Astronomy and Computing 10 (2015) 61–72

Contents lists available at ScienceDirect

Astronomy and Computing

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

Full length article

The overlooked potential of Generalized Linear Models in astronomy-II: Gamma regression and photometric redshifts

J. Elliott [a,*], R.S. de Souza [b], A. Krone-Martins [c], E. Cameron [d], E.E.O. Ishida [e], J. Hilbe [f,g] for the COIN collaboration

[a] Max-Planck-Institut für extraterrestrische Physik, Giessenbachstraße 1, 85748, Garching, Germany
[b] MTA Eötvös University, EIRSA "Lendulet" Astrophysics Research Group, Budapest 1117, Hungary
[c] SIM, Faculdade de Ciências, Universidade de Lisboa, Ed. C8, Campo Grande, 1749-016, Lisboa, Portugal
[d] Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford, OX1 3PS, United Kingdom
[e] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85748 Garching, Germany
[f] Arizona State University, 873701, Tempe, AZ 85287-3701, USA
[g] Jet Propulsion Laboratory, 4800 Oak Grove Dr., Pasadena, CA 91109, USA

ARTICLE INFO

Article history:
Received 22 September 2014
Accepted 7 January 2015
Available online 16 January 2015

Keywords:
Techniques: photometric
Methods: statistical
Methods: analytical
Galaxies: distances and redshifts

ABSTRACT

Machine learning techniques offer a precious tool box for use within astronomy to solve problems involving so-called *big data*. They provide a means to make accurate predictions about a particular system without prior knowledge of the underlying physical processes of the data. In this article, and the companion papers of this series, we present the set of Generalized Linear Models (GLMs) as a fast alternative method for tackling general astronomical problems, including the ones related to the machine learning paradigm. To demonstrate the applicability of GLMs to inherently positive and continuous physical observables, we explore their use in estimating the photometric redshifts of galaxies from their multi-wavelength photometry. Using the gamma family with a log link function we predict redshifts from the *PHoto-z Accuracy Testing* simulated catalogue and a subset of the *Sloan Digital Sky Survey* from Data Release 10. We obtain fits that result in catastrophic outlier rates as low as ∼1% for simulated and ∼2% for real data. Moreover, we can easily obtain such levels of precision within a matter of seconds on a normal desktop computer and with training sets that contain merely thousands of galaxies. Our software is made publicly available as a user-friendly package developed in `Python`, `R` and via an interactive web application. This software allows users to apply a set of GLMs to their own photometric catalogues and generates publication quality plots with minimum effort. By facilitating their ease of use to the astronomical community, this paper series aims to make GLMs widely known and to encourage their implementation in future large-scale projects, such as the Large Synoptic Survey Telescope.

1. Introduction

Generalized Linear Models (GLMs), as introduced by Nelder and Wedderburn (1972), offer a well established statistical framework for robust modelling and prediction making. It allows the application of regression analysis when the observed quantity originates from an *exponential family* distribution rather than a Gaussian (or Normal; e.g., Hardin and Hilbe, 2012; Hilbe, 2014). As a result, GLMs offer a readily interpretable and physically-motivated approach (via family distributions) to machine learning (ML) that can be applied to a variety of astronomical data sets. Despite being widely used across a range of scientific disciplines, such as biology (Brown and Rothery et al., 1993; Ahrestani et al., 2013), medicine (Lindsey, 1999), and economics (Pindyck and Rubinfeld, 1998; de Jong and Heller, 2008), and its availability within the overwhelming majority of contemporary statistical software packages (e.g., `R`, R Core Team 2014; `SAS`, Inc. 2003; and `STATA`, StataCorp 2009), GLMs remain almost *terra incognita* within the astronomical community (de Souza et al., 2014a).

One particular problem which presents itself as a candidate for the implementation of GLMs is the photometric redshift

* Corresponding author.
E-mail addresses: jonnyelliott@mpe.mpg.de (J. Elliott), rafael.2706@gmail.com (R.S. de Souza), algol@sim.ul.pt (A. Krone-Martins), dr.ewan.cameron@gmail.com (E. Cameron), emille@mpa-garching.mpg.de (E.E.O. Ishida), j.m.hilbe@gmail.com (J. Hilbe).

http://dx.doi.org/10.1016/j.ascom.2015.01.002
2213-1337/

(photo-z) estimation of galaxies. Although precise redshifts can in principle be directly determined through identification of known absorption or emission lines in the optical and/or near-infrared spectrum of each target galaxy, the observational cost of this procedure can quickly become prohibitive for large scale surveys. The only feasible alternative in such cases is to use available multi-wavelength photometry to infer approximate photo-zs instead, but this is not always a simple task.

There exist a plethora of different spectra emitted from galaxies throughout the Universe. Their characteristic features carry signatures from the galaxy's morphology, age, metallicity, star formation history, merging history, and a host of other confounding factors in addition to its redshift, thus, making photo-z estimation a far from trivial task. There exist several techniques which are commonly used to estimate redshifts from photometry and can be divided into: (i) template fitting techniques (e.g., Benítez, 2000; Bolzonella et al., 2000; Ilbert et al., 2006), and (ii) ML (or empirical) techniques (e.g. Connolly et al., 1995; Collister and Lahav, 2004; Wadadekar, 2005; Miles et al., 2007; O'Mill et al., 2011; Reis et al., 2012; Krone-Martins et al., 2014). In template fitting techniques, a set of synthetic spectra are determined from synthesised stellar population models for a given set of metallicities, star formation histories and initial mass functions, among other properties. The photo-z is calculated by determining the synthetic photometry (and thus spectral template and redshift) which best fits the photometric observations. ML techniques, on the other hand, usually require a data set with spectroscopically measured redshifts to train the chosen method.

Many studies have examined the individual advantages of each photo-z code (for a glimpse on the diversity of existent methods, see Hildebrandt et al., 2010; Abdalla et al., 2011; Zheng and Zhang, 2012; Sánchez et al., 2014, and references therein). Abdalla et al. (2011) investigated the differences between five commonly used template fitting codes and a neural network. The neural network proved to be more reliable in redshift ranges with a higher density of training data, while the template fitting methods depended heavily on the underlying templates. Despite these caveats, the overall performance of all codes was, to first order, consistent and displayed catastrophic errors ranging from 5%–9%, which is considered good in terms of photo-z estimates (Abdalla et al., 2011). More recently, methods which combine several photo-z techniques in a Bayesian approach, coined ensemble learning, have begun to be implemented with the hope that they can complement each other's drawbacks (Carrasco Kind and Brunner, 2014).

One of the largest practical difficulties for the current photo-z methods is the time necessary to either fit the templates or train the underlying ML method; on top of that, the required size of the training set is often highly influential for empirical methods (Firth et al., 2003). *Big data* catalogues expected from large sky surveys, like the *Large Synoptic Survey Telescope*[1] (LSST Science Collaboration et al., 2009), *EUCLID*[2] (Refregier et al., 2010) or the *Wide-Field Survey Infrared Telescope*[3] (Green et al., 2012), warrant the need for fast and reliable photo-z methods that are capable of processing such large volumes of data in minutes to days rather than years, thereby facilitating higher level analyses and model refinements for downstream data products.

In this work, we introduce a new technique based on robust principal component analysis (PCA) and GLMs to estimate photo-zs. The method runs in a matter of seconds on a single core computer, even for millions of objects. In addition, we achieved very low levels of catastrophic errors when using training sets of a few thousands of objects. The combination of short computational run time, moderate training set size, and small catastrophic errors makes GLMs a robust and implementable technique for future large scale surveys.

The outline of this article is as follows. In Section 2, we give a broad overview of GLMs, in Section 3 we provide a description of the data set utilised. The methodology implemented is outlined in Section 4. We then present our results and compare with the recent literature in Section 5 and summarise our conclusions in Section 6.

2. Overview of regression methods

Before we delve into the details of GLMs and the gamma family, we make a brief overview of linear regression, a common tool used within astrophysics. Afterwards, we explicitly outline the details of GLMs with the gamma family and explain how it can be applied to determine photo-zs for a particular data set.

2.1. Overview of linear regression

Consider a given data set containing N (distinct objects; e.g., galaxies),

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\},$$

where the x_i are observations of the independent variable, X, and the y_i are observed values of a dependent random variable (RV), Y, which is a function of X, $Y = f(X)$. Traditionally, X is called the *explanatory variable* and Y the *response variable*. The expected value and variance of Y are denoted by $\mathrm{E}(Y)$ and $\mathrm{var}(Y)$, respectively. In this context, a linear model describes the response variable (Y) as a linear function of the explanatory variable (X):

$$Y = \beta_0 + \beta_1 X + \epsilon = \eta + \epsilon, \tag{1}$$

where $\{\beta_0, \beta_1\}$ are scalars called *slope coefficients* or *covariates*, $\eta = \beta_0 + \beta_1 X$ is the linear component (or *predictor*) of this simple model. Finally, ϵ is an error term considered to be independent and identically distributed, $\epsilon \sim N(0, \sigma^2)$.

When a standard linear regression approach is applied, the linear predictor in Eq. (1) is assumed to fully describe the response variable. The measured values are used to determine the covariates of the linear predictor that uniquely identify a straight line through the chosen data set minimising the error term. Having the scalar coefficients determined, the model provides a direct relation between X and Y, allowing one to predict the mean value of Y for a given measurement of X.

In order to clarify the procedure described in the next subsections, we invite the reader to approach this simple linear regression problem from an alternative perspective. Consider now each measurement, $\{x_i, y_i\}$, as a realisation of different variables $\{X_i, Y_i\}$ from a common family of probability density functions (PDFs), but with distinct parameters μ_i for each index i. The underlying PDF driving the behaviour of the response variable (Y_i) will be denoted by $f(y_i; \kappa_i)$, where κ_i is the parameter vector of the PDF underlying the ith measurement. If Y_i follows a Normal PDF with mean μ_i and variance σ_i^2, then

$$f(y_i; \kappa_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\frac{(y_i - \mu_i)^2}{\sigma_i^2}\right], \tag{2}$$

where $\kappa_i = \{\mu_i, \sigma_i\}$. This is summarised as $Y_i \sim N(\mu_i, \sigma_i)$. For reasons which will be clarified later, we consider σ_i a fixed value and, thus, determining μ_i is enough to completely characterise $f(y_i; \kappa_i)$. In this context, we can relate the measured x_i to the expected value of the corresponding response variable, $y_i = \mathrm{E}(Y_i)$, through the

[1] http://www.lsst.org/lsst.
[2] http://sci.esa.int/euclid.
[3] http://wfirst.gsfc.nasa.gov.