fit, and set  $X(t_i) = Y(t_i) - a(t_i)$ .
- (3) Estimate  $\theta$  as  $\hat{\theta} = \arg \min_{\theta} \sum_{i=0}^{n-1} (X(t_{i+1}) - \theta^{\Delta_i} X(t_i))^2$ , and  $\sigma$  using  $\hat{\sigma}$  in Equation (4.5).
- (4) Since  $a_p(t)$  is deterministic, set  $\mathbf{cov}_Y(\Delta) = \mathbf{cov}_X(\Delta) = \theta^{\Delta} \frac{\sigma^2}{1 - \theta^{\Delta}}$ .
- (5) Prediction:

$$\begin{aligned} \hat{Y}(t + \Delta) &= a_p(t + \Delta) + \hat{X}(t + \Delta) \\ &= a_p(t + \Delta) + \mathbf{E}[X(t + \Delta) | X(t)] \\ &= a_p(t + \Delta) + \theta^{\Delta} X(t). \end{aligned}$$

---

## 5 A model for non-stationary $X(t)$

The model introduced in Definition 4 can be used if  $X(t)$  is stationary. Mean and variance of a stationary process are time independent; and the covariance between any two observations  $X(t)$  and  $X(s)$  depends only on their time difference  $|t - s|$ . This allowed us to derive a simple expression for the auto-covariance function. In practice, however, one may not have stationarity all the time. If  $X(t)$  is not stationary, its statistical properties may vary with time. Thus it is natural to focus on estimating these properties in the near future instead of trying to obtain some global, time-independent values. To achieve this goal, we model  $X(t + \Delta)$  as a function of  $t$ ,  $\Delta$ , and  $X(t)$ , plus a random noise whose variance also depends on  $t$ ,  $\Delta$ , and  $X(t)$ .

As before, after fitting a polynomial of degree  $p$

to  $Y(t)$  we get  $X(t) = Y(t) - a_p(t)$ . We first introduce a simple model for  $X(t)$  to show the intuition behind the general model, and then present the general model.

**5.1 A simple IN-AR(1) process** We define a simple, non-stationary, irregularly sampled AR(1) process that allows us to model  $X(t + \Delta)$  as a function of only  $X(t)$  and  $\Delta > 0$ :

$$(5.9) \quad X(t + \Delta) = X(t) - \theta(X(t) - c)\Delta + \sqrt{\Delta}\sigma\epsilon_{t+\Delta},$$

where  $\epsilon_{t+\Delta} \sim \mathcal{N}(0, 1)$  and  $\mathbf{cov}(\epsilon_t, \epsilon_s) = 0$  for all  $t \neq s$ . Here  $\theta > 0$ ,  $\sigma > 0$ , and  $c$  are the unknown parameters that will be estimated from observations. The above expression implies that if  $\Delta$  is sufficiently small, then  $X(t + \Delta)$  can be approximated by a linear function of  $\Delta$ . Such approximation is very common when estimating parameters of continuous time processes.

The stochastic process given in Equation 5.9 is a mean reverting process with mean  $c$ , mean reversion parameter  $\theta$ , and variance  $\sigma$ . If  $X(t) > c$  (i.e., the value of  $X(t)$  is above the mean reversion level), the drift term  $-\theta(X(t) - c)\Delta$  is negative for all  $\Delta > 0$ , and the value of the process will tend to decrease. If  $X(t) < c$ , the drift term  $-\theta(X(t) - c)\Delta$  is positive, resulting in an upward influence on  $X$ . This way, if the process drifts away from its mean  $c$ , it will eventually be pulled back towards the mean  $c$ , at a speed determined by the mean reversion rate  $\theta$ . The term  $\sqrt{\Delta}\sigma\epsilon_{t+\Delta}$  adds some noise to this process.

Figure 1 gives an example of a mean reverting process with parameters  $c = 0.5$ ,  $a = 2.5$ ,  $\sigma = 0.01$ .

Mean reverting stochastic processes are widely used in practice to model interest rates, energy prices, option pricing, heat transfer, and many other processes. The model is very intuitive and supported by our observations about these processes.

**5.2 A Generalization of the IN-AR(1) Process** In model (5.9) the value of  $X(t + \Delta)$  given  $X(t)$  is a linear function of only  $X(t)$  and  $\Delta$ . In practice, however, time series may have more complicated dependences. For example, since  $X(t)$  may be non-stationary,  $X(t + \Delta)$  may depend on  $t$  or  $t + \Delta$ , or – since the data is irregularly sampled,  $\Delta$  may be large, so the linear approximation we used in 5.9 may not be adequate. To handle more general cases, we introduce the following (still linear in  $\theta$  and  $\sigma$ ) model.

Let  $\theta \in \mathbf{R}^m$  be an  $m$ -dimensional drift parameter vector and  $\sigma \in \mathbf{R}$  be a scalar variance parameter. Let  $\alpha(\Delta, t, X(t)) : [0, \infty) \times [0, \infty) \times \mathbf{R} \rightarrow \mathbf{R}^m$  be

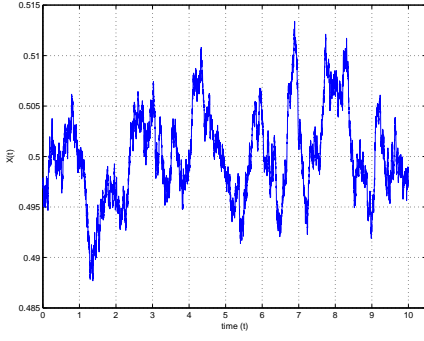


Figure 1: Mean reverting process with  $c = 0.5$ ,  $a = 2.5$ ,  $\sigma = 0.01$

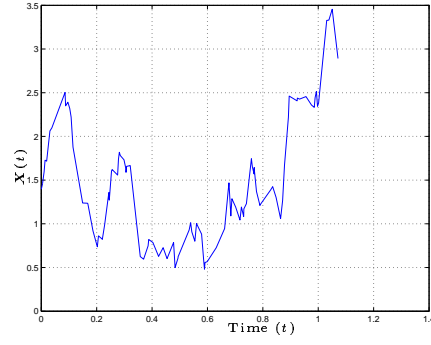


Figure 2: A general dependence on  $X(t)$  and  $\Delta$

a function of  $\Delta$ ,  $t$ , and  $X(t)$ . Similarly, define  $\beta(\Delta, t, X(t)) : [0, \infty) \times [0, \infty) \times \mathbf{R} \rightarrow \mathbf{R}$ .

**Definition 5.** (GENERAL IN-AR(1) PROCESS) *A general irregularly sampled non-stationary time series (General IN-AR(1)) process is defined as*

$$X(t + \Delta) = X(t) + \boldsymbol{\theta}^T \boldsymbol{\alpha}(\Delta, t, X(t)) + \sigma \beta(\Delta, t, X(t)) \epsilon_{t+\Delta},$$

where  $\epsilon_{t+\Delta} \sim \mathcal{N}(0, 1)$ ,  $\text{cov}(\epsilon_t, \epsilon_s) = 0$  for all  $t \neq s$ ,  $\boldsymbol{\theta}$  is the vector of drift parameters,  $\boldsymbol{\alpha}(\Delta, t, X(t))$  is the drift function,  $\sigma$  is the variance parameter, and  $\beta(\Delta, t, X(t))$  is the variance function. In addition, if  $\Delta = 0$  then the functions  $\boldsymbol{\alpha}$  and  $\beta$  satisfy

$$\boldsymbol{\alpha}(0, t, X(t)) = \mathbf{0} \text{ and } \beta(0, t, X(t)) = 0.$$

Since the above condition is the only assumption on the structure of  $\boldsymbol{\alpha}$  and  $\beta$ , the model covers a wide range of irregularly sampled time series.

Observe that if we set  $\boldsymbol{\theta} = [a; ac]$ ,  $\boldsymbol{\alpha}(\Delta, t, X(t)) = [-X(t)\Delta; \Delta]$ , and  $\beta(\Delta, t, X(t)) = \sqrt{\Delta}$  in Definition 5, we get the simple model given by (5.9).

**EXAMPLE 1. [A general dependence on  $X(t)$  and  $\Delta$ ]** Consider an irregularly sampled time series given in Figure 2. This time series is not stationary and does not have any specific pattern. Also, it has a number of jumps with different magnitude. However, the time series in Figure 2 is a sample path of  $X(t)$  whose state space equation is

$$X(t + \Delta) = X(t) + 0.5X(t)\Delta - 0.01\Delta + 1.4\sqrt{X(t)}\Delta^2 + 2\sqrt{\Delta}\epsilon_{t+\Delta} \quad (5.10)$$

Note that if we set  $\boldsymbol{\theta} = [0.5; -0.01; 1.4]$ ,  $\boldsymbol{\alpha}(\Delta, t, X(t)) = [X(t)\Delta; \Delta; \sqrt{X(t)}\Delta^2]$ , and  $\sigma = 2$ ,  $\beta(\Delta, t, X(t)) = \sqrt{\Delta}$  in Definition (5) we get (5.10).

These examples show that a wide range of irregularly sampled time series can be modeled as an IN-AR(1) process.

**5.3 Parameter Estimation** We first show how to find the maximum likelihood estimators of  $\boldsymbol{\theta}$  and  $\sigma$  for given  $\boldsymbol{\alpha}(\Delta, t, X(t))$  and  $\beta(\Delta, t, X(t))$ , and then discuss how to appropriately select  $\boldsymbol{\alpha}(\Delta, t, X(t))$  and  $\beta(\Delta, t, X(t))$ .

Let  $\{x(t_i), i = 0, \dots, n\}$  be the values of  $X(t)$  at times  $0 \leq t_0 < t_1 < \dots < t_n$  and define  $\Delta_i = t_{i+1} - t_i$  for  $i = 0, \dots, n-1$ . Since  $\epsilon_t$  and  $\epsilon_s$  are independent for all  $t \neq s$ , the distribution of  $X(t_{i+1})$  given  $X(t_i)$  is normal with mean  $X(t_i) + \boldsymbol{\theta}^T \boldsymbol{\alpha}(\Delta_i, t_i, X(t_i))$  and variance  $\beta^2(\Delta_i, t_i, X(t_i))\sigma^2$ . Therefore,  $\boldsymbol{\theta}$  and  $\sigma$  can be estimated by maximizing the Gaussian likelihood of observations  $x(t_1), \dots, x(t_n)$  at times  $t_1, \dots, t_n$ .

**PROPOSITION 5.1.** *The maximum likelihood estimators of  $\boldsymbol{\theta}$  and  $\sigma$  are given by*

$$\hat{\boldsymbol{\theta}} = \left( \sum_{i=1}^n \frac{\boldsymbol{\alpha}(\Delta_i, t_i, x(t_i)) \boldsymbol{\alpha}^T(\Delta_i, t_i, x(t_i))}{\beta^2(\Delta_i, t_i, x(t_i))} \right)^{-1} \cdot \sum_{i=1}^n \frac{(x(t_{i+1}) - x(t_i)) \boldsymbol{\alpha}(\Delta_i, t_i, x(t_i))}{\beta^2(\Delta_i, t_i, x(t_i))},$$

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{(x(t_{i+1}) - x(t_i) - \hat{\boldsymbol{\theta}}^T \boldsymbol{\alpha}(\Delta_i, t_i, x(t_i)))^2}{\beta^2(\Delta_i, t_i, x(t_i))}}.$$

*Proof.* See Appendix A.

We suggest the following strategy for choosing  $\boldsymbol{\alpha}(\Delta, t, X(t))$  and  $\beta(\Delta, t, X(t))$ . First, a set of candidate functions is chosen, by gathering more information about the data generating process, plotting the data, and/or using expert views. In fact, this should

be the first step in any statistical analysis. After identifying a set of possible functions, we use Proposition 5.1 to estimate the parameters for each one. Finally, we choose the pair that fits the data best, i.e., has the lowest mean squared error.

**5.4 Prediction using the general IN-AR(1) model** We assumed that  $Y(t)$ , the time series we are analyzing, can be decompsed as  $Y(t) = a_p(t) + X(t)$ . Since  $a_p(t)$  is deterministic, we only need to predict  $X(t + \Delta)$ . In particular, we need to estimate the following terms.

$$\begin{aligned}\widehat{Y}(t + \Delta) &= \mathbf{E}[Y(t + \Delta)|Y(t)] \\ &= a_p(t + \Delta) + \mathbf{E}[X(t + \Delta)|X(t)], \\ \mathbf{var}[Y(t + \Delta)|Y(t)] &= \mathbf{var}[X(t + \Delta)|X(t)], \\ \mathbf{cov}[Y(t + \Delta_1 + \Delta_2), Y(t + \Delta_1)|Y(t)] \\ &= \mathbf{cov}[X(t + \Delta_1 + \Delta_2), X(t + \Delta_1)|X(t)].\end{aligned}$$

Since we did not assume that  $X(t)$  is stationary, we might not have a time independent expression for the mean, variance, and the auto-covariance function. Therefore, we are interested in estimations about the near future given the current value of the process.

Using Definition 5, the independence of  $\epsilon_{t+\Delta}$  and  $X(t)$ , and the assumption that  $\mathbf{E}[\epsilon_t] = 0$  for all  $t$ , we can write the conditional expectation of  $X(t + \Delta)$  given  $X(t)$  as

$$\begin{aligned}\mathbf{E}[X(t + \Delta)|X(t)] &= X(t) + \boldsymbol{\theta}^T \boldsymbol{\alpha}(\Delta, t, X(t)) + \\ &\quad \sigma \beta(\Delta, t, X(t)) \mathbf{E}[\epsilon_{t+\Delta}] \\ (5.11) \quad &= X(t) + \widehat{\boldsymbol{\theta}}^T \boldsymbol{\alpha}(\Delta, t, X(t))\end{aligned}$$

and the conditional variance as

$$\begin{aligned}\mathbf{var}[X(t + \Delta)|X(t)] &= \mathbf{E}\left[(X(t + \Delta) - \right. \\ &\quad \left. - \mathbf{E}[X(t + \Delta)|X(t)])^2 | X(t)\right] = \sigma^2 \beta^2(\Delta, t, X(t))\end{aligned}$$

Let  $\Delta_1, \Delta_2 > 0$ . The conditional covariance between  $X(t + \Delta_1 + \Delta_2)$  and  $X(t + \Delta_1)$  given  $X(t)$  is

$$\begin{aligned}\mathbf{cov}[X(t + \Delta_1 + \Delta_2), X(t + \Delta_1)|X(t)] \\ &= \sigma^2 \beta^2(\Delta, t, X(t)) \\ &\quad + \mathbf{E}[\boldsymbol{\theta}^T \boldsymbol{\alpha}(\Delta_2, t + \Delta_1, X(t + \Delta_1)) X(t + \Delta_1) | X(t)] \\ &\quad - (X(t) + \widehat{\boldsymbol{\theta}}^T \boldsymbol{\alpha}(\Delta, t, X(t))) \cdot \\ (5.12) \quad &\quad \cdot \mathbf{E}[\boldsymbol{\theta}^T \boldsymbol{\alpha}(\Delta_2, t + \Delta_1, X(t + \Delta_1)) | X(t)]\end{aligned}$$

It can be further simplified using the structure of  $\boldsymbol{\alpha}$ . The expressions inside the expectation operators are functions of  $\epsilon_{t+\Delta_1}$ , and thus are independent of  $X(t)$ , so the operators can be removed. Finally,  $\mathbf{cov}[X(t + \Delta_1 + \Delta_2), X(t + \Delta_1)|X(t)]$  is a function of  $\boldsymbol{\alpha}, \beta, \boldsymbol{\theta}, \sigma$  and  $X(t)$ . Since  $X(t)$  is given, if we replace  $\boldsymbol{\theta}$  and  $\sigma$  by their estimators  $\widehat{\boldsymbol{\theta}}$  and  $\widehat{\sigma}$ , we can estimate  $\mathbf{cov}[X(t + \Delta_1 + \Delta_2), X(t + \Delta_1)|X(t)]$ .

A one-step predictor of  $X(t + \Delta)$  given  $X(t)$  for any  $\Delta > 0$  is given by (5.11):

$$\begin{aligned}\widehat{X}(t + \Delta) &= \mathbf{E}[X(t + \Delta)|X(t)] \\ &= X(t) + \widehat{\boldsymbol{\theta}}^T \boldsymbol{\alpha}(\Delta, t, X(t)).\end{aligned}$$

**5.5 Analyzing  $Y(t)$  with a non-stationary component  $X(t)$**  The following algorithm can be used to for estimating the auto-covariance function of irregularly sampled time series  $Y(t)$  and for predicting  $Y(t + \Delta)$  given  $Y(t)$ .

---

#### Algorithm 2

---

**Given:**  $\{Y(t_i), t_i\}, i = 0, 1, \dots, n$ , with  $t_i < t_{i+1}$ ;  
values of  $\delta_1, \delta_2$ ;

functions  $\boldsymbol{\alpha}(\Delta, t, X(t)), \beta(\Delta, t, X(t))$

**Output:** a predictor  $\widehat{Y}(t_n + \delta_1)$ ,  
an estimate of  $\mathbf{var}[Y(t_n + \Delta)|Y(t_n)]$  and  
 $\mathbf{cov}[Y(t + \delta_1 + \delta_2), Y(t + \delta_1)|Y(t)]$ .

- (1) Set  $\Delta_i = t_{i+1} - t_i$  for  $i = 0, 1, \dots, n - 1$ .
- (2) Fit a polynomial  $a_p(t)$  to  $Y(t)$  as in Algorithm 1.
- (3) Estimate  $\widehat{\boldsymbol{\theta}}$  and  $\widehat{\sigma}$  using Proposition 5.1.
- (4) Prediction:  

$$\begin{aligned}\widehat{Y}(t + \delta_1) &= a_p(t + \delta_1) + \widehat{X}(t + \delta_1) \\ &= a_p(t + \delta_1) + X(t) + \widehat{\boldsymbol{\theta}}^T \boldsymbol{\alpha}(\delta_1, t, X(t)).\end{aligned}$$
- (5) Estimate variance and covariance:

$$\begin{aligned}\mathbf{var}[Y(t + \delta)|Y(t)] &= \widehat{\sigma}^2 \beta^2(\delta, t, X(t)), \\ \mathbf{cov}[Y(t + \delta_1 + \delta_2), Y(t + \delta_1)|Y(t)] &= \\ &\quad \mathbf{cov}[X(t + \delta_1 + \delta_2), X(t + \delta_1)|X(t)],\end{aligned}$$

where the last equation can be estimated using Equation 5.12.

## 6 Resampling

Let  $X(t) = Y(t) - a_p(t)$  as in the previous sections. In this section, we propose a model to resample a regularly sampled  $\widehat{X}(t)$  from irregular time series  $\{X(t_i), t_i, i = 0, 1, \dots, n\}$ . Once we obtain a regularly

sampled version  $\hat{X}(t)$ , we can use techniques developed for the regular time series, and analyze multiple irregularly sampled time series.

We illustrate our approach using the stationary model introduced in Section 4. This resampling approach can easily be modified for the non-stationary model of Section 5. Before we present details of our resampling algorithm, we give an example illustrating the main idea.

Suppose that we are given the values  $X(0)$  and  $X(4)$  and we want to construct a regularly sampled version  $\hat{X}(t)$  of  $X(t)$  for  $t \in \{0, 1, 2, 3, 4\}$ . We set  $\hat{X}(0) = X(0)$  and  $\hat{X}(4) = X(4)$ . According to the model in Equation (4.4),

$$X(t+1) = \hat{\theta}X(t) + \hat{\sigma}_1\epsilon_{t+1} \quad \text{for } i \in \{0, 1, 2, 3\},$$

where  $\hat{\theta}$  and  $\hat{\sigma}$  are the estimators of  $\theta$  and  $\sigma$ . Since we only have  $X(0)$  and  $X(4)$ , we cannot estimate the noises  $\epsilon_t$ ,  $t \in \{1, 2, 3, 4\}$ . Instead, we estimate an auxiliary error  $\varepsilon$  and set  $\epsilon_t = \varepsilon$ , for  $t \in \{1, 2, 3, 4\}$ . After replacing  $\epsilon_t = \varepsilon$  for  $t \in \{1, 2, 3, 4\}$  in the above equation we write  $X(4)$  in terms of  $X(0)$  as

$$X(4) = \hat{\Delta}^4 X(0) + \varepsilon \hat{\sigma} \sum_{j=0}^3 \hat{\theta}^j,$$

which implies that

$$\varepsilon = \frac{X(4) - \hat{\Delta}^4 X(0)}{\hat{\sigma} \sum_{j=0}^3 \hat{\theta}^j}.$$

Then we set recursively

$$\hat{X}(t+1) = \hat{\theta}\hat{X}(t) + \hat{\sigma}\varepsilon \quad \text{for } t \in \{0, 1, 2, 3\}.$$

Note that the above recursion sets  $\hat{X}(4) = X(4)$  therefore values of the resampled process  $\hat{X}(t)$  coincides with the values of the irregular time series at sampling times.

Let  $\{X(t_i), t_i\}$  be an irregular time series sample. Without loss of generality we assume that  $\{t_i\}$  are integers and suppose we want to construct a regularly sampled version  $\{\tilde{X}(s_k), s_k\}$ ,  $k = 0, \dots, N$  with sampling times  $s_k$  satisfying  $s_0 = t_0$ ,  $s_N = t_n$ . Since we want  $\{\tilde{X}(s_k), s_k\}$  be a regular time series we require that  $s_{k+1} = s_k + 1$  for all  $k = 0, \dots, N - 1$ . We also want the values of the resampled time series and the original time series be the same at the times  $\{t_i\}$ ,  $i = 0, 1, \dots, n$ .

---

### Algorithm 3

---

**Given:** Irregularly sampled time series  $\{X(t_i), t_i\}$ .

**Output:** Regularly sampled time series  $\{\tilde{X}(s_k), s_k\}$ .

---

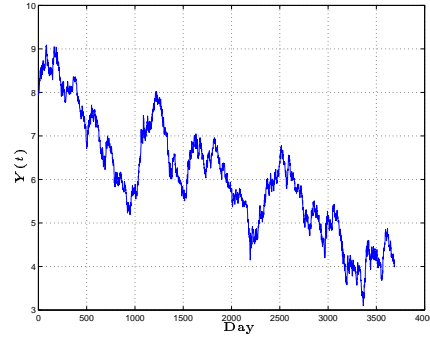


Figure 3: 10 year treasury bond data

- (1) Set  $k = 0$ ,  $s_k = t_0$  and  $\tilde{X}(s_0) = X(t_0)$ .
- (2) FOR  $i = 0$  TO  $n - 1$  DO
  - (2.1) Set  $\delta = t_{i+1} - t_i$ ,  $\varepsilon = \frac{X(t_{i+1}) - \hat{\theta}^\delta X(t_i)}{\hat{\sigma} \sum_{j=0}^{\delta-1} \hat{\theta}^j}$ .
  - (2.2) FOR  $j = 1$  to  $\delta - 1$  DO
    - (2.2.1)  $s_{k+1} = s_k + 1$
    - (2.2.2) Set  $\tilde{X}(s_{k+1}) = \hat{\theta}\tilde{X}(s_k) + \hat{\sigma}\varepsilon$
    - (2.2.3)  $k = k + 1$
  - END FOR
- END FOR

## 7 Computational Experiments

### 7.1 Computational experiments with IS-AR(1)

We analyzed daily prices of 10 year Treasury bonds between January 2, 1990 and September 27, 2004. The regular time series data  $Y(t)$  is given in Figure 3. We estimated parameters of the regular time series by fitting a polynomial of degree 1 and assuming that the residuals follow a regular AR(1) process. The estimated polynomial was  $a_1(t) = \rho_0 + \rho_1 t = 8.0450 - 0.0011t$  and the parameter values were  $\hat{\theta}_r = 0.9955$  and  $\hat{\sigma}_r = 0.0602$ . The subscripts  $r$  denote that these parameters are estimated using regular AR(1) process. Then we randomly removed points from this data to generate an irregularly sampled time series and estimated the parameters using techniques introduced in Section 4 to compare them with the parameters estimated by regular time series. In Figures 4 and 5 we show value of the estimators as a function of the data points of the irregular time series and compare them with  $\hat{\theta}_r$ ,  $\hat{\sigma}_r$ . (the lines in the corresponding graphs represent  $\hat{\theta}_r$ , and  $\hat{\sigma}_r$ ). As seen in the Figure 4, the value of the estimator  $\hat{\theta}$  is very close to the one estimated by regular time series for all irregular time series. However,  $\hat{\theta}$

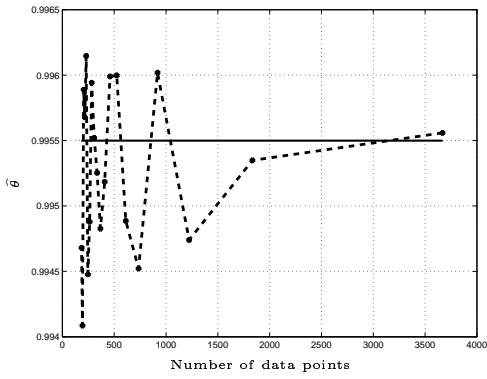


Figure 4: Estimators of  $\theta$

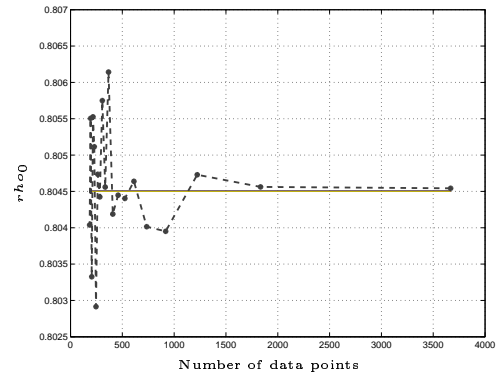


Figure 6: Estimators of  $\rho_0$

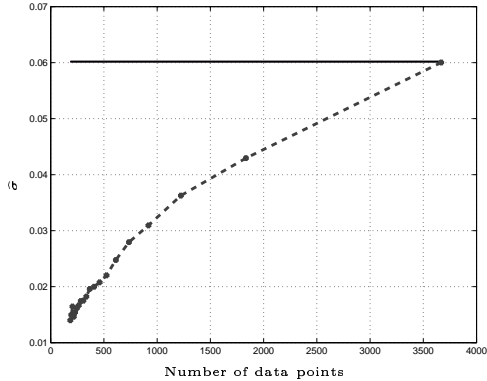


Figure 5: Estimators of  $\sigma$

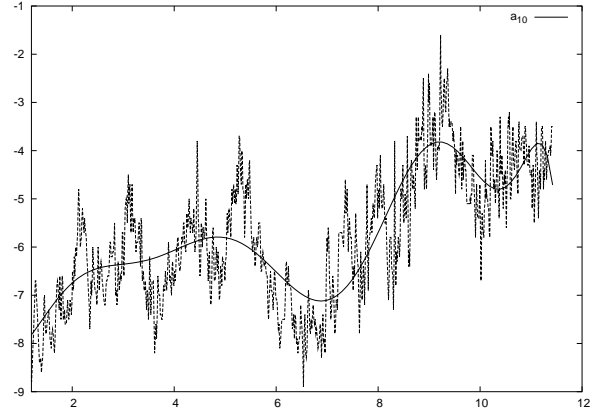


Figure 7: Fitting a polynomial to the data

decreases as the number of data points decrease, Figure 4. We also present the estimator of the constant term of  $a_1(t)$ ,  $\rho_0$  in Figure 6. Since  $\rho_1$  was estimated perfectly we do not show its estimator to save space.

**7.2 Computational experiments with IN-AR(1)** We tested prediction abilities of our IN-AR(1) model on several real datasets. Figure 7 shows a dataset containing certain historical isotopic temperature record from the Vostok ice core, due to Petit et al. [9]. The dataset contains about 1K irregularly sampled points. The figure also shows a 10-degree polynomial fit to the data. Figure 9 overlays the dataset with a 10-point prediction given by the model (see Algorithm 1). More precisely, given a history of 100 points, we estimate the model and use it to predict the next 10 points (supplying it with the desired delta values), then we advance by 10 points, and repeat. For comparison, we did a similar prediction using a vanilla algorithm that always predicts the last value it sees. The corresponding curve is given in Figure 8. The vanilla algorithm produces a

step function with a good overall fit to the data, but with no attempt to give accurate short-term predictions. The curve produced by the IN-AR(1) model provides a smoother, much more accurate fit to the data. However, the model appears to be more conservative about predicting spikes, while the vanilla algorithm has a better chance of exploring them just because it may be randomly seeded with a spiked value.

## 8 Conclusion

In this paper, we addressed the issue of analyzing irregularly sampled time series. We developed two auto-regression order 1 (AR(1)-type) models. Our first model assumed that the irregularly sampled time series is stationary, and our second model relaxed this assumption. Then we discussed how to estimate parameters of these models very efficiently and how to utilize them to conduct prediction. We showed that under the assumption of stationarity, our first model

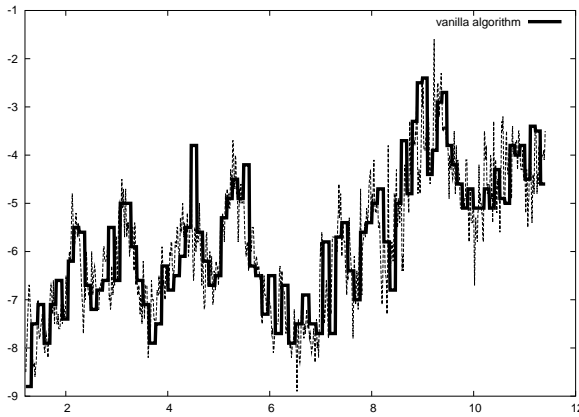


Figure 8: A 10-step vanilla prediction

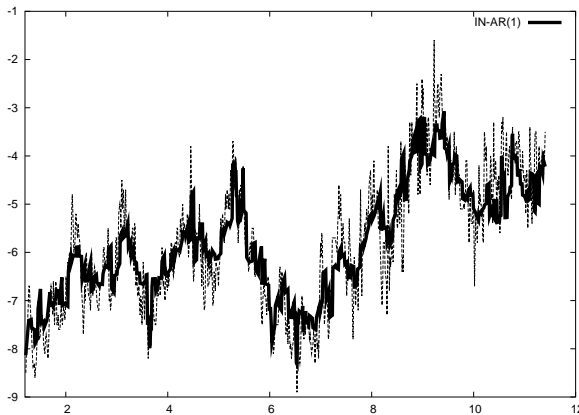


Figure 9: A 10-step prediction with the proposed model

can be used to derive the auto-covariance function of the irregularly sampled time series. We also suggested an interpolation like resampling algorithm to construct a regularly sampled version of the irregularly sampled time series. This resampling algorithm can be used for analyzing multiple time series by constructing a regularly sampled version of each of the time series first and then using the techniques developed for regularly sampled time series.

Our key idea in this paper is to model the value of an observation as a function of the observation just before it and the time difference between these two observations. Then assuming Gaussian noises we derived the maximum likelihood estimators of the model parameters. For prediction, we used conditional expectation arguments. We demonstrated our algorithms on the real-world data. Our computational experiments showed that the models introduced in this paper can accurately estimate the model parameters.

Finally, we note that our approach can be extended to higher order auto-regression processes (AR(p)), to moving average processes, (MA(q)) and to autoregressive moving average processes, (ARMA(p,q)). One of the open questions is how one can develop a model to analyze irregularly sampled time series when the random noise is not Gaussian. These are some of the subjects of our future research.

## References

- [1] H.M. Adorf. Interpolation of irregularly sampled data series—a survey. In Shaw R.A, Payne H.E., and Hayes J.J.E., editors, *Astronomical data analysis software and systems IV*, volume 77 of *APS conference series*, 1995. Available at <http://www.stsci.edu/stsci/meetings/adassIV/adorfhm2.html>.
- [2] Y. Ait-Sahalia. The effects of random and discrete sampling when estimating continuous time diffusions. *Econometrica*, 71(2):483–549, March 2003.
- [3] P.J. Brockwell. Lévy driven carma processes. *Ann.Inst.Statist.Math.*, 53(1):113–124, 2001.
- [4] P.J. Brockwell and R.A. Davis. *Time series: Theory and Methods*. Springer-Verlag Series in Statistics, 1991.
- [5] R. H. Jones and P. V. Tryon. Continuous time series models for unequally spaced data applied to modelling atomic clocks. *SIAM J. Sci. Stat. Comput.*, 4(1):71–81, January 1987.
- [6] R.H. Jones. *Time series analysis with unequally spaced data*. Handbook of statistics 5 (eds. Hannan, E.J. and Krishnaiah, P.R. and Rao, M.M.), pages 157–178. North Holland, Amsterdam, 1985.
- [7] R.E. Kalman. A new approach to linear filtering and prediction problems. *ASME Trans. Part D (J.Basic Engineering)*, 83:35–45, 1960.

- [8] J.H. Krolik. *in Statistical Challenges in Modern Astronomy, eds. E.D. Feigelson & G.J. Babu*, page 349. Springer-Verlag, New York, 1992.
- [9] J. R. Petit, J. Jouzel, D. Raynaud, N.I. Barkov, J.-M. Barnola, I. Basile, M. Bender, J. Chappel-laz, M. Davis, G. Delayque, M. Delmotte, V.M. Kotlyakov, M. Legrand, V.Y. Lipenkov, C. Lorius, L. Pepin, C. Ritz, E. Saltzman, and M. Stievenard. Climate and atmospheric history of the past 420,000 years from the vostok ice core, antarctica. *Nature*, (399):429–436, 1999.
- [10] D.B. Rubin. *Statistical Analysis With Missing Data*. Wiley, New York, 2nd edition, 2002.

nonsingular, then  $\hat{\boldsymbol{\theta}}$  is given by

$$\hat{\boldsymbol{\theta}} = \left( \sum_{i=1}^n \frac{\boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^T}{\beta_i^2} \right)^{-1} \sum_{i=1}^n \frac{(x(t_{i+1}) - x(t_i)) \boldsymbol{\alpha}_i}{\beta_i^2}.$$

If  $\mathbf{A}$  is singular, the system has multiple solutions, and  $\hat{\boldsymbol{\theta}}$  can be set to any such solution. Once we obtain  $\hat{\boldsymbol{\theta}}$ , we can substitute it in the second optimality condition to solve for  $\sigma$ . The estimator  $\hat{\sigma}$  is given by

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{(x(t_{i+1}) - x(t_i) - \boldsymbol{\alpha}_i^T \hat{\boldsymbol{\theta}})^2}{\beta_i^2}}$$

## A Proof of Proposition 5.1

Let  $x(t_1), \dots, x(t_n)$  be the values of  $X$  at  $t_1, t_2, \dots, t_n$  and define  $\Delta_i = t_{i+1} - t_i$  for  $i = 0, \dots, n-1$ . Since  $\epsilon_t$  and  $\epsilon_s$  are independent for any distinct  $t$  and  $s$ , the distribution of  $X(t_{i+1})$  given  $X(t_i)$  is normal with mean  $X(t_i) + \boldsymbol{\alpha}^T(\Delta_i, t_i, X(t_i))\boldsymbol{\theta}$  and with variance  $\beta^2(\Delta_i, t_i, X(t_i))\sigma^2$ . Therefore, the Gaussian likelihood of data is

$$g(x(t_1), \dots, x(t_n); \boldsymbol{\theta}, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \frac{1}{\beta(\Delta_i, t_i, x(t_i))\sigma} \exp \left\{ -\frac{(x(t_{i+1}) - x(t_i) - \boldsymbol{\alpha}^T(\Delta_i, t_i, x(t_i))\boldsymbol{\theta})^2}{2\beta^2(\Delta_i, t_i, x(t_i))\sigma^2} \right\}.$$

Taking the natural logarithm of both sides,

$$\begin{aligned} \mathbf{L}(\boldsymbol{\theta}, \sigma) &= \ln g(x(t_1), \dots, x(t_n); \boldsymbol{\theta}, \sigma) \\ &= -\frac{n}{2} \ln(2\pi) - \sum_{i=1}^n \left[ \ln(\beta(\Delta_i, t, x(t_i))\sigma) \right. \\ &\quad \left. + \frac{(x(t_{i+1}) - x(t_i) - \boldsymbol{\alpha}^T(\Delta, t, x(t_i))\boldsymbol{\theta})^2}{2\beta^2(\Delta_i, t, x(t_i))\sigma^2} \right]. \end{aligned}$$

and the taking partial derivatives of  $\mathbf{L}(\boldsymbol{\theta}, \sigma)$ , equating them to zero

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \mathbf{L}(\boldsymbol{\theta}, \sigma) &= \frac{1}{\sigma^2} \sum_{i=1}^n \frac{(x(t_{i+1}) - x(t_i) - \boldsymbol{\alpha}_i^T \boldsymbol{\theta}) \boldsymbol{\alpha}_i}{\beta_i^2} = \mathbf{0} \\ \nabla_{\sigma} \mathbf{L}(\boldsymbol{\theta}, \sigma) &= \sum_{i=1}^n -\frac{1}{\sigma} + \frac{(x(t_{i+1}) - x(t_i) - \boldsymbol{\alpha}_i^T \boldsymbol{\theta})^2}{\beta_i^2 \sigma^3} = \mathbf{0} \end{aligned}$$

where  $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}^T(\Delta, t, x(t_i))$  and  $\beta_i = \beta^2(\Delta_i, t_i, x(t_i))$ . Solving the resulting equations we get

$$\sum_{i=1}^n \frac{(x(t_{i+1}) - x(t_i)) \boldsymbol{\alpha}_i}{\beta_i^2} = \sum_{i=1}^n \frac{\boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^T}{\beta_i^2} \boldsymbol{\theta}.$$

Then any solution  $\hat{\boldsymbol{\theta}}$  to the above equation is an estimator of  $\boldsymbol{\theta}$ . If the matrix  $\mathbf{A} = \sum_{i=1}^n \frac{\boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^T}{\beta_i^2}$  is