Journal of Biomedical Informatics 76 (2017) 69–77

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Probabilistic broken-stick model: A regression algorithm for irregularly sampled data with application to eGFR

Norman Poh[a,b,*], Santosh Tirunagari[a,c], Nicholas Cole[d], Simon de Lusignan[d]

[a] Department of Computer Science, University of Surrey, UK
[b] QuintilesIMS, London, UK
[c] Surrey Clinical Research Center, Guildford, Surrey, UK
[d] Department of Clinical and Experimental Medicine, University of Surrey, UK

ARTICLE INFO

Keywords:
Chronic kidney disease
Electronic medical records
Estimated glomerular filtration rate
eGFR
Regression
Broken-sticks
Clinical time series

ABSTRACT

In order for clinicians to manage disease progression and make effective decisions about drug dosage, treatment regimens or scheduling follow up appointments, it is necessary to be able to identify both short and long-term trends in repeated biomedical measurements. However, this is complicated by the fact that these measurements are irregularly sampled and influenced by both genuine physiological changes and external factors. In their current forms, existing regression algorithms often do not fulfil all of a clinician's requirements for identifying short-term (acute) events while still being able to identify long-term, chronic, trends in disease progression. Therefore, in order to balance both short term interpretability and long term flexibility, an extension to broken-stick regression models is proposed in order to make them more suitable for modelling clinical time series. The proposed probabilistic broken-stick model can robustly estimate both short-term and long-term trends simultaneously, while also accommodating the unequal length and irregularly sampled nature of clinical time series. Moreover, since the model is parametric and completely generative, its first derivative provides a long-term non-linear estimate of the annual rate of change in the measurements more reliably than linear regression. The benefits of the proposed model are illustrated using estimated glomerular filtration rate as a case study used to manage patients with chronic kidney disease.

1. Introduction

The trend in measurements of clinical interest such as blood sugar, cholesterol or kidney function can provide insight into the change over time in a patient's condition. For patients with chronic illnesses such as diabetes and chronic kidney disease (CKD), monitoring of these measurements is necessary in order to effectively manage the condition. For example, in order for clinicians to make effective decisions about drug dosage, treatment regimens or when scheduling follow up appointments, it is necessary to know not only the value of these indicators, but also to have an idea of both the short- and long-term trajectory they are following. However, modelling the trend of biomedical measurements over the long-term can be complicated by both practical, e.g. the irregular taking of measurements and lengthy gaps between them, and biological considerations. For example, the primary indicator of kidney function, the estimated glomerular filtration rate (eGFR), can be influenced by, amongst other things, the level of protein in the diet, changes in muscle breakdown and the level of hydration [1]. This can lead to substantive variability in a patient's eGFR measurements [2,3]. Unfortunately, existing regression algorithms such as linear, polynomial and Gaussian process regression (GPR) [4] either cannot account for these challenges or do not satisfy the key clinical requirements of providing an easily interpretable model that can elucidate short- and long-term trends.

Biomedical measurements are irregularly sampled, posing an additional challenge to analysis. Prior work in time series analysis has strongly emphasised regularly sampled data, resulting in fewer methods that exist specifically for analysing irregularly sampled data. Despite methods for analysing irregular time series data directly having been employed successfully [5–7], the most common approach is still to transform the data to enforce regularity using either interpolation techniques or regression analysis [8]. However, with biomedical time series interpolation can present its own problems due to the measurements not always being taken at random, but rather requested at specific times by clinicians, e.g. as part of routine monitoring or as follow up to treatment. On the other hand, regression imposes a number of assumptions on both the variables and their relationships. For example, linear regression assumes a linear relationships between the dependent

[*] Corresponding author at: QuintilesIMS, London, UK.
E-mail address: norman.poh@quintilesims.com (N. Poh).

https://doi.org/10.1016/j.jbi.2017.10.006
Received 23 November 2016; Received in revised form 3 October 2017; Accepted 10 October 2017
Available online 16 October 2017

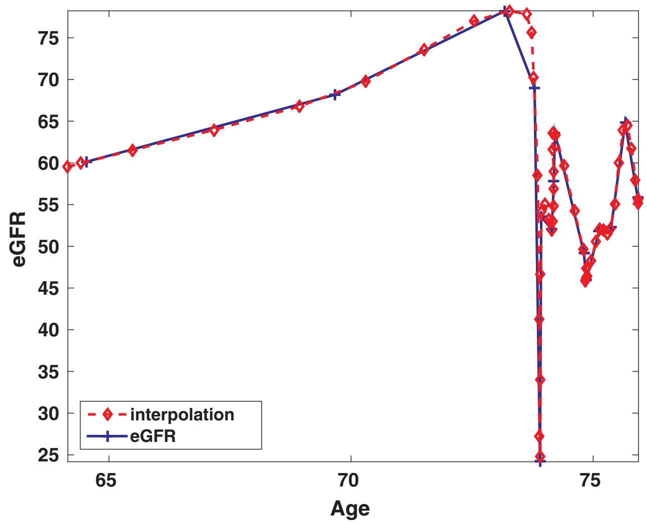

Fig. 1. An eGFR time series (blue) modelled using linear interpolation in order to produce a fixed-size vector of 50 observations (red) over the age range for which the patient has eGFR measurements [13]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and independent variables and independence of the residuals (no autocorrelation); assumptions which are usually violated in biomedical time series. Often linearity is violated due to an acute episode. For example, when a patient suffers an acute kidney injury (AKI) [9–12] their eGFR will drop sharply and potentially recover a short time after (as seen in Fig. 1). Long-term trends may therefore exhibit local fluctuations due to genuine physiological changes as well as external factors.

More flexible models such as Gaussian process regression (GPR) [14], multivariate adaptive regression splines [15] and multivariate additive models [16] can be used instead to provide the desired flexibility. For example, through the use of a kernel function GPR can avoid making the assumptions of linear regression. However, when there are gaps between the data, as is often the case with biomedical time series, the estimated variance of the predicted output can ‘explode’ [13] (Fig. 2). Consequently, these models are less interpretable, and therefore lose out in situations where a clinician simply needs to know whether a patient’s condition is progressing or improving.

In order to strike a balance between interpretability and flexibility, broken-stick regression, also known as segmented or piece-wise regression, can be used to linearly model local trends [17–20]. However, in this formulation local discontinuities are introduced at the segment boundaries, resulting in a loss of smoothness and consequently in the inability to infer trends in the boundary regions reliably. To address this we take a Bayesian approach to derive a long-term trend by enforcing a smooth transition between the locally linear line segments, while still preserving the local trends. The ability to capture both long- and short-term trends makes this approach ideally suited to modelling biomedical time series in a clinical context. Additionally, by enforcing smoothness local rates of change can be derived, giving clinicians an indication of whether a patient’s condition is progressing or not. Finally, a broken-stick model can accommodate gaps in a time series through choosing the length of each line segment in a manner that ensures that there are a sufficient number of measurements within each segment and can mitigate overfitting as it fits only locally linear line segments.

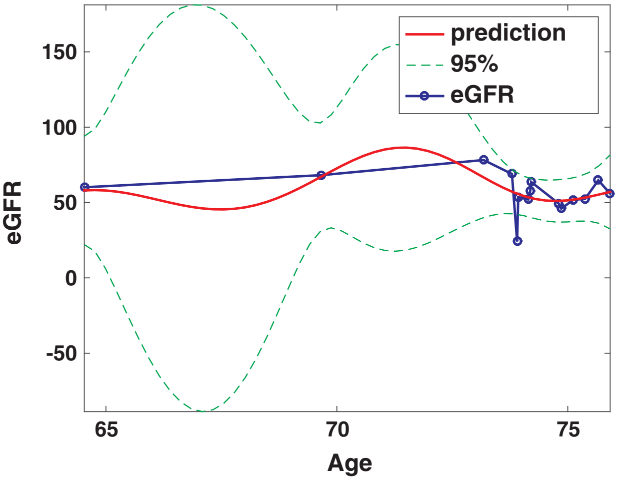

Fig. 2. The GPR model shows relatively low variance when the gap between measurements is small, but the variance increases markedly when the measurements are sparse.

Table 1
Notation.

Variable	Domain	Meaning
$\mathbf{T}$	Vector of real numbers	The time domain
t	Integer	Enumerator of the time domain, from 1 to T
w	Integer	Enumerator of the window, from 1 to W
$\mathbf{L}$	Vector of integers	Indices storing the beginning of window
$\mathbf{U}$	Vector of integers	Indices storing the end of a window
θ_w	Model parameters	Parameters of the w-th line segment
θ	All model parameters	$\{\theta_w \mid \forall_w\}$
Δ_d	Integer	Window interval of length d in year
W	Integer	Number of windows
$\omega_1^{(w)}$	Integer	Line segment gradient
$\omega_0^{(w)}$	Integer	Line segment intercept
$\mu_t^{(w)}$	Integer	Mean value of the time window

2. Methodology

Here, $\mathbf{X}$ is used to denote a vector and $\mathbf{X}[t]$ to denote the element in the vector indexed by t. The remainder of the notation used is given in Table 1.

2.1. Windowing

The first step in fitting the broken-stick model is the division of a time series into a number of windows. Here, windows of equal length d were used across all time series, although there is no constraint requiring the windows to be of equal length across or within individual time series. The window length was determined from the data based on the intervals between measurements, as there should be at least three measurements within each window in order to avoid overfitting line segments. In general, having more measurements within each window is preferable. However, it is only possible to influence the number of measurements within a window by increasing d, as the number of measurements in each time series is fixed. Given that larger values of d may result in local fluctuations going undetected, while smaller values of d may lead to measurement noise dominating the model, the window length must be optimised for each application.

2.2. Local fitting

Given d and a specified interval to slide the window by, Δ_d, the number of windows W is also determined. For each window, a linear regression is performed by:

$$\mu^{(w)}(t) = \omega_1^{(w)} \times t + \omega_0^{(w)}, \tag{1}$$

where $\omega_1^{(w)}$ is the gradient and $\omega_0^{(w)}$ is the intercept for the w-th window