Vision Research 122 (2016) 105–123

Contents lists available at ScienceDirect

Vision Research

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

Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data

Heiko H. Schütt [a,b,c,*], Stefan Harmeling [d], Jakob H. Macke [e,f,g], Felix A. Wichmann [a,g,h,*]

[a] Neural Information Processing Group, University of Tübingen, Tübingen, Germany
[b] Department of Psychology, Universität of Potsdam, Potsdam, Germany
[c] Graduate School for Neural and Behavioural Sciences IMPRS, Tübingen, Germany
[d] Institut für Informatik, Heinrich-Heine-Universität Düsseldorf, Germany
[e] Research Center Caesar, Bonn, Germany
[f] Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[g] Bernstein Center for Computational Neuroscience, Tübingen, Germany
[h] Max Planck Institute for Intelligent Systems, Tübingen, Germany

A R T I C L E I N F O

Article history:
Received 28 June 2015
Received in revised form 22 February 2016
Accepted 27 February 2016
Available online 2 May 2016
Number of Reviews = 4

Keywords:
Psychometric function
Bayesian inference
Beta-binomial model
Overdispersion
Non-stationarity
Confidence intervals
Credible intervals
Psychophysical methods

A B S T R A C T

The *psychometric function* describes how an experimental variable, such as stimulus strength, influences the behaviour of an observer. Estimation of psychometric functions from experimental data plays a central role in fields such as psychophysics, experimental psychology and in the behavioural neurosciences. Experimental data may exhibit substantial overdispersion, which may result from non-stationarity in the behaviour of observers. Here we extend the standard binomial model which is typically used for psychometric function estimation to a *beta-binomial* model. We show that the use of the beta-binomial model makes it possible to determine accurate credible intervals even in data which exhibit substantial overdispersion. This goes beyond classical measures for overdispersion—goodness-of-fit—which can detect overdispersion but provide no method to do correct inference for overdispersed data. We use Bayesian inference methods for estimating the posterior distribution of the parameters of the psychometric function. Unlike previous Bayesian psychometric inference methods our software implementation—*psignifit 4*—performs numerical integration of the posterior within automatically determined bounds. This avoids the use of Markov chain Monte Carlo (MCMC) methods typically requiring expert knowledge. Extensive numerical tests show the validity of the approach and we discuss implications of overdispersion for experimental design. A comprehensive MATLAB toolbox implementing the method is freely available; a python implementation providing the basic capabilities is also available.

1. Introduction

In psychophysics, experimental psychology and the behavioural neurosciences, researchers attempt to measure detection or discrimination behaviour as a function of stimulus level, i.e. some changeable aspect of the stimulus or experimental setup controlled by the researcher. The range of applications is vast, from simple detection of spots of lights or Gabor patches to categorical perception of faces in experimental psychology and from discrimination performance of a single neuron up to the behaving animal in neuroscience. After data collection, researchers frequently fit a psychometric function to their data—almost always an appropriately scaled cumulative probability density function—relating the independent variable on the abscissa to the observer's behaviour on the ordinate. Researchers then obtain the "threshold" and, sometimes, the slope from the estimated psychometric function. Detection or discrimination behaviour, or performance, is thus summarised using one or two values, namely the threshold and the slope.

Thus fitting the psychometric functions to experimental data is of central importance for many fields. Given this importance, much research was conducted to either investigate the efficiency and reliability of the data collection (e.g. Blackwell, 1952; Watson & Pelli, 1983; Green, 1990; Treutwein, 1995; García-Pérez, 1998; Kontsevich & Tyler, 1999; Jäkel & Wichmann, 2006; Shen & Richards, 2012) or how to obtain accurate estimates of the psychometric function parameters (e.g. O'Regan & Humbert, 1989; Treutwein & Strasburger, 1999; Wichmann & Hill, 2001a; Knoblauch & Maloney, 2012).

* Corresponding authors at: Neural Information Processing Group, University of Tübingen, Tübingen, Germany.

E-mail addresses: heiko.schuett@uni-tuebingen.de (H.H. Schütt), felix.wichmann@uni-tuebingen.de (F.A. Wichmann).

http://dx.doi.org/10.1016/j.visres.2016.02.002
0042-6989/© 2016 The Authors. Published by Elsevier Ltd.

However, unless one has collected infinitely many trials per psychometric function, the parameters of the psychometric function are not fully constrained by the data and there remains uncertainty regarding the estimated parameters. To be able to draw valid conclusions when comparing thresholds and slopes from different experimental conditions, it is essential that this uncertainty is quantified. Typically, the uncertainty is expressed in the form of confidence intervals around the point estimates. Unfortunately, a reliable and accurate characterisation of this uncertainty is more difficult to obtain than the estimates themselves, partly due to the small size of typical datasets collected during behavioural experiments.[1]

The bootstrap (Efron, 1979; Efron & Tibshirani, 1994) was the first numerical sampling method applied to psychophysical data in order to characterise the uncertainty of the point estimates, i.e. to obtain confidence intervals (Foster & Bischof, 1987; Maloney, 1990; Foster & Bischof, 1991, 1997; Wichmann & Hill, 2001b).[2] Hill (2002) showed, however, that bootstrapped confidence intervals in the context of psychometric function estimation can be too small, a result confirmed by both Kuss et al. (2005) and Fründ, Haenel, and Wichmann (2011).[3]

As an alternative to the bootstrap, Bayesian statistics,[4] is centred on the notion of how to quantify uncertainty, and thus Bayesian statistics, too, offers a suitable theoretical framework to analyse data obtained in psychophysics, experimental psychology and the behavioural neurosciences. Bayesian statistics is, furthermore, especially suited for the small datasets (sample sizes) typically gathered in behavioural experiments. Kuss et al. (2005) provide a detailed and tutorial-style introduction to Bayesian inference for psychometric functions, and show results from numerical simulations suggesting that credible intervals obtained from Bayesian inference are more accurate than those obtained using the bootstrap. Similar results were later obtained by Fründ et al. (2011).

Bayesian inference for psychometric functions cannot be performed analytically, and instead has to rely on numerical methods to obtain the posterior distribution of the parameters given the data. Both Kuss et al. (2005) and Fründ et al. (2011) use Markov chain Monte Carlo (MCMC) methods to generate samples from the posterior distribution over parameters. MCMC is a standard method in Bayesian inference in general, and, in principle, allows Bayesian inference to be performed on many statistical problems. Unfortunately MCMC requires considerable statistical expertise from the user to fine tune the proposal distribution and the sampling step size, and especially to detect when the sampling fails. MCMC methods thus rarely work "automatically" with no or little user intervention the way analytical methods and the bootstrap do. Kuss et al. aptly summarise the problem in their paper: *A difficulty of the proposed method is that using Markov chain Monte Carlo methods is nontrivial and requires the Markov chains to be inspected and parameters to be set by the user. In practice, the parameters are found in a trial-and-error procedure.* Kuss et al. (2005, p. 491). For many researchers in psychophysics, experimental psychology and the behavioural neurosciences this difficulty precludes the use of the MCMC-based Bayesian methods introduced by Kuss et al. (2005) and Fründ et al. (2011), and they still have to rely on the easier to use, albeit less accurate, bootstrap-based methods, e.g. the psignifit 2.5 toolbox by Wichmann and Hill (2001a, 2001b).

Finally, there is one more hurdle for inferring the uncertainty about the psychometric function parameters: overdispersion. Overdispersion means that the variance of the measured data is larger than expected from the binomial model, which may happen due to fluctuations in attention, vigilance, criteria or unmodelled aspects of the stimulus. Consequently all estimates of the uncertainty based on the binomial model become too small, whether based on Bayesian or on frequentist statistics if the data are overdispersed. To prevent this, early approaches used goodness-of-fit measures like deviance to detect overdispersion but could only suggest to reject overdispersed datasets (Wichmann & Hill, 2001b). Later Fründ et al. (2011) presented a method to perform a post hoc corrections of error bars for overdispersed datasets. However there has been no method which directly incorporated overdisperion in psychometric function fitting, despite the fact that the beta-binomial model for overdispersed binomial data has been well established for many years (Williams, 1982; McCullagh & Nelder, 1989, chap. 4.5, exercise 4.17; also see Venables & Ripley, 2013, chap. 7.5).

1.1. Contributions of this paper

The contributions of the current paper are fourfold:

1. We extend psychometric function modelling from the standard binomial to a beta-binomial model to capture overdispersion. We show that this model not only allows statistical inference from overdispersed data from a beta-binomial observer, but yields reasonable results for other sources of overdispersion, e.g. stemming from several types of serial dependencies (Sections 3.1 and 3.2).
2. We show that fitting a beta-binomial model provides a way of detecting overdispersion consistent with goodness-of-fit measures. In contrast with these approaches which can merely reject overdispersed data, this method allows valid statistical inference even for overdispersed data.[5]
3. We introduce a pain-free method for Bayesian inference for psychometric functions. First, we compute the posterior distribution of the parameters using numerical integration without the need for MCMC sampling techniques and any user intervention. Second, we suggest default priors and parameters for the Bayesian inference which in our simulations and experience yield good results, again without user intervention.[6] Third, we provide an implementation of the method, *psignifit 4*, coded in pure MATLAB[7] without dependencies on external code (such as mex-files) or other toolboxes, which eases the installation for the user, and helps the platform-independence.

[1] Much of conventional statistics relies on the asymptotic behaviour of estimators and probability distributions, i.e. relies on—ultimately infinitely—large datasets. Wichmann and Hill (2001a, 2001b) showed that, for the typical size of psychophysical datasets, methods based on asymptotic theory are not always reliable.

[2] Note that the "confidence intervals" estimated in the frequentist statistical framework, e.g. via the bootstrap, are not the same as the "credible intervals" obtained from Bayesian statistics. For a discussion of this difference in the context of psychometric functions see Kuss, Jäkel, and Wichmann (2005), p. 480–481. We always calculate credible intervals in what follows. For readers unfamiliar with the distinction, Bayesian credible intervals are what most people intuit when they think about confidence intervals, whereas the frequentist confidence intervals do not provide this Hoekstra, Morey, Rouder, and Wagenmakers (2014).

[3] Knoblauch and Maloney (2012) provide a comprehensive and clearly presented different approach to psychometric function estimation using the well-established framework of generalised linear models (GLMs). Their GLM approach benefits from a broad array of existing tests, confidence intervals, and software implementations. However, fitting asymptotes requires an alternation between the fitting of the GLM and fitting the asymptotes and the methods available to calculate confidence intervals are either based on asymptotic distributions for the parameter estimates or on bootstrapping. Thus the GLM approach provides no alternative approach for uncertainty assessment and, thus, no principled treatment of overdispersion.

[4] Detailed treatments of Bayesian statistics be found in many available textbooks, for example in O'Hagan (1994), Gelman et al. (2013), Jaynes (2003), and Kruschke (2014).

[5] Assuming, of course, that the data are reasonably well modelled using a sigmoidal function.

[6] Expert users can, of course, override any of the default choices in the software implementation, see the third sense of being pain-free.

[7] Similarly the python implementation does not require the user to compile code or link-in compiled binaries; furthermore, it does not require the user to install "exotic" packages.