RAS Techniques and Instruments

RASTI **4,** 1 (2025) https://doi.org/10.1093/rasti/rzaf035
Advance Access publication 2025 August 13

An approach to robust Bayesian regression in astronomy

William Martin [1]★ and Daniel Mortlock [1,2]

[1]*Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK*
[2]*Department of Mathematics, Imperial College London, London SW7 2AZ, UK*

Accepted 2025 July 31. Received 2025 July 29; in original form 2024 November 21

ABSTRACT

Model mis-specification (e.g. the presence of outliers) is commonly encountered in astronomical analyses, often requiring the use of ad hoc algorithms which are sensitive to arbitrary thresholds (e.g. sigma-clipping). For any given data set, the optimal approach will be to develop a bespoke statistical model of the data generation and measurement processes, but these come with a development cost; there is hence utility in having generic modelling approaches that are both principled and robust to model mis-specification. Here we develop and implement a generic Bayesian approach to linear regression, based on Student's t-distributions, that is robust to outliers and mis-specification of the noise model. Our method is validated using simulated data sets with various degrees of model mis-specification; the derived constraints are shown to be systematically less biased than those from a similar model using normal distributions. We demonstrate that, for a data set without outliers, a worst-case inference using t-distributions would give unbiased results with $\lesssim$10 per cent increase in the reported parameter uncertainties. We also compare with existing analyses of real-world data sets, finding qualitatively different results where normal distributions have been used and agreement where more robust methods have been applied. A Python implementation of this model, t-cup, is made available for others to use.

Key words: Data Methods – Software – Statistics – Bayesian methods – Regression analysis.

1 INTRODUCTION

Linear regression is a common problem in astronomy, arising in fields as diverse as galaxy formation and evolution [e.g. the supermassive black hole (SMBH) mass–stellar velocity dispersion correlation; Ferrarese & Merritt 2000; Gebhardt et al. 2000], stellar physics (e.g. the Leavitt law linking the luminosity and pulsation period for Cepheid variable stars; Leavitt & Pickering 1912), and cosmology (e.g. the original formulation of Hubble's law; Hubble 1929). For this reason, there are a plethora of techniques used by astronomers for linear regression when both the dependent and independent variables are measured with error (e.g. Press et al. 1992; Akritas & Bershady 1996; Tremaine et al. 2002; Kelly 2007 – see Andreon & Hurn 2013 and Andreon & Weaver 2015 for reviews including further examples).

Kelly (2007) illustrates that common ad-hoc estimators such as FITEXY (Press et al. 1992; Tremaine et al. 2002) and BCES (Akritas & Bershady 1996) suffer from biases and can underestimate intrinsic scatter. Similar algorithms exist for removing outliers from data (e.g. sigma clipping) have been used previously but these can lead to controversy (e.g. the reanalysis of the data set from Riess et al. 2011 by Efstathiou 2014). A more principled approach is to model the entire measurement process.

Bayesian hierarchical models (BHMs) are a natural way to model such data sets – these allow astronomers to account for, e.g. measurement errors, selection effects, interlinked parameters, censored data, and many other effects common to astronomical problems. BHMs have seen increasing use in astrophysics over the past few decades, from the distance–redshift relation in cosmology (e.g. Feeney, Mortlock & Dalmasso 2018; Avelino et al. 2019) and photometric redshift estimation (e.g. Leistedt, Mortlock & Peiris 2016) to exoplanet characterization (e.g. Sestovic, Demory & Queloz 2018) and population-level inference (e.g. Kelly, Vestergaard & Fan 2009) – see Feigelson et al. (2021) for a recent review including further examples of BHMs.

Kelly (2007) presented a general BHM for linear regression with measurement errors and censored data, demonstrating several advantages over the other methods considered: no bootstrapping was required to obtain uncertainties on parameters; the Bayesian approach was easily extensible to truncated or censored data; and other methods would sometimes severely underestimate intrinsic scatter in the data. This formulation of Bayesian regression, sometimes known as LINMIX_ERR, is now commonly used in astronomy (e.g. Andrews et al. 2013; Bentz et al. 2013; McConnell & Ma 2013). This approach has been extended and refined by others (e.g. Mantz 2016; Sereno 2016; Bartlett & Desmond 2023; Jing & Li 2024). These models assume parameters are normally distributed throughout; however, scientific uncertainties are often empirically not normally distributed, leading to more frequent outliers (Bailey 2017). Inference that relies on normal distributions can be unduly affected by outliers (see Section 4 for an exploration of this effect).

The problem of outliers within data sets can be thought of as model mis-specification: these objects do not fit the distributions used to model them. At very least, results should be checked by

★ E-mail: w.martin19@imperial.ac.uk

performing analyses using different models for the noise distribution (e.g. Andreon & Weaver 2015; McElreath 2020). However, it would be preferable to use robust inference methods, which are designed to work irrespective of the actual generative distribution (Berger et al. 1994). One approach is to use distributions that are leptokurtic (i.e. have heavier tails than a normal distribution) for inference, with suggestions including Student's t-distributions (Andrews & Mallows 1974), Gaussian mixture models (Box & Tiao 1968; Aitkin & Tunnicliffe Wilson 1980), and log-regularly varying distributions (LRVD; Gagnon, Desgagné & Bédard 2020; Hamura, Irie & Sugasawa 2022), all of which have been successful in practice (e.g. Berger et al. 1994; Sivia & Skilling 2006; Gelman et al. 2013; Gagnon & Hayashi 2023; Hamura, Irie & Sugasawa 2024). Student's t-distributions, parametrized by the shape parameter ν, have seen use in bespoke astronomical and cosmological inference, both by fixing ν (e.g. Andreon et al. 2008; Jontof-Hutter et al. 2016; Andreon 2020) and treating ν as a free parameter (e.g. Park et al. 2017; Feeney et al. 2018). The advantage of the latter approach is that we can marginalize over ν so that we only consider models supported by the data (Gelman et al. 2013). While there are code examples in which Student's t-distributions are used for the scatter term in the dependent variable of regression models (e.g. Appendix C.2 of Gelman et al. 2013, Section 9.1.3 of Andreon & Weaver 2015, or Example 7.35 of McElreath 2020), there is not currently a general method for robust Bayesian regression in astronomy which incorporates uncertainties in all quantities.

We propose a development of a generic approach for robust astronomical data analysis. From the review of previous methods, we can identify properties that we would like to see in our regression model:

(i) A BHM – we favour a hierarchical approach because it naturally encodes the hierarchical structure of astronomical regression problems (i.e. objects are drawn from a high-level population; the objects intrinsically obey some relationship; the objects are measured with error and only the measured values are known). We adopt a Bayesian approach both for practical reasons – the resultant posterior distributions provide full uncertainty quantification – and due to their logical consistency (e.g. Cox 1946; Van Horn 2003; Knuth & Skilling 2012).

(ii) A robust model – we desire a model that is robust to both outliers (i.e. data points that, whether intrinsically or by virtue of measurement errors, do not lie sufficiently close to our regression relation) and model mis-specification (i.e. where the underlying distribution of data does not match up with the distribution assumed for modelling).

(iii) A general method – we seek a model that does not require case-by-case optimization for application to different regression problems (e.g. no manual outlier identification and removal, no need to rescale prior distributions for different problems, etc.). While a bespoke model tailored to a specific problem will perform better than a general method, there is still value in making a generic approach available for analysis.

We implement these ideas in this paper, beginning with a discussion of our model in Section 2. In Section 3, we outline the methods that we use to validate our model; the results of these validation checks are presented in Section 4. In Section 5, we compare the performance of our model on real-world data sets with the models outlined in Kelly (2007) and Park et al. (2017), before summarizing our conclusions in Section 6.

2 FORMALISM

Here we establish notation and set out our model. Our data set has N astronomical objects, each with a K-dimensional vector of associated independent quantities $\{\boldsymbol{x}_i\}$ and a dependent quantity $\{y_i\}$. For example, if we were estimating SMBH mass using measurements of luminosity and line width, we would have $K = 2$ dimensional vectors of independent quantities $\{\boldsymbol{x}_i\} = \{(L_i, \Delta V_i)^T\}$, and the dependent quantity $\{y_i\}$ would correspond to the black hole mass.

We assume that the independent quantities $\{\boldsymbol{x}_i\}$ and dependent quantities $\{y_i\}$ are related by a regression relation

$$y_i = f(\boldsymbol{x}_i; \boldsymbol{\theta}_f) + \delta_i, \tag{1}$$

$$\delta_i \sim \mathcal{P}_{\rm int}(\boldsymbol{\theta}_{\rm int}), \tag{2}$$

where $f(\boldsymbol{x}_i; \boldsymbol{\theta}_f)$ is a function relating $\{\boldsymbol{x}_i\}$ to $\{y_i\}$, with parameters $\boldsymbol{\theta}_f$, and $\mathcal{P}_{\rm int}$ is an unknown probability distribution with parameters $\boldsymbol{\theta}_{\rm int}$.

These objects are then observed, resulting in the measured data

$$\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \boldsymbol{\epsilon}_{x,i}, \tag{3}$$

$$\hat{y}_i = y_i + \epsilon_{y,i}, \tag{4}$$

$$\boldsymbol{\epsilon}_{x,i} \sim \mathcal{P}_{\rm obs}(\boldsymbol{\theta}_{\rm obs}), \tag{5}$$

$$\epsilon_{y,i} \sim \mathcal{P}_{\rm obs}(\boldsymbol{\theta}_{\rm obs}), \tag{6}$$

where $\mathcal{P}_{\rm obs}$ is an unknown probability distribution with parameters $\boldsymbol{\theta}_{\rm obs}$.

When building a model to infer parameters of interest for our model, we must choose probability distributions to represent the unknown $\mathcal{P}_{\rm int}$ and $\mathcal{P}_{\rm obs}$ – we label the chosen distributions as $\tilde{\mathcal{P}}_{\rm int}$ and $\tilde{\mathcal{P}}_{\rm obs}$.

In a Bayesian framework, we can extend this model to deal with, e.g. censored data or selection effects, but this is beyond the scope of this work – see Kelly (2007) for an overview of an approach that would incorporate these effects.

2.1 Building a robust model

In the setup outlined above, the form of the 'true' distributions $\mathcal{P}_{\rm int}$ and $\mathcal{P}_{\rm obs}$ are unknown; we can only choose the forms of $\tilde{\mathcal{P}}_{\rm int}$ and $\tilde{\mathcal{P}}_{\rm obs}$. A common assumption in analysis is to use normal distributions to model both of these. However, in the case of model mis-specification (where, e.g. we assume $\tilde{\mathcal{P}}_{\rm int}$ is a normal distribution but the true $\mathcal{P}_{\rm int}$ is a different distribution), the results obtained under this assumption can be biased. Making the assumption that $\tilde{\mathcal{P}}_{\rm int}$ follows a different distribution can give results that are robust to this model mis-specification. This means that, even though we do not believe that the distribution we choose to model $\tilde{\mathcal{P}}_{\rm int}$ is the same as the underlying, unknown $\mathcal{P}_{\rm int}$, we can have greater confidence in the inferences about the regression model and the properties of individual objects.

In choosing the form of $\tilde{\mathcal{P}}_{\rm int}$, some consideration must be given to the type of outliers which are expected. In the case of extreme outliers, it is possible to use distributions with very heavy tails such as LRVD to provide whole robustness (Gagnon et al. 2020; Hamura et al. 2022). However, the focus here is on the less extreme situation of more numerous but moderate outliers that cannot be identified as easily but will none the less affect the inference. For this reason, we use Student's t-distributions, which have sufficient flexibility to represent any noise model which is unimodal and approximately symmetric. The model presented will not alleviate all possible

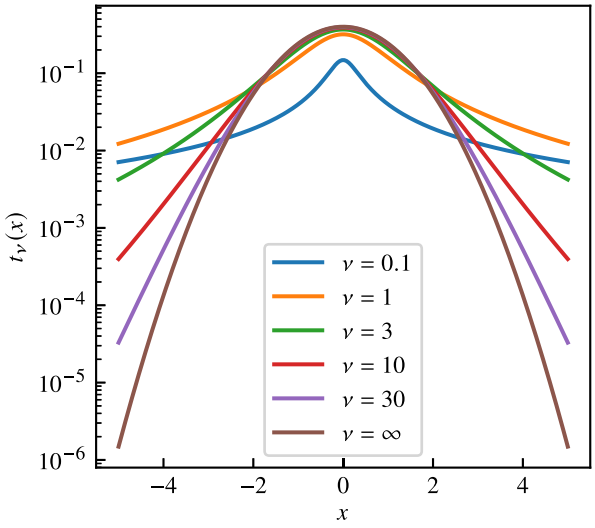

Figure 1. The t-distribution probability density function for different values of shape parameter, ν.

issues of model mis-specification, such as a strongly mis-specified regression relation $f(\mathbf{x}_i; \theta_f)$, but, as we shall show in Section 4, provides markedly improved performance over a similar model with normal distributions, while maintaining constraining power.

2.2 Sampling distribution

For robust inference, we seek a sampling distribution that can have heavier tails than a normal distribution, but that can reduce to a normal distribution when the underlying data set is normally distributed. We further seek an identifiable model, and a differentiable distribution which can be fit using Hamiltonian Monte Carlo (HMC). These constraints lead us naturally to Student's t-distributions, which fulfil all of these criteria.

Student's t-distributions are encountered when estimating the mean of a normal distribution with unknown variance from a limited number of samples. The number of samples is a parameter of the distribution: for n samples, the corresponding Student's t-distribution will have $\nu \equiv n - 1$ 'degrees-of-freedom'. This value of ν parametrizes how heavy-tailed the distribution is. While the interpretation of ν as 'degrees-of-freedom' only makes sense for $\nu \in \mathbb{Z}^+$, the distribution is normalizable for any positive, real ν; for this reason, we shall refer to ν as the shape parameter.

The Student's t-distribution, with location μ and scale σ, has the probability density function

$$t_\nu\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{\pi\nu}\sigma} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{1}{\nu}\frac{(x-\mu)^2}{\sigma^2}\right)^{-\frac{\nu+1}{2}}. \tag{7}$$

This is shown for a range of ν in Fig. 1: $\nu = 1$ gives a Cauchy distribution; and $\nu \to \infty$ tends to a normal distribution. The distribution has mean

$$\mathbb{E}(x) = \begin{cases} \mu & \nu > 1, \\ \text{undefined} & \text{otherwise} \end{cases} \tag{8}$$

and variance

$$\text{Var}(x) = \begin{cases} \frac{\nu}{\nu-2}\sigma^2 & \nu > 2, \\ \infty & 1 < \nu \leq 2, \\ \text{undefined} & \text{otherwise.} \end{cases} \tag{9}$$

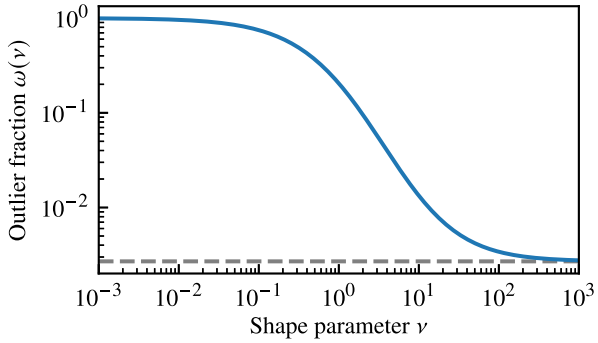

Figure 2. The outlier fraction, ω, for different values of shape parameter, ν. The outlier fraction of a normal distribution ($\approx 2.70 \times 10^{-3}$) is indicated by the dashed grey line.

In this paper, we have found it useful to define two quantities for comparison with normal distributions. The first, $\sigma_{68}(\nu)$, is the width of the highest density interval for a t-distribution with scale parameter $\sigma = 1$ such that the probability contained in the interval is equal to that of a 1σ interval for a normal distribution. This is given by

$$\sigma_{68}(\nu) = \sqrt{\nu\left(\frac{1}{I^{-1}(\Phi_{\sigma=1}; \frac{\nu}{2}, \frac{1}{2})} - 1\right)}, \tag{10}$$

where $\Phi_{\sigma=1} = \text{erf}(2^{-1/2}) \approx 0.6827$ is the posterior density contained within $\pm 1\sigma$ of the mean for a normal distribution, and I^{-1} is the inverse of the regularized incomplete beta function, which we define as

$$I(x; a, b) = \frac{\int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t}{\int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t}. \tag{11}$$

This quantity is useful for defining a 'scale' parameter that is less tightly coupled to ν than the σ parameter of the t-distribution.

The second quantity we use is an 'outlier fraction' ω, which is defined as the fraction of points expected to lie more than 3σ from the distribution mean μ. For a t-distribution, this is given by

$$\omega(\nu) = P\left(|x - \mu| > 3\sigma \mid x \sim t_\nu(\mu, \sigma)\right) \tag{12}$$

$$= 2F_\nu(\mu - 3\sigma), \tag{13}$$

where $F_\nu(x)$ is the cumulative distribution function for the Student's t-distribution with shape parameter ν. Fig. 2 shows the relationship between outlier fraction ω and shape parameter ν, with $\omega \to 2.70 \times 10^{-3}$ in the limit of a normal distribution, i.e. as $\nu \to \infty$. Under this definition approximately 1 in 370 data-points would be outliers for data that follows a normal distribution; for Cauchy-distributed data (i.e. $\nu = 1$), every fifth data-point to be an outlier.

2.3 Regression model

Our regression model, represented as a directed acyclic graph in Fig. 3, is specified here. We assume a linear relationship between $\{\boldsymbol{x}_i\}$ and $\{y_i\}$, i.e.

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}_f) = \alpha + \boldsymbol{\beta} \cdot \boldsymbol{x}_i, \tag{14}$$

where $\{\alpha, \boldsymbol{\beta}\} \equiv \boldsymbol{\theta}_f$. In this paper, we only consider this linear relationship, but this apparatus could be used for other, non-linear relationships between variables with different parameters.

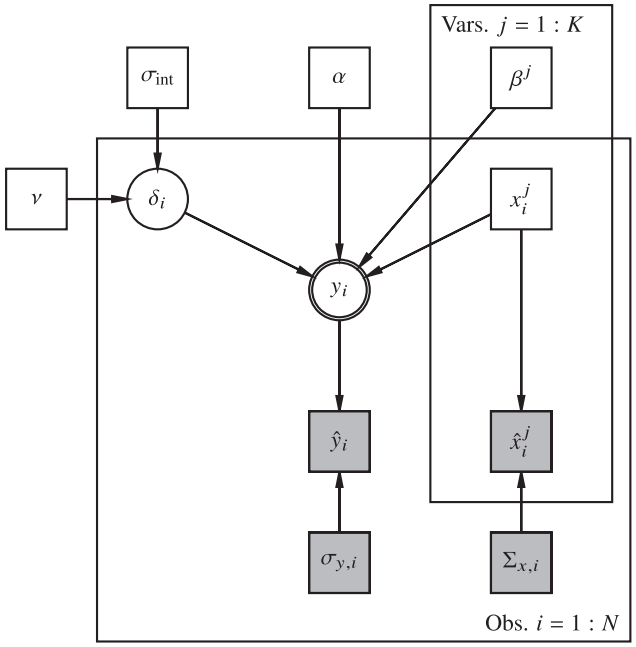

Figure 3. A directed acyclic graph representing the t-cup model for the function $f(\boldsymbol{x};\boldsymbol{\theta}) = \alpha + \boldsymbol{\beta}\cdot\boldsymbol{x}_i$.

Similarly to Kelly (2007), we define a Bayesian hierarchical model to reflect the nature of the regression problem. First, we assume that we can represent the probability distribution $\mathcal{P}_{\rm int}$ with a Student's t-distribution with shape parameter ν and scale parameter $\sigma_{\rm int}$ – i.e.

$$y_i \sim t_\nu\left(\alpha + \beta\boldsymbol{x}_i, \sigma_{\rm int}^2\right). \tag{15}$$

We use a Student's t-distribution here not because we believe that the data intrinsically follows this distribution, but because the resulting model is robust to model mis-specification – such as if a galaxy were to be an outlier from a particular relation as a result of a recent merger.

We assume that we can represent the probability distribution $\mathcal{P}_{\rm obs}$ as a normal distribution with scale parameter given by the error associated with the measured quantity. These measurements are modelled as

$$\hat{\boldsymbol{x}}_i \sim \mathcal{N}\left(\boldsymbol{x}_i, \Sigma_{x,i}^2\right) \tag{16}$$

$$\hat{y}_i \sim \mathcal{N}\left(y_i, \sigma_{y,i}^2\right). \tag{17}$$

We experimented with assuming $\mathcal{P}_{\rm obs}$ to be t-distributed, but found that the resultant model was difficult to sample as a result of its geometry, and that posterior predictive checks often included large outliers and did not resemble the data sets that gave rise to them. Fortunately, for robust inference the presence of a single heavy-tailed component is sufficient.

2.4 Priors

To ensure that our model is generically applicable to astronomical linear regression, we ensure that both independent and dependent quantities are scaled to have zero mean and unit variance. This allows us to set generic priors on the scaled intercept, gradients and scatter $\{\tilde{\alpha}, \tilde{\beta}, \tilde{\sigma}_{\rm int}\}$ that do not require rescaling for different units or data sets (though these priors can be revised to incorporate prior information).

Our default priors on the regression parameters $\{\tilde{\alpha}, \tilde{\beta}, \tilde{\sigma}_{\rm int}\}$ are

$$\tilde{\alpha} \sim \mathcal{N}(\mu = 0, \sigma^2 = 4), \tag{18}$$

$$\tilde{\beta} \sim \mathcal{N}(\mu = 0, \sigma^2 = 4), \tag{19}$$

$$\tilde{\sigma}_{68} \sim \Gamma(\alpha = 1.1, \theta = 5). \tag{20}$$

The motivation behind the prior on $\tilde{\alpha}$ is that, as the data is pre-scaled to have zero mean and the relationship between quantities is assumed to be linear, we would expect the data to have zero intercept; we adopt a broad prior by choosing a variance of 4. Similarly, the prior on $\tilde{\beta}$ has been chosen because the pre-scaling of the data suggests a gradient between -1 and 1, with variance chosen to give a weakly-informative prior. The prior on $\tilde{\sigma}_{68}$ is informed by Chung et al. (2013), who demonstrate that a prior with density at $\tilde{\sigma}_{\rm int} = 0$ will lead to a prediction of no intrinsic scatter in data sets where intrinsic scatter is present; in addition, there is a reasonable physical argument that most astrophysical processes will not produce objects with no population scatter. The chosen prior balances this constraint with difficulties arising in testing as a result of the pre-scaling, which led to insufficient prior density at small levels of intrinsic scatter. We found that choosing a small value of the gamma distribution power-law slope placed more weight at smaller intrinsic scatter, but required $\alpha > 1$ to ensure zero prior density at $\tilde{\sigma}_{68} = 0$; similarly, we found that a scale parameter of $\theta = 5$ ensured most of the prior density fell in the range $0 < \tilde{\sigma}_{68} < 1$, which we would expect because of the pre-scaling.

The prior on the latent values of the independent quantities $\{\boldsymbol{x}_i\}$ is a Gaussian mixture model prior, similar to that used in Kelly (2007). Although Gaussian mixture models may not seem suited to modelling distributions that have strong power-law gradients, previous work has found them to be more than sufficient to reproduce standard astronomical distributions (e.g. Blanton et al. 2003; Kelly, Fan & Vestergaard 2008 – see also Johnston 2011 for a review of approaches). To illustrate the approximation, we fit a 10-component Gaussian mixture model to two common population models in astronomy:

(i) the Schechter (1976) function

$$\Phi(L;\alpha,L_*,\phi_*) = \phi_*\left(\frac{L}{L_*}\right)^\alpha \exp\left(-\frac{L}{L_*}\right) \tag{21}$$

(ii) the double power-law function (first used to model the quasar population in Boyle et al. 1987 and then presented explicitly in Boyle, Shanks & Peterson 1988)

$$\Phi(L;\alpha,\beta,L_*,\phi_*) = \phi_*\left(\left(\frac{L}{L_*}\right)^\alpha + \left(\frac{L}{L_*}\right)^\beta\right)^{-1}. \tag{22}$$

Fig. 4 shows that, while these functions are well-approximated at the bright end of the luminosity function, if the sample is flux-limited then the mixture model prior will fall off at the faint end when the true population prior may increase. If sample selection effects are not accounted for (as in this work), this mismatch could lead to Eddington bias (Eddington 1913) affecting the inferred parameters – see Andreon & Hurn (2010) for an exploration of this effect. Similarly, using this mixture prior for a flux-limited sample could lead to Malmquist bias (Malmquist 1922), as brighter objects would be over-represented in the sample, though this is unlikely to affect the inferred regression relation. In such cases, another prior should be considered (and can be provided to the t-cup package) – for instance, if the luminosity function of the population is well-characterized, this can be provided as an alternative prior and sidestep these biases. An extension of the model that handles selection effects (e.g. Kelly 2007) would eliminate these biases, but only in cases where the selection

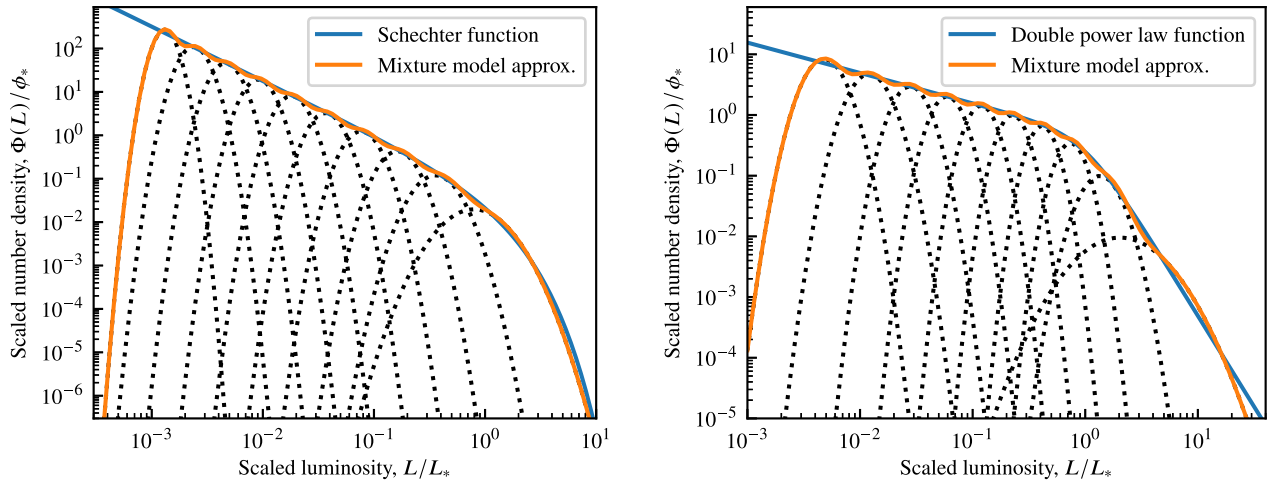

Figure 4. A mixture model approximation to a Schechter function (left) and a double power-law function (right).

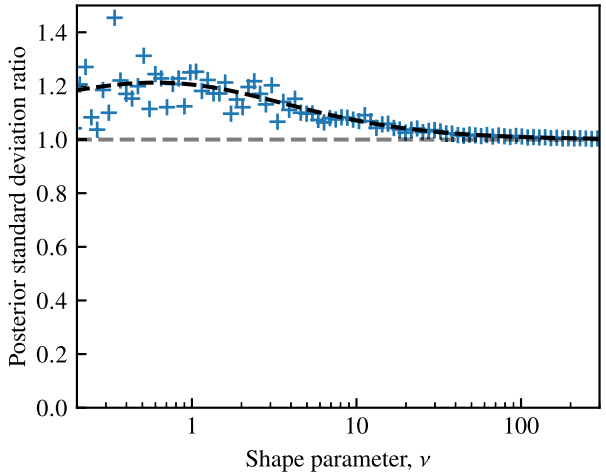

Figure 5. The ratio of the posterior standard deviation for a single location parameter when analysed with a t-distribution, compared to when analysed with a normal distribution. Each point represents a data set with $N = 100$ data points. The black dashed line shows the expected ratio of posterior standard deviations as calculated using the Fisher information of the t-likelihood.

function is well-defined; we defer a thorough treatment of a broad range of selection functions to future work.

While Kelly infers the complete prior as part of the model, we found that such a prior led to sampling issues as a result of the lack of identifiability inherent in a multicomponent Gaussian mixture model. Therefore, we use extreme deconvolution (Bovy, Hogg & Roweis 2011) to estimate the parameters of a Gaussian mixture model approximating the latent distribution of $\{\boldsymbol{x}_i\}$, and use these parameters for the prior.

In principle, the exact choice of prior for ν is unimportant, as the model is designed to be robust to model mis-specification; the only key element is that ν can take on values which correspond to both normal and heavy-tailed distributions. While difficulties in sampling ν according to a given prior distribution would not necessarily invalidate the results of the model as long as there is coverage at both small and large values, we use a prior that can be efficiently sampled using HMC. The prior we adopt on the shape parameter ν is

$$\nu \sim \text{Inv-}\Gamma(4, 15), \tag{23}$$

which balances flexibility (such that the model could accommodate both heavily leptokurtic distributions – e.g. the Cauchy distribution – and normally-distributed data) and numerical considerations (as $\nu \to 0$, the posterior geometry becomes highly curved, which leads to numerical issues; as $\nu \to \infty$, different values of ν become rapidly indistinguishable; both are down-weighted by this prior). The reasoning behind this choice of prior is expanded upon in Appendix A.

2.5 Asymptotic normality

One potential disadvantage of adopting a heavy-tailed likelihood in the statistical model is that the resultant inferences would, in general, be less constraining than under the assumption of a normal model (see e.g. the motivation behind the mixture model proposed in Tak, Ellis & Ghosh 2019). If there are no outliers, the implication would be that the robust approach is inferior. This effect is illustrated in Fig. 5, which shows results for data sets of $N = 100$ points drawn from a normal distribution with zero mean and unit variance. The points show the ratio of the posterior standard deviations for the mean when analysed using a t-distribution to that obtained using a normal distribution. These agree well with the approximate form of this ratio obtained using the Fisher information of a t-distribution likelihood (see Appendix B). This ratio increases with decreasing ν until $\nu \sim 0.6$, beyond which it falls again; this is a result of the increasingly narrow peak of the t-distribution (see Fig. 1). While the uncertainty can be increased by up to ~ 20 per cent (in the Cauchy regime, $\nu \approx 1$), we consider this an acceptable trade-off to reduce bias in cases of model mis-specification. This uncertainty trade-off can be mitigated by a flexible approach in which ν is inferred (such as the one presented here) as, if there are no outliers, higher values of ν corresponding to a normal distribution would be favoured. The increased numerical difficulty of treating ν as a parameter is justified by the reduction in the uncertainty relative to fixed ν analyses.

This effect would be appreciable for the smallest data sets – the posteriors would have heavier tails than under a normal model – but, in this case, meaningful constraints are unlikely irrespective of the adopted model. For larger data sets, of more than a few tens of points, these tails are effectively multiplied out, typically leaving just a single region of high probability. Moreover, for the regression problem considered here, the posterior in the regression parameters satisfies the requirements for asymptotic normality (e.g. Ghosh, Delampady

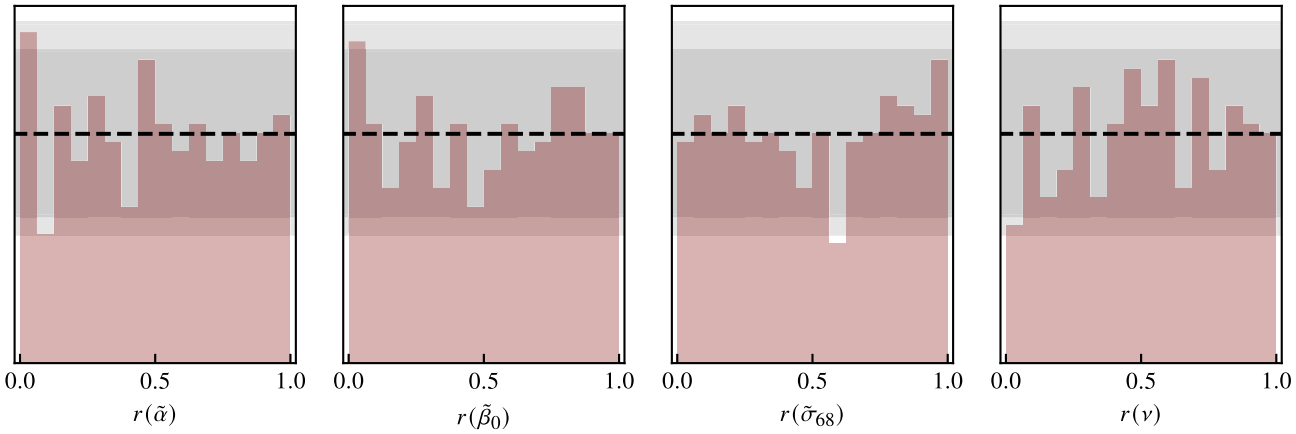

Figure 6. SBC runs for the t distribution under t-cup. If the inference procedure is working as expected, the histograms for each parameter should be distributed uniformly (as indicated by the black dashed line). The dark (light) grey regions correspond to the 94 per cent (98 per cent) confidence interval of uniformity [i.e. we expect one histogram bin per panel (figure) to lie outside of this range].

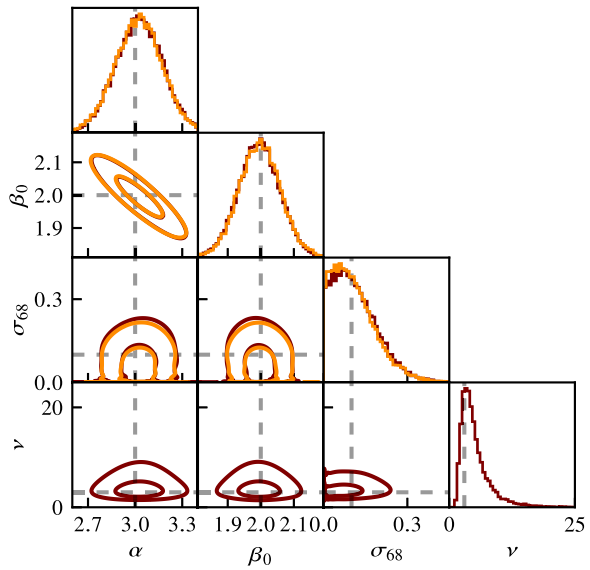

Figure 7. The posterior for each of the regression coefficients and the shape parameter ν for the t-distributed data set, with ground-truth values indicated by the grey dashed lines. Constraints from t-cup (red) are compared with those derived with ν fixed to the true value (orange). Contours indicate 39.3 per cent and 86.5 per cent highest posterior density regions, corresponding to 1σ and 2σ contours for a bivariate normal distribution.

& Samanta 2006), so the core of the posterior is Gaussian in form, just as would be the case under a normal model. (There is not even a significant numerical cost as the posterior has to be evaluated by sampling anyway.)

2.6 Implementation

The model described in Section 2.3 is implemented in NumPyro (Bingham et al. 2019; Phan, Pradhan & Jankowiak 2019), which is used to draw samples via a HMC No U-Turn Sampler. For the most part, we implement the model as presented previously; however, we reparametrize the Student's t-distribution using a mixture model of normal distributions (e.g. Stan Development Team 2024). We found this to be necessary for any model with non-negligible prior density below $\nu \lesssim 1$, as HMC struggles with the extreme gradients that would be present if the model were implemented using Student t-distributions directly (Neal 2003; Betancourt & Girolami 2013). The implementation is packaged as t-cup, available as a Python package.[1]

For the purposes of comparison, we have also implemented a model that mirrors the above structure, but uses normal distributions at each stage – we refer to this as n-cup.

3 VALIDATION

In this section, we outline the methods used to validate the model set out in Section 2. We run two types of tests to validate the performance of t-cup under different types of model mis-specification: simulation-based calibration tests (Cook, Gelman & Rubin 2006; Talts et al. 2018); and fixed value calibration tests.

3.1 Simulation-based calibration

Simulation-based calibration tests (Cook et al. 2006; Talts et al. 2018) are used to diagnose sampling issues. Parameters of interest θ_0 are drawn from the model prior $\pi(\theta)$, and a data set is drawn following the prescription of the model. We can then run a single Markov chain Monte Carlo (MCMC) chain on this data set until we have drawn L independent samples $\{\tilde{\theta}_i\}$ [see Talts et al. (2018) for a discussion of independence criteria]. We then compute the rank statistic

$$r(\theta_0, \tilde{\boldsymbol{\theta}}_{\mathrm{MCMC}}) = \sum_{i=1}^{L} \mathbb{I}(\tilde{\theta}_i < \theta_0), \tag{24}$$

where $\mathbb{I}(\cdot)$ is the indicator function which evaluates to 1 if its (logical) argument is true and 0 if it is false. If the generative model matches the sampling model and the sampling algorithm correctly draws from the posterior, the rank statistic will be uniformly distributed across the integers $[0, L]$. We can therefore repeat this process many times to build a histogram of rank statistics; any shortfalls in the sampling algorithm will manifest as deviations from uniformity.

While this procedure has been developed to diagnose sampling issues, we can also use the method to build a heuristic measure of how robust a model is to mis-specification. If the generative model and the sampling model no longer match, we would expect deviations from uniformity in the rank statistic histogram. We can compare how

[1] https://github.com/wm1995/tcup

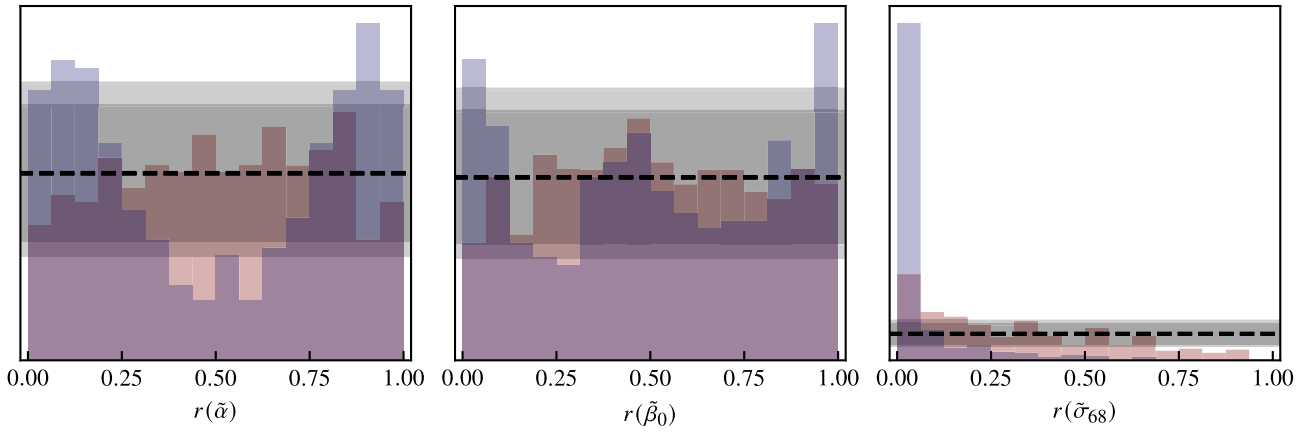

Figure 8. A rank statistic histogram for 400 simulation-based calibration runs for the outlier data set under n-cup (blue) and t-cup (red). If there were a perfect match between the generative model and the inference model, the histograms for each parameter will be distributed uniformly (as indicated by the black dashed line). The dark (light) grey regions correspond to the 94 per cent (98 per cent) confidence interval of uniformity [i.e. we expect one histogram bin per panel (figure) to lie outside of this range].

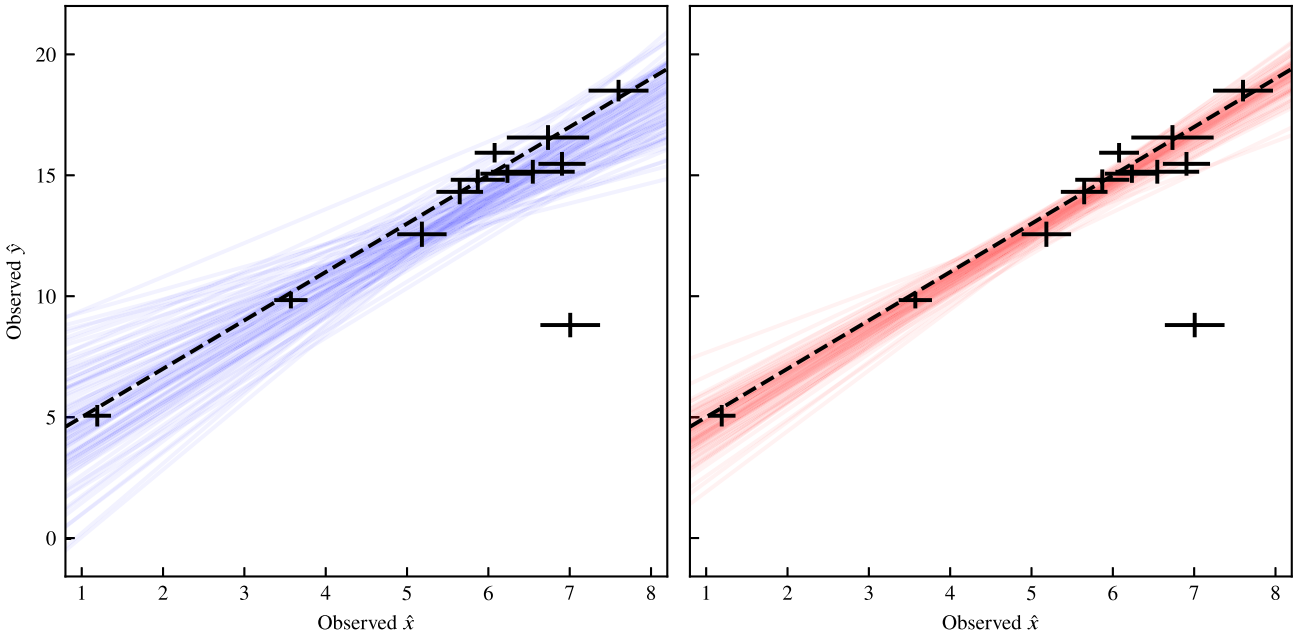

Figure 9. 100 draws from the posterior of regression lines from the normal model (left panel, blue) and t-cup (right panel, red). The data set is illustrated by the black points, with the ground-truth regression line illustrated by the black dashed line.

well different sampling models deal with model mis-specifcation by building the rank statistic histogram and assessing the deviation from uniformity.

As our priors are defined in a scaled space, our simulation-based calibration tests are also conducted in this scaled space; in this way, we are solely testing the performance of the MCMC models, and not of the scaling before fitting the models.

3.2 Fixed-value calibration

For our fixed-value calibration tests, we simulate data sets from known models with true fixed values for our regression parameters. The purpose of these tests is to demonstrate accurate recovery of these values, and to compare with the results given when using normal distributions.

We generate multiple data sets with the same ground-truth parameters to verify the results across multiple runs, producing plots of the composite cumulative distribution function across each data set. In each instance, full specifications of the data models are given in Appendix C.

We aim to do a full end-to-end test of our inference pipeline in the fixed-value calibration tests; therefore these data sets are generated in the unscaled space and are a test not only of the sampler but also of the scaling.

4 RESULTS ON SIMULATED DATA SETS

In the previous section, we proposed a general-purpose, robust statistical model for linear regression; in this section, we investigate the performance of the model on a series of simulated data sets

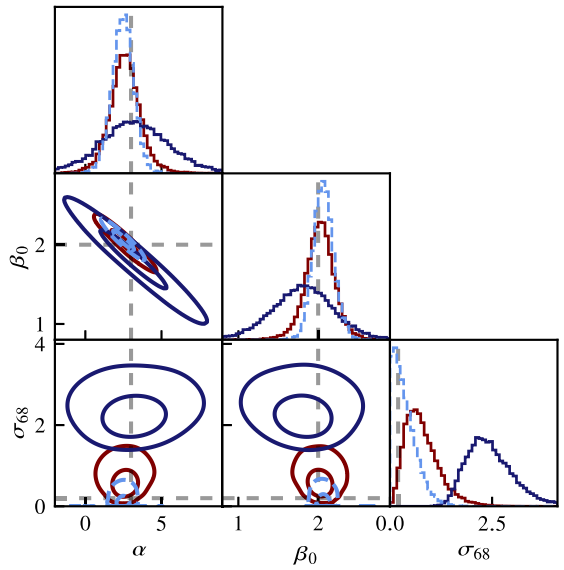

Figure 10. The posterior for each of the regression coefficients under the normal model with the outlier included (dark blue, solid) and excluded (light blue, dashed), and for t-cup with the outlier included (dark red, solid). Ground-truth values are indicated by the grey dashed lines.

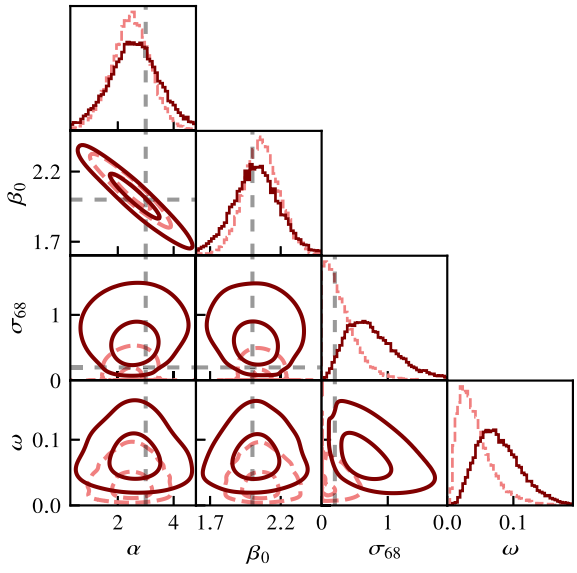

Figure 11. The posterior for each of the regression coefficients and the shape parameter ν under the t-cup model with the outlier included (dark red, solid) and excluded (light red, dashed). Ground-truth values are indicated by the grey dashed lines.

with known parameters. Code that reproduces the data sets in this section is available online.[2]

4.1 *t*-distributed data

For our first test, we start with a data set that matches our model perfectly (i.e. $\mathcal{P}_{\rm int} = t_\nu$, $\mathcal{P}_{\rm obs} = \mathcal{N}$), and choose $K = 1$ independent variables. The simulation-based calibration tests indicate that the rank statistic is consistent with being uniformly distributed (see Fig. 6), and therefore that our inference procedure is working as expected.

For our fixed-value tests, we drew $N = 20$ data points with $(\alpha, \beta, \sigma_{68}, \nu) = (3, 2, 0.1, 3)$; the full model used to generate the data is specified in Appendix C1. The chosen value of $\nu = 3$ corresponds to an outlier fraction of $\omega(\nu = 3) \approx 5.8$ per cent. As shown in Fig. 7, we recover the values of the parameters that were used to generate the data set; these values are consistent with a model where the shape parameter is fixed to $\nu = 3$.

We then generated 400 data sets from the same ground-truth parameters, and ran MCMC against each data set. 95 per cent highest posterior density credible intervals were constructed for each run; the true parameter values were contained in these credible intervals 97 per cent, 96 per cent, and 98 per cent of the time for the intercept α, the slope β, and the intrinsic scatter σ_{68}. For the nuisance parameter ν, the credible intervals contained the true parameter value across all runs; this may be caused by this parameter being only weakly constrained for these data sets.

4.2 Normally distributed data

In this test, we compare t-cup with an equivalent model that employs normal distributions to check that:

(i) t-cup reduces to a normal model in the absence of outliers
(ii) t-cup gives less biased results when an extreme outlier is present.

We conducted a simulation-based calibration test, where data sets were generated using normal distributions throughout (i.e. $\mathcal{P}_{\rm int} = \mathcal{P}_{\rm obs} = \mathcal{N}$) and with a single independent variable (i.e. $K = 1$). The results (shown in Fig. 8) indicate that, while the estimates of the intercept and slope are significantly biased when analysed by the normal model (as indicated by the rank statistic's distribution deviating from normal), the estimates derived by t-cup are unbiased. Additionally, while both n-cup and t-cup systematically overestimate the intrinsic scatter σ_{68} (as defined in equation 10), the bias is much smaller for t-cup than for n-cup.

A data set of $N = 12$ points was generated with $(\alpha, \beta, \sigma_{\rm int}) = (3, 2, 0.2)$, and one of the points was modified to be a $\sim 20\,\sigma$ outlier (the full generative model is given in Appendix C2). While such an extreme outlier could easily be identified and removed, the purpose here is to demonstrate that t-cup does not require this to obtain sensible inferences.

Fig. 9 illustrates how the estimates for the true parameter models are biased in the normal model, but less affected in the t-cup model.

In Fig. 10, we compare the constraints on parameters derived under the normal model (including and excluding the outlier from the data set), and the t-cup model (including the outlier only). The constraints from the t-cup model including the outlier are consistent with those derived under the normal model when the outlier is excluded, obviating the need to remove the outlier manually. While this outlier is particularly extreme, this example illustrates the utility of the t-cup model in data sets with outliers.

For completeness, in Fig. 11 we illustrate that the t-cup model recovers consistent constraints regardless of whether the outlier is included or excluded.

We then generated 400 data sets using the same procedure, and combined posterior samples from each run to build an effective cumulative distribution function across all runs for both the normal model and for t-cup. The results (illustrated in Fig. 12) indicate that constraints under t-cup are significantly less biased than those calculated under the normal model.

[2] https://github.com/wm1995/tcup-paper

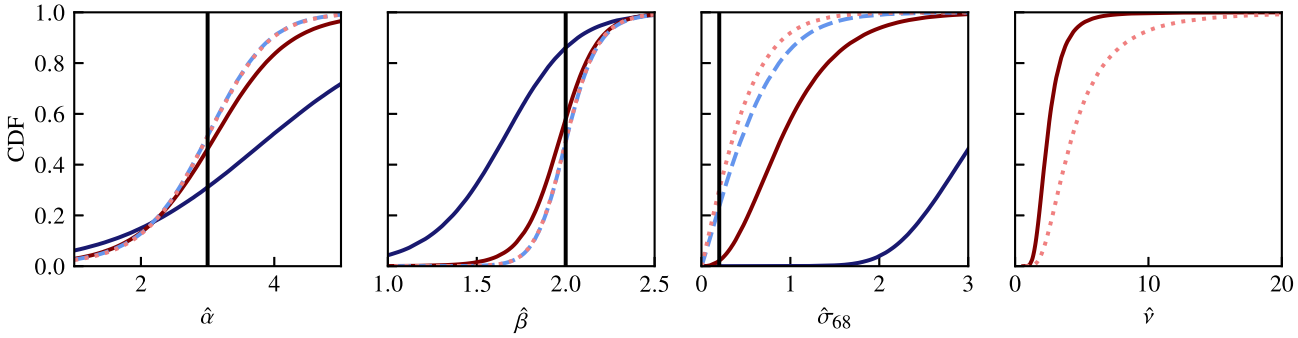

Figure 12. The combined cumulative distribution function of the regression parameters for 400 normal data sets with an outlier under the normal model (dark blue, solid) and t-cup model (dark red, solid), and with the outlier removed for the normal model (light blue, dashed) and t-cup model (light red, dashed).

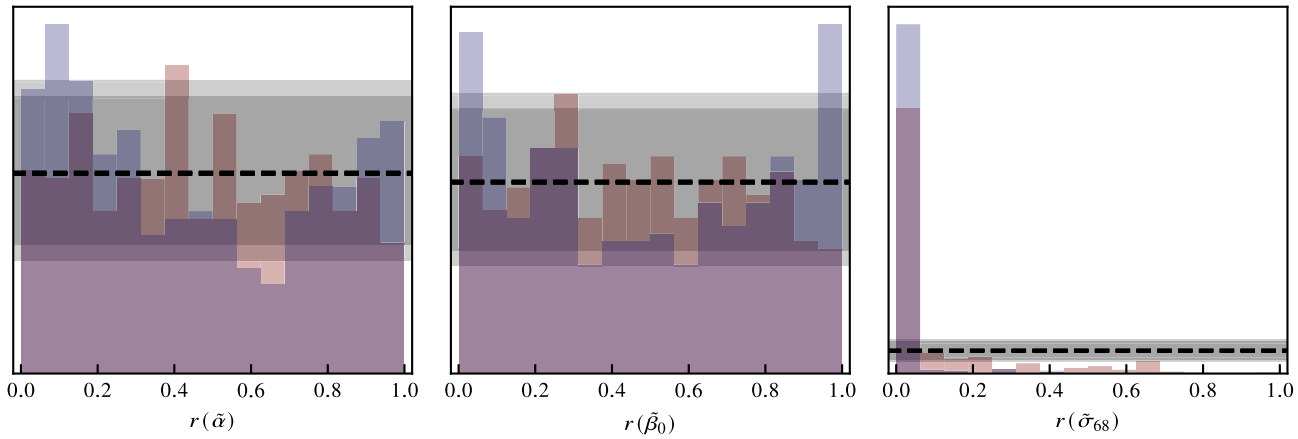

Figure 13. A rank statistic histogram for 400 simulation-based calibration runs for the heavy-tailed observation data set under n-cup (blue) and t-cup (red). If there were a perfect match between the generative model and the inference model, the histograms for each parameter would be distributed uniformly (as indicated by the black dashed line). The dark (light) grey regions correspond to the 94 per cent (98 per cent) confidence interval of uniformity [i.e. we expect one histogram bin per panel (figure) to lie outside of this range].

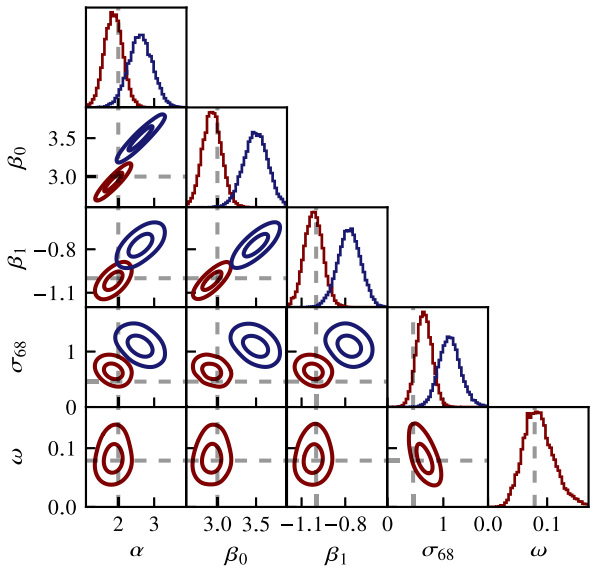

Figure 14. The posterior for each of the regression coefficients and the outlier fraction ω for the normal mixture model data set for n-cup (blue) and t-cup (red), with ground-truth values indicated by the grey dashed lines.

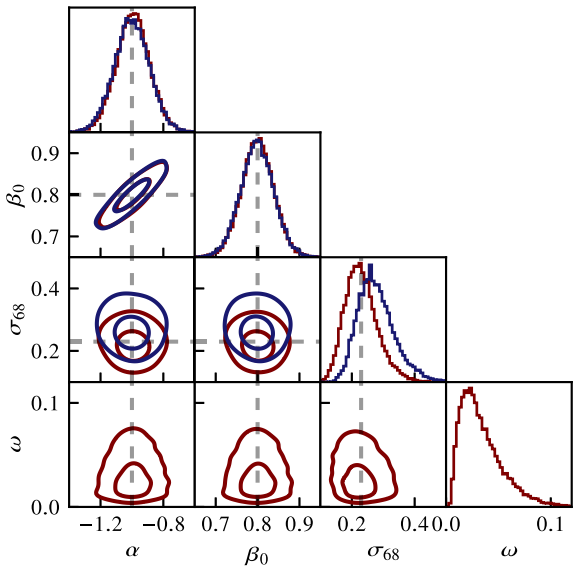

Figure 15. The posterior for each of the regression coefficients and the shape parameter ν for the Laplace-distributed data set under the normal model (blue) and t-cup (red), with ground-truth values indicated by the grey dashed lines.

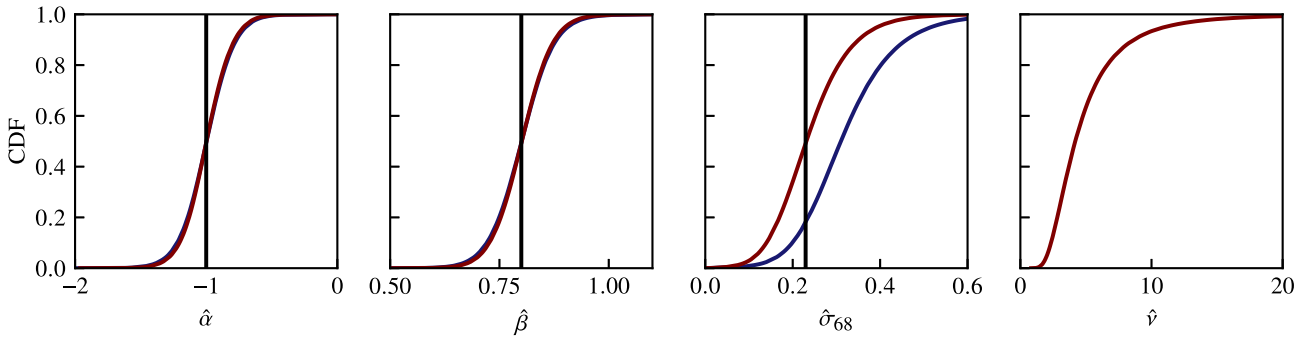

Figure 16. The combined cumulative distribution function of the regression parameters for 400 Laplace data sets under the normal model (blue) and t-cup model (red).

4.3 Mis-specified observational error

This test explores how t-cup performs when the distribution $\mathcal{P}_{\rm obs}$ is mis-specified. We conducted a simulation-based calibration test where $\mathcal{P}_{\rm int}$ is a normal distribution, and used $\mathcal{P}_{{\rm obs},x} = $ Laplace, $\mathcal{P}_{{\rm obs},y} = $ Cauchy for observational error. The results (see Fig. 13) show that t-cup is able to recover both the intercept and slope without bias, while n-cup produces biased estimates of both. As with the outlier case above, both t-cup and n-cup show significant bias in recovering the intrinsic scatter, but the bias is smaller for t-cup.

4.4 2D normal mixture model with outliers

This test is designed to further explore how t-cup performs when the model is mis-specified. In this case, we are looking at a normally distributed population which has a 10 per cent contamination rate with another normally distributed population of the same mean but 10 times the standard deviation. Equivalently, this test can be thought of as investigating how the model performs when there is a significant fraction of outliers.

The intrinsic scatter distribution $\mathcal{P}_{\rm int}$ is a mixture of two normal distributions with zero mean; 90 per cent of points are drawn from a core distribution with standard deviation $\sigma_{\rm int}$, and 10 per cent of points are drawn from an outlier distribution with standard deviation $10\sigma_{\rm int}$. The observation distribution $\mathcal{P}_{\rm obs}$ is a normal distribution. For fixed-value tests, the true values of the regression parameters were fixed to $(\alpha, \beta_0, \beta_1, \sigma_{\rm int}) = (2, 3, 1, 0.4)$. The full generative model for the fixed-value tests is given in Appendix C3.

The results (see Fig. 14) show that, while the constraints from the normal model are significantly biased for this generative distribution, t-cup is able to recover the correct parameter values.

4.5 Laplace-distributed data

As the underlying sampling distributions can never be known with certainty it is important to test the method on data generated on multiple distributions that do not fall into the family of t-distributions. Here we use a Laplace distribution as $\mathcal{P}_{\rm int}$, giving another example of performance under explicit model mis-specification. The Laplace distribution has probability density

$$\text{Laplace}\,(x;\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right), \tag{25}$$

with scale parameter b related to σ_{68} as $b(\sigma_{68}) = -\sigma_{68}/\ln\left(1-\text{erf}(2^{-1/2})\right) \approx 0.87\,\sigma_{68}$, where σ_{68} is defined in equation (10).

We generate $N = 25$ data points from a model with Laplacian intrinsic scatter (i.e. $\mathcal{P}_{\rm int} = $ Laplace), normally distributed observation noise (i.e. $\mathcal{P}_{\rm obs} = \mathcal{N}$), and $K = 1$ independent variables. For the fixed-value tests, we set $(\alpha, \beta, b) = (-1, 0.8, 0.2)$. The full generative model can be found in Appendix C4.

As we can see in Fig. 15, while both n-cup and t-cup successfully recover the gradient and intercept that were used to generate the data set, n-cup overpredicts the true intrinsic scatter when t-cup constrains it accurately. To confirm this, we generated 400 data sets using the same procedure, and analysed each with both n-cup and t-cup. The results (illustrated in Fig. 16) show the same pattern – that t-cup is able to accurately constrain the intrinsic scatter, while the estimate from the normal model is biased high.

5 DEMONSTRATION ON REAL DATA

Having seen t-cup's performance on simulated data in the previous section, we now compare the performance of the t-cup model with a generic astronomical Bayesian linear regression model (LINMIX_ERR; Kelly 2007) and a tailored approach using t-distributions (Park et al. 2017).

5.1 LINMIX_ERR

We use the same data set that is used in section 8 of Kelly (2007) – a data set of $N = 39$ quasars with measured Eddington ratio $L/L_{\rm Edd}$ and X-ray spectral index Γ. Performance is compared with a Python implementation[3] of the original LINMIX_ERR paper proposed by Kelly – see Fig. 17.

While the parameter estimates are broadly consistent (see Table 1), the posterior distributions in Fig. 18 can be seen to differ, with the posterior inferred by LINMIX being more diffuse than that inferred by t-cup. The tighter constraints of t-cup suggest that the inference presented by Kelly (2007) may be biased by the enforced assumption of normally distributed data, which may not be accurate. The explicit assumption of normality in LINMIX is in tension with the outlier fraction estimated by t-cup (68 per cent CI $\omega = 0.05^{+0.04}_{-0.01}$). This showcases that analysing data with robust procedures gives materially different answers on real-world data, and highlights the importance of carefully considering the implications when assuming normality.

[3] https://github.com/jmeyers314/linmix

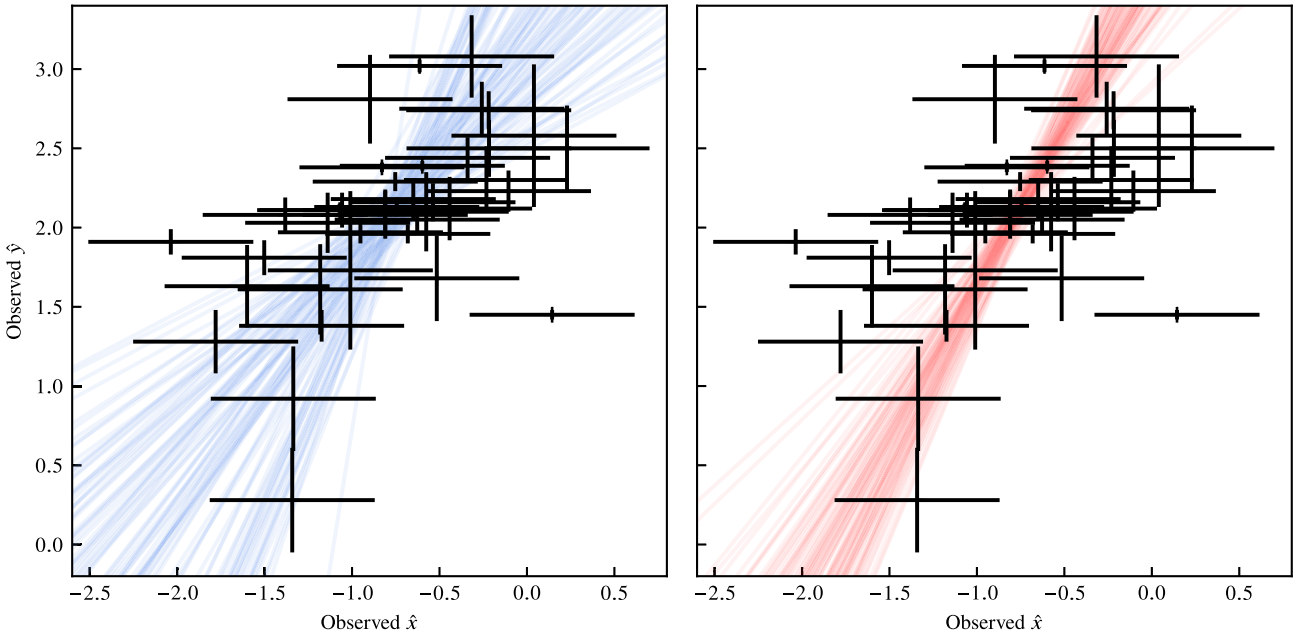

Figure 17. 100 draws from the posterior of regression lines from LINMIX_ERR (left panel, blue) and t-cup (right panel, red) for the Kelly (2007) data set.

Table 1. Estimates for the intercept, slope, and intrinsic scatter inferred for the relationship between Eddington ratio $L/L_{\rm Edd}$ and X-ray spectral index Γ for the sample of quasars analysed by Kelly (2007). The parameter estimates reported in Kelly (2007) are the posterior median and 'a robust estimate of the standard deviation'; for linmix and t-cup, we report the posterior median and an estimate of the standard deviation as $\sigma = 1.4826\,\mathrm{MAD}$, where MAD is the median absolute deviation.

	Kelly (2007)	linmix	t-cup
Intercept α	3.12 ± 0.41	3.18 ± 0.48	3.62 ± 0.26
Slope β	1.35 ± 0.54	1.40 ± 0.60	2.00 ± 0.33
Int. scatter σ_{68}	0.26 ± 0.11	0.25 ± 0.12	0.08 ± 0.07
Outlier fraction ω	–	–	0.04 ± 0.03

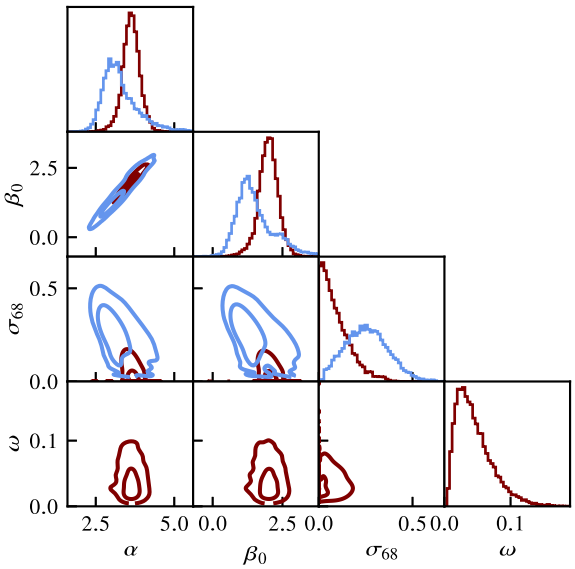

Figure 18. The posterior distributions for the data from Kelly (2007) under the linmix (blue) and t-cup (red) models.

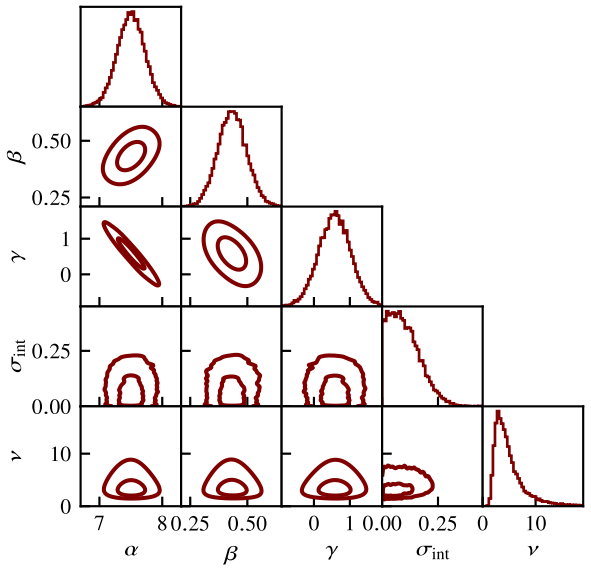

Figure 19. The posterior distributions for the data from Park et al. (2017) under the t-cup model. This figure is directly comparable with fig. 9 from Park et al. (2017).

5.2 Park et al. (2017)

Park et al. (2017) presents a bespoke t-distribution-based linear regression model for estimating SMBH mass; this allows us to compare the results of their bespoke model with our generic one. Their data set consists of $N = 31$ AGN with reverberation-mapped mass estimates, $M_{\rm BH}$, C IV emission line width, ΔV, and continuum luminosities, λL_λ. Three different C IV line width measurements are compared in Park et al. (2017): the full-width at half maximum (FWHM), the line dispersion, σ_l, and the median absolute deviation.

Table 2. A comparison between the constraints on the parameters in equation (26) derived by Park et al. (2017) and using t-cup.

	FWHM		σ_l		MAD	
	Park et al. (2017)	t-cup	Park et al. (2017)	t-cup	Park et al. (2017)	t-cup
Intercept α	$7.54^{+0.26}_{-0.27}$	$7.51^{+0.22}_{-0.22}$	$6.90^{+0.35}_{-0.34}$	$6.91^{+0.30}_{-0.31}$	$7.15^{+0.24}_{-0.25}$	$7.15^{+0.22}_{-0.22}$
Slope β	$0.45^{+0.08}_{-0.08}$	$0.43^{+0.06}_{-0.06}$	$0.44^{+0.07}_{-0.07}$	$0.43^{+0.05}_{-0.06}$	$0.42^{+0.07}_{-0.07}$	$0.42^{+0.05}_{-0.06}$
Slope γ	$0.50^{+0.55}_{-0.53}$	$0.58^{+0.44}_{-0.45}$	$1.66^{+0.65}_{-0.66}$	$1.66^{+0.57}_{-0.58}$	$1.65^{+0.61}_{-0.62}$	$1.65^{+0.55}_{-0.55}$
Int. scatter σ	$0.16^{+0.10}_{-0.08}$	$0.10^{+0.03}_{-0.10}$	$0.12^{+0.09}_{-0.06}$	$0.08^{+0.02}_{-0.08}$	$0.12^{+0.09}_{-0.06}$	$0.07^{+0.02}_{-0.07}$

This data set is used to fit the regression relation

$$\log_{10}\left(\frac{M_{\rm BH}}{10^8 M_\odot}\right) = \alpha + \beta \log_{10}\left(\frac{\lambda L_\lambda}{10^{44}{\rm erg\ s^{-1}}}\right) + \gamma \log_{10}\left(\frac{\Delta V}{10^3{\rm km\ s^{-1}}}\right). \quad (26)$$

The t-cup posterior for parameters α, β, and γ, as well as intrinsic scatter σ_{68} and shape parameter ν, is shown in Fig. 19.

The constraints on the intercept, α, and slopes, β and γ, are consistent with those derived by Park et al. (2017) for all three measures of emission line width – see Table 2. Park et al. (2017) find a systematically higher intrinsic scatter than that obtained using t-cup; however, these figures cannot be compared directly, as we do not know the exact statistical model used by Park et al. (2017).

6 CONCLUSIONS

We have presented a general-purpose approach to linear regression, implemented as t-cup, that is robust to model mis-specification, with a model laid out in Section 2. In Section 4, we demonstrated that the model recovers constraints consistent with those used to generate the data sets, including in cases where there is a mismatch between the generative model used to create the data set and our regression model. In Section 5, we compared the method to Kelly (2007) and Park et al. (2017) on real-world data, illustrating that the models derive consistent constraints.

It may be fruitful to re-examine some of the priors assumed in the t-cup model; while we focused on producing a model that was applicable to a large range of data sets, t-cup predicted lower intrinsic scatter than both Kelly (2007) and Park et al. (2017). While results between the three are not directly comparable as a result of the different assumptions in each model, this difference could suggest that the prior on σ_{68} has too much density near $\sigma_{68} = 0$. In addition, setting the prior on $\{\mathbf{x}_i\}$ using extreme deconvolution, while effective, is not theoretically motivated; a prior similar to that presented in Bartlett & Desmond (2023) may be more appropriate in this case.

In our next paper, we will apply the robust inference techniques presented here to exploring the single-epoch mass estimators used in estimating the masses of high-redshift quasars.

ACKNOWLEDGEMENTS

The authors would like to thank Alan Heavens, Stephen Feeney, Boris Leistedt, Jonathan Pritchard, Justin Alsing, Tamas Budavari, and Chad Schafer for useful conversations, and Tianning Lyu for a typographical correction. We also thank the anonymous referees for their comments and suggestions.

WM is supported by the Science and Technology Facilities Council (STFC) grant ST/T506151/1.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

DATA AVAILABILITY

The data in this paper is sourced from Kelly (2007) and Park et al. (2017). Scripts to generate the simulated data sets used to validate the model in Section 4 are available from the GitHub repository for this paper, at https://github.com/wm1995/tcup-paper.

REFERENCES

Aitkin M., Tunnicliffe Wilson G., 1980, Technometrics, 22, 325
Akritas M. G., Bershady M. A., 1996, ApJ, 470, 706
Andreon S., 2020, A&A, 640, A34
Andreon S., Hurn M. A., 2010, MNRAS, 404, 1922
Andreon S., Hurn M., 2013, Stat. Anal. Data Mining: ASA Data Sci. J., 9, 15
Andreon S., Weaver B., 2015, Bayesian Methods for the Physical Sciences. Learning from Examples in Astronomy and Physics, Vol. 4. Springer, Cham
Andreon S., Puddu E., de Propris R., Cuillandre J. C., 2008, MNRAS, 385, 979
Andrews D. F., Mallows C. L., 1974, J. R. Stat. Soc. Ser. B (Methodological), 36, 99
Andrews S. M., Rosenfeld K. A., Kraus A. L., Wilner D. J., 2013, ApJ, 771, 129
Avelino A., Friedman A. S., Mandel K. S., Jones D. O., Challis P. J., Kirshner R. P., 2019, ApJ, 887, 106
Bailey D. C., 2017, R. Soc. Open Sci., 4, 160600
Bartlett D. J., Desmond H., 2023, Open J. Astrophys., 6, 42
Bentz M. C. et al., 2013, ApJ, 767, 149
Berger J. O. et al., 1994, Test, 3, 5
Betancourt M. J., Girolami M., 2013, preprint (arXiv:1312.0906)
Bingham E. et al., 2019, J. Mach. Learn. Res., 20, 28:1
Blanton M. R. et al., 2003, ApJ, 592, 819
Bovy J., Hogg D. W., Roweis S. T., 2011, Ann. Appl. Stat., 5, 1657
Box G. E. P., Tiao G. C., 1968, Biometrika, 55, 119
Boyle B. J., Fong R., Shanks T., Peterson B. A., 1987, MNRAS, 227, 717
Boyle B. J., Shanks T., Peterson B. A., 1988, MNRAS, 235, 935
Chung Y., Rabe-Hesketh S., Dorie V., Gelman A., Liu J., 2013, Psychometrika, 78, 685
Cook S. R., Gelman A., Rubin D. B., 2006, J. Comput. Graph. Stat., 15, 675
Cox R. T., 1946, Am. J. Phys., 14, 1
Cramér H., 1946, Mathematical Methods of Statistics, Princeton Mathematical Series; 9. Princeton Univ. Press, Princeton
Ding P., 2014, J. Multivariate Anal., 124, 451
Eddington A. S., 1913, MNRAS, 73, 359
Efstathiou G., 2014, MNRAS, 440, 1138
Feeney S. M., Mortlock D. J., Dalmasso N., 2018, MNRAS, 476, 3861
Feigelson E. D., de Souza R. S., Ishida E. E. O., Jogesh Babu G., 2021, Annu. Rev. Stat. Appl., 8, 493
Ferrarese L., Merritt D., 2000, ApJ, 539, L9
Gagnon P., Hayashi Y., 2023, Stat. Probab. Lett., 193, 109693

Gagnon P., Desgagné A., Bédard M., 2020, Bayesian Anal., 15, 389
Gebhardt K. et al., 2000, ApJ, 539, L13
Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., Rubin D. B., 2013, Bayesian Data Analysis. Chapman and Hall/CRC, Boca Raton
Ghosh J. K., Delampady M., Samanta T., 2006, An Introduction to Bayesian Analysis. Springer-Verlag, Berlin
Hamura Y., Irie K., Sugasawa S., 2022, Comput. Stat. Data Anal., 173, 107517
Hamura Y., Irie K., Sugasawa S., 2024, Stat. Probab. Lett., 210, 110130
Hubble E., 1929, Proc. Natl. Acad. Sci., 15, 168
Jing T., Li C., 2025, AJ, 170,45
Johnston R., 2011, A&AR, 19, 41
Jontof-Hutter D. et al., 2016, ApJ, 820, 39
Juárez M. A., Steel M. F. J., 2010, J. Appl. Econom., 25, 1128
Kelly B. C., 2007, ApJ, 665, 1489
Kelly B. C., Fan X., Vestergaard M., 2008, ApJ, 682, 874
Kelly B. C., Vestergaard M., Fan X., 2009, ApJ, 692, 1388
Knuth K. H., Skilling J., 2012, Axioms, 1, 38
Leavitt H. S., Pickering E. C., 1912, Harvard Coll. Obs. Circ., 173, 1
Leistedt B., Mortlock D. J., Peiris H. V., 2016, MNRAS, 460, 4258
Malmquist K. G., 1922, Meddel. Lunds Astron. Obs. Ser. I, 100, 1
Mantz A. B., 2016, MNRAS, 457, 1279
McConnell N. J., Ma C.-P., 2013, ApJ, 764, 184
McElreath R., 2020, Statistical Rethinking: A Bayesian Course with Examples in R and Stan. Chapman and Hall/CRC, New York
Neal R. M., 2003, Ann. Stat., 31, 705
Park D., Barth A. J., Woo J.-H., Malkan M. A., Treu T., Bennert V. N., Assef R. J., Pancoast A., 2017, ApJ, 839, 93
Phan D., Pradhan N., Jankowiak M., 2019, preprint (arXiv:1912.11554)
Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical Recipes in C. The Art of Scientific Computing. Cambridge Univ. Press, Cambridge
Rao C. R., 1945, Bull. Calcutta Math. Soc., 37, 81
Riess A. G. et al., 2011, ApJ, 730, 119
Schechter P., 1976, ApJ, 203, 297
Sereno M., 2016, MNRAS, 455, 2149
Sestovic M., Demory B.-O., Queloz D., 2018, A&A, 616, A76
Sivia D. S., Skilling J., 2006, Data Analysis: A Bayesian Tutorial, 2nd edn. Oxford Science Publications, Oxford Univ. Press, Oxford
Stan Development Team, 2024, Stan User's Guide, v2.36. Available at: https://mc-stan.org
Tak H., Ellis J. A., Ghosh S. K., 2019, J. Comput. Graph. Stat., 28, 415
Talts S., Betancourt M., Simpson D., Vehtari A., Gelman A., 2018, preprint (arXiv:1804.06788)
Tremaine S. et al., 2002, ApJ, 574, 740
Van Horn K. S., 2003, Int. J. Approx. Reason., 34, 3

APPENDIX A: PRIOR CHOICE FOR SHAPE PARAMETER

As our model relies on Student's t-distributions, we review notation and priors used by previous works and justify our reasoning for our prior choice. One approach is to adopt a fixed value of ν, which is equivalent to setting a Dirac delta function prior: a common choice is $\nu = 4$ (e.g. Berger et al. 1994; Gelman et al. 2013). Another approach is to adopt a more flexible approach by allowing ν to vary (e.g. Juárez & Steel 2010; Gelman et al. 2013; Ding 2014; Park et al. 2017; Feeney et al. 2018).

We sought a flexible prior for this work that could reduce to a (nearly) normal distribution, but had sufficient flexibility to accommodate heavy-tailed distributions as well. Priors meeting this criterion include:

(i) priors of the form $\nu \sim \Gamma(\alpha, \beta)$ for some shape parameter, α, and rate parameter, β – e.g. Juárez & Steel (2010) uses $\{\alpha = 2, \beta = 0.1\}$; Ding (2014) uses $\{\alpha = 1, \beta = 0.1\}$

(ii) A uniform prior in $\frac{1}{\nu} \sim U(0, 1)$ (Gelman et al. 2013)

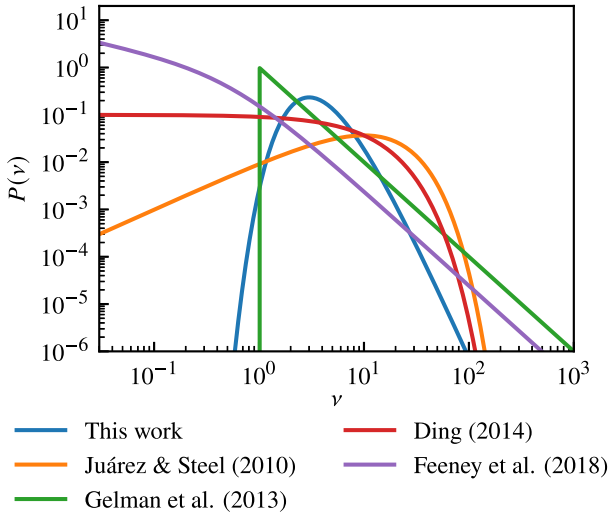

Figure A1. A selection of different priors on ν in works that have used t distributions.

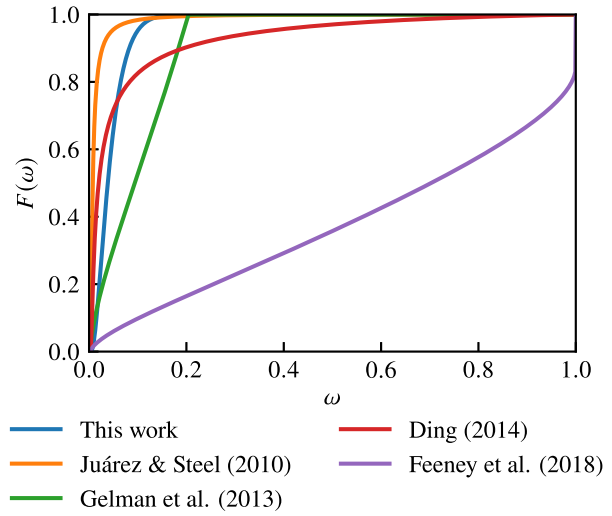

Figure A2. The CDF of priors on ν used in works that have used t-distributions. The priors are expressed in terms of outlier fraction, ω, as defined in equation (12).

(iii) A uniform prior[4] in distribution peak height relative to a normal distribution, t, such that

$$t \equiv \sqrt{\frac{2}{\nu}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \sim U(0, 1). \tag{A1}$$

These priors, along with the prior we adopted, are illustrated in Fig. A1. The cumulative distribution functions for the priors in terms of outlier fraction, ω, are illustrated in Fig. A2.

[4]Strictly, Feeney et al. (2018) approximate this prior with the closed form prior on shape parameter, ν, of

$$\mathcal{P}(\nu) \propto \frac{\Theta(\nu)}{\left(\left(\frac{\nu}{\nu_0}\right)^{1/(2a)} + \left(\frac{\nu}{\nu_0}\right)^{2/a}\right)^a},$$

where $\Theta(\cdot)$ is the Heaviside step function, and ν_0 and a are constants with values ~ 0.55 and ~ 1.2 respectively.

When testing different priors, we found that priors with significant density in the range $0 < \nu < 1$ led to sampling difficulties as ν approached 0; these low values of $\nu \sim 0.1$ can correspond to outliers that are more than 20 orders of magnitude larger than σ. In theory, these unphysical regions of parameter space ought to be excluded during the process of inference. However, the use of HMC in such cases can lead to divergences in the sampling process or inefficient sampling as the sampler struggles with regions of high curvature – see the discussion of this phenomenon in Neal (2003) and Betancourt & Girolami (2013). As discussed in Section 2.6, reparametrizing the Student's t-distribution as a mixture of normals was found to help with, but not alleviate, this problem.

Another option would be to place a prior is on outlier fraction ω. It could be argued that the term outlier loses its meaning when the majority of a data set is composed of so-called 'outliers'; therefore, a natural choice of prior might be a uniform distribution ranging from the normally distributed outlier fraction of ∼0.00270 to this 'maximum' outlier fraction of 0.5, corresponding to $\nu \sim 0.302$. This still leads to large outliers (some of 8 orders of magnitude for 100 draws from the distribution), which are unphysical and continue to present difficulties when sampling with HMC.

To limit the number of unphysical outliers, we can instead limit the prior to consider only distributions that are less heavy-tailed than the Cauchy distribution – i.e. all those with $\nu > 1$. This is the approach taken by Gelman et al. (2013), rendering regions of parameter space inaccessible [in the case of the Gelman et al. (2013) prior, $\nu = 1$ is the cutoff, which corresponds to a Cauchy distribution]. On the other hand, the priors used in Juárez & Steel (2010), Ding (2014), Feeney et al. (2018) have disproportionate prior density at low values of ν, which corresponds to models with outliers several orders of magnitude larger than the predicted effect size.

In this work, we use the prior

$$\nu \sim \text{Inv-}\Gamma(4, 15), \tag{A2}$$

where Inv-$\Gamma(\alpha, \beta)$ is an inverse gamma distribution with shape parameter, α, and scale parameter, β.

This prior was chosen for two reasons:

(i) the prior is smooth in ν with no sharp boundaries [unlike Gelman et al. (2013)]

(ii) the prior has density at a larger range of outlier fractions ω than that in Juárez & Steel (2010) but insignificant density at unrealistically high outlier fractions, in contrast with the priors in Ding (2014) and Feeney et al. (2018).

APPENDIX B: ESTIMATING VARIANCE TRADE-OFFS

In Section 2.5, we examined the increased standard deviation in parameter estimates using a toy model, showing the results of repeated trials alongside the Cramér-Rao bound (Rao 1945; Cramér 1946) in Fig. 5; here, we derive the formula for this bound.

For the toy model, we looked at the posterior distribution of the mean μ for a series of N points drawn from a normal distribution with zero mean and unit variance. The log likelihood, $\ell(\mu; \nu, \{x_i\})$, when analysed with a t-distribution, is

$$\ell(\mu; \nu, \{x_i\}) = -\sum_i \frac{\nu+1}{2} \log\left(1 + \frac{(x_i-\mu)^2}{\nu}\right) + C, \tag{B1}$$

where C is a constant. Thus, the Fisher information is

$$\mathcal{F}(\mu; \nu, \{x_i\}) = (\nu+1) \sum_i \frac{\nu - (x_i-\mu)^2}{\nu + (x_i-\mu)^2}. \tag{B2}$$

We can then evaluate the Cramér-Rao bound at the true mean $\mu = 0$, marginalizing over the particular data set $\{x_i\}$, to derive the variance in the estimate

$$\text{Var}(\hat{\mu}) \geq \frac{1}{N(\nu+1)\left(1 - \sqrt{\frac{\nu}{2}\pi}\exp\left(\frac{\nu}{2}\right)\text{erfc}\left(\sqrt{\frac{\nu}{2}}\right)\right)}, \tag{B3}$$

where $\text{erfc}(x) = 1 - 2/\sqrt{\pi}\int_0^x \exp(-t^2)\,dt$ is the complementary error function. This bound is used to derive the posterior standard deviation ratio plotted in Fig. 5.

APPENDIX C: FIXED-VALUE CALIBRATION DATA MODELS

Here, we fully specify the data set models used for fixed-value calibration tests in Section 4.

C1 t-distributed data

In Section 4.1, the fixed-value test data sets have $N = 20$ data points drawn from the following distribution:

$$x_i \sim \mathcal{N}(\mu = 2, \sigma^2 = 4) \tag{C1}$$

$$y_i \sim t_3(\mu = 3 + 2x_i, \sigma^2 = 0.01/\sigma_{68}^2(\nu = 3)) \tag{C2}$$

$$\log_{10}\sigma_{x,i} \sim \mathcal{N}(\mu = -1, \sigma^2 = 0.01) \tag{C3}$$

$$\log_{10}\sigma_{y,i} \sim \mathcal{N}(\mu = -0.7, \sigma^2 = 0.01) \tag{C4}$$

$$\hat{x}_i \sim \mathcal{N}(\mu = x_i, \sigma^2 = \sigma_{x,i}^2) \tag{C5}$$

$$\hat{y}_i \sim \mathcal{N}(\mu = y_i, \sigma^2 = \sigma_{y,i}^2). \tag{C6}$$

C2 Normally distributed data with an outlier

In Section 4.2, we generated data sets of $N = 12$ points using the model:

$$x_i \sim \mathcal{N}(\mu = 5, \sigma^2 = 9) \tag{C7}$$

$$y_i \sim \begin{cases} \mathcal{N}(\mu = 3 + 2x_i - 10, \sigma^2 = 0.04) & \text{for the second-largest } x_i \\ \mathcal{N}(\mu = 3 + 2x_i, \sigma^2 = 0.04) & \text{otherwise} \end{cases} \tag{C8}$$

$$\log_{10}\sigma_{x,i} \sim \mathcal{N}(\mu = -0.5, \sigma^2 = 0.01) \tag{C9}$$

$$\log_{10}\sigma_{y,i} \sim \mathcal{N}(\mu = -0.3, \sigma^2 = 0.01) \tag{C10}$$

$$\hat{x}_i \sim \mathcal{N}(\mu = x_i, \sigma^2 = \sigma_{x,i}^2) \tag{C11}$$

$$\hat{y}_i \sim \mathcal{N}(\mu = y_i, \sigma^2 = \sigma_{y,i}^2). \tag{C12}$$

C3 2D normal mixture model

We introduce the parameter O_i to indicate whether the ith data point is drawn from the core distribution (in which case, $O_i = 0$) or from the outlier distribution (for which $O_i = 1$).

In Section 4.4, we generated a data set of $N = 200$ points using the model:

$$\boldsymbol{x}_i \sim \begin{cases} \mathcal{N}\left(\mu = \begin{pmatrix} -3 \\ 2 \end{pmatrix}, \Sigma^2 = \begin{pmatrix} 0.5 & -1 \\ -1 & 4 \end{pmatrix}^2\right) & 1 \leqslant i \leqslant 140 \\ \mathcal{N}\left(\mu = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \Sigma^2 = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}^2\right) & 140 < i \leqslant 200 \end{cases} \tag{C13}$$

$$O_i \sim \text{Bernoulli}(0.1) \tag{C14}$$

$$y_i \sim \begin{cases} \mathcal{N}(\mu = 2 + (3,1)^T \cdot x_i, \sigma^2 = 0.16) & O_i = 0 \\ \mathcal{N}(\mu = 2 + (3,1)^T \cdot x_i, \sigma^2 = 16.0) & O_i = 1 \end{cases} \tag{C15}$$

$$\Sigma_{x,i} \sim \mathcal{W}_2(\boldsymbol{V} = 0.1\mathbb{I}, n = 3) \tag{C16}$$

$$\log_{10}\sigma_{y,i} \sim \mathcal{N}(\mu = -1, \sigma^2 = 0.01) \tag{C17}$$

$$\hat{x}_i \sim \mathcal{N}(\mu = \boldsymbol{x}_i, \Sigma^2 = \Sigma^2_{x,i}) \tag{C18}$$

$$\hat{y}_i \sim \mathcal{N}(\mu = y_i, \sigma^2 = \sigma^2_{y,i}), \tag{C19}$$

where $\mathcal{W}_2$ denotes a Wishart distribution over 2x2 matrices and $\mathbb{I}$ denotes the 2x2 identity matrix.

C4 Laplace-distributed data

In Section 4.5, we generate $N = 25$ data points under the following model:

$$x_i \sim \mathcal{U}(-5, 5) \tag{C20}$$

$$y_i \sim \text{Laplace}(\mu = -1 + 0.8x_i, b = 0.2) \tag{C21}$$

$$\log_{10}\sigma_{x,i} \sim \mathcal{N}(\mu = -1, \sigma^2 = 0.01) \tag{C22}$$

$$\log_{10}\sigma_{y,i} \sim \mathcal{N}(\mu = -1, \sigma^2 = 0.01) \tag{C23}$$

$$\hat{x}_i \sim \mathcal{N}(\mu = x_i, \sigma^2 = \sigma^2_{x,i}) \tag{C24}$$

$$\hat{y}_i \sim \mathcal{N}(\mu = y_i, \sigma^2 = \sigma^2_{y,i}). \tag{C25}$$

This paper has been typeset from a TEX/LATEX file prepared by the author.