Calibration Concordance for Astronomical Instruments via Multiplicative Shrinkage

Yang Chen,[*] Xiao-Li Meng,[†] Xufei Wang,[‡] David A. van Dyk,[§]
Herman L. Marshall,[¶] Vinay L. Kashyap[‖]

September 21, 2018

Abstract

   Calibration data are often obtained by observing several well-understood objects simultaneously with multiple instruments, such as satellites for measuring astronomical sources. Analyzing such data and obtaining proper concordance among the instruments is challenging when the physical source models are not well understood, when there are uncertainties in "known" physical quantities, or when data quality varies in ways that cannot be fully quantified. Furthermore, the number of model parameters increases with both the number of instruments and the number of sources. Thus, concordance of the instruments requires careful modeling of the mean signals, the intrinsic source differences, and measurement errors. In this paper, we propose a log-Normal model and a more general log-t model that respect the multiplicative nature of the mean signals via a half-variance adjustment, yet permit imperfections in the mean modeling to be absorbed by residual variances. We present analytical solutions in the form of power shrinkage in special cases and develop reliable Markov chain Monte Carlo (MCMC) algorithms for general cases, both of which are available in the `Python` module *CalConcordance*. We apply our method to several datasets including a combination of observations of *active galactic nuclei* (AGN) and spectral line emission from the *supernova remnant* E0102, obtained with a variety of X-ray telescopes such as *Chandra*, XMM-*Newton*, *Suzaku*, and *Swift*. The data are compiled by the *International Astronomical Consortium for High Energy Calibration* (IACHEC). We demonstrate that our method provides helpful and practical guidance for astrophysicists when adjusting for disagreements among instruments.

Keywords: Adjusting attributes; shrinkage estimator; Bayesian hierarchical model; log-Normal model; half-variance adjustment; log-t model.

[*]Yang Chen is Assistant Professor, Department of Statistics and Michigan Institute for Data Science (MIDAS), University of Michigan, Ann Arbor, MI 48109; email: ychenang@umich.edu.

[†]Xiao-Li Meng is Whipple V. N. Jones Professor of Statistics, Harvard University, Cambridge, MA 02138.

[‡]Xufei Wang was a Ph.D. candidate, Department of Statistics, Harvard University, Cambridge, MA 02138.

[§]David A. van Dyk is a Professor of Statistics and Head of the Department of Mathematics at Imperial College London, London, UK SW7 2AZ.

[¶]Herman Marshall is Astrophysicist, MIT Kavli Institute, Cambridge, MA 02139.

[‖]Vinay Kashyap is Astrophysicist, Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138.

1 Introducing Calibration Concordance

The calibration of instruments is fundamental for comparing or combining measurements obtained with different instruments. Typically, calibration is conducted by using each of several instruments to measure one or more well-understood objects, e.g., astronomical sources. The resulting data are used to develop adjustments that can be applied to future observations to obtain reliable absolute measurements. Convenient adjustments, such as ad hoc affine or ratio adjustments, however, often result in poor calibration, and without justifiable quantification of the calibration error that is essential for assessing the uncertainty of the final estimates of interest. The main difficulty of deriving reliable adjustments for instruments springs from the variations that are intrinsic to the sources and to the instruments, in addition to individual measurement errors.

First, the physical models, derived using various approximations based on scientists' current understandings of the instruments, may not be as reliable as we hope. Second, "known" physical quantities are typically estimates themselves; even when their estimated errors are available, standard plug-in estimators and error propagation techniques may lead to biased and often overly optimistic results. Third, data quality varies in ways that cannot be fully quantified, especially across instruments or in the presence of outliers. Last, the number of unknown model parameters increases with the number of instruments and the number of sources, leading to well-known model challenges. Together these challenges and subtleties expose that, although calibration problems have a long history, principled statistical adjustments are not in routine use or even understood.

This paper attempts to fill this gap for a variety of astronomical instruments, by developing hierarchical models that respect the physical models for the mean signals, while permitting the modeling imperfections to be captured by residual variances. We build effective fitting algorithms and a software package, *CalConcordance*, which are used to test our models via simulated data, and then applied to several datasets from the *International Astronomical Consortium for High Energy Calibration* (IACHEC). The intended readers are both statisticians and astrophysicists.

1.1 Calibration Concordance for Astronomical Instruments

In astrophysics, various instruments such as telescopes are used by different teams of scientists to understand intrinsic properties of astronomical objects, i.e., sources such as stars. Although it

is possible to make relative comparisons of different sources observed with the same instrument, unless the instruments are properly calibrated (Sembay et al., 2010), we cannot make reliable absolute measurements or make comparisons of sources observed with different instruments. Therefore, calibration of different instruments is an important, and on-going, problem for astrophysicists (e.g., Seward, 1992; Matthews and Havey, 2010; Nevalainen et al., 2010; Tsujimoto et al., 2011; Read et al., 2014; Schellenberger et al., 2015; Madsen et al., 2016).

As an example, space-based (e.g., X-ray) telescope calibration (Schwartz, 2014) is handled in two phases: first, under controlled laboratory conditions ("ground" calibration), and second, while in space using astrophysical sources ("in-flight" calibration; see Guainazzi et al. (2015)). At each phase, the same set of well-understood sources is observed with multiple instruments. The intent of in-flight calibration is usually to verify ground calibration but imperfect laboratory conditions and evolving instrument characteristics (while in-flight) may result in discrepancies between different telescopes. The task of developing reliable adjustments for astronomical instruments based on observing multiple sources with multiple instruments is known as the *calibration concordance problem*, which aims to develop a *concordance* in the calibration among these astronomical instruments. This paper aims to provide a statistically principled solution and it is a joint effort between statisticians and astrophysicists, both of whose expertise are critical for appropriately quantifying the uncertainties while incorporating scientific knowledge and judgments.

For the calibration problem discussed in this paper, the following two concepts are essential.

- **Flux of an astronomical source.** The absolute flux is the quantity of luminous energy incident upon the aperture of a telescope per unit area per unit time. The absolute flux of an astronomical source depends on the luminosity of the object and its distance from the Earth, both of which are intrinsic to the object. For a fixed source spectrum, i.e., the distribution of photon energies, the *measured flux* is proportional to the number of photons *detected* by an astronomical instrument. If the spectrum changes, or the detector on the instrument changes, then so do the number of detected photons and the measured flux.

- **Effective Area for an instrument.** The geometric area of a telescope (instrument) is an upper bound on its capacity to collect photons. Many factors can reduce the efficiency of photon collection, including mirror reflectivity, structural obscuration, filter transmission,

detector sensitivity, etc. This reduction in efficiency is also photon-energy dependent. The *Effective Area* is the equivalent geometric size of an ideal detector that would have the same collection capability and it is empirically measured or theoretically calculated and tabulated as a function of energy. An instrument's Effective Area is used to estimate the absolute flux of an astronomical source given its measured flux: the estimated absolute flux is the measured flux divided by the Effective Area. Since the Effective Area varies with photon-energy, astronomers often compare different energy bands in the way that we describe comparing different instruments, a convention we also adopt (George et al., 1992; Graessle et al., 2006).

The calibration problem arises because the Effective Areas of the instruments are not known precisely (Drake et al., 2006; Kashyap et al., 2008; Lee et al., 2011; Xu et al., 2014), and hence different instruments can yield substantially different *estimates of absolute fluxes* for the same unvarying source even after accounting for the measurement uncertainties in measured fluxes. This is manifested in Figure 1, which shows the logarithm of estimates of absolute fluxes of three sources (panels 1-3) using three instruments, "pn", "MOS1" and "MOS2", taken from the XCAL data that we describe in detail in Section 4.3. Therefore, the problem of calibration concordance among
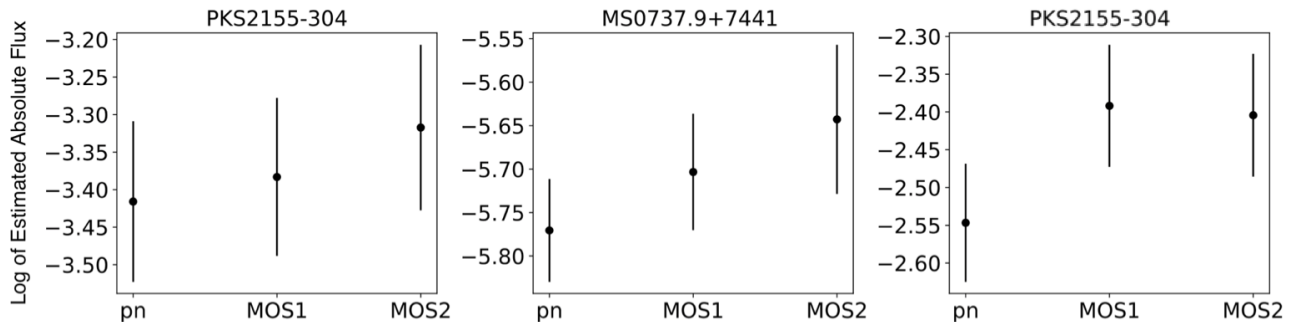

Figure 1: Using (natural) logarithm of measured fluxes, correcting for existing "known" Effective Areas, to estimate log absolute fluxes. Measured fluxes are collected with three instruments "pn, MOS1, MOS2" for each of three sources, labeled on top of each panel; see Section 4.3. Estimates are given by the dots with approximate 95% confidence intervals. The differences among the estimates are particularly pronounced in the third panel. The estimates from "pn" is systematically smaller than those from "MOS1" and "MOS2", illustrating the need for adjustments.

different instruments is equivalent to reliably estimating the Effective Area of each instrument. By *reliably* we mean that, after proper adjustments of the Effective Areas, instruments measuring a common source should agree within *stated and scientifically acceptable* statistical uncertainty on the absolute flux of each source.

1.2 A Multiplicative Physical Model

Suppose we observe photon counts, $\{c_{ij}\}$, where i indexes N instruments and j indexes M objects/sources. The observed photon count c_{ij} is known to follow a Poisson model with intensity C_{ij}, which is affected by the Effective Area A_i and flux F_j as follows. Because source fluxes have units of expected photons per second and per square centimeter, they are multiplied by instrument Effective Areas and a known factor, denoted by T_{ij}, to obtain expected photon counts:

$$C_{ij} = T_{ij}A_iF_j, \quad 1 \le i \le N, \quad 1 \le j \le M. \tag{1.1}$$

The multiplicative constant T_{ij} contains the exposure time, as well as other factors that can be calculated approximately by astrophysicists such as corrections for enclosed energy fractions and spectral shape correction factors; see Marshall et al. (2018) for details. With this in mind, we can regard T_{ij} as a fixed known constant, and any real uncertainties related to T_{ij} can be partially captured by our residual modeling discussed later in this paper and in subsequent work. Intuitively, the A_i can be regarded as a measure of the efficiency of instrument i in terms of photon collection. Fundamentally, (1.1) presumes that the Effective Area for a particular instrument remains the same regardless of which source it is applied to (and vice versa). By using a more homogeneous subgroup of sources (see Section 1.3), we can increase the applicability of (1.1), as we illustrate in Section 4.1. Of course, there is no free lunch – by using a subgroup instead of all M of the sources, we have less data and hence higher variability of our estimator, a bias-variance trade-off.

Prior to observing $\{c_{ij}\}$, astronomers obtain initial estimates a_i of A_i from ground-based or in-flight calibration measurements, and hence it is safe to assume these measurements are independent of $\{c_{ij}\}$. Comparing with estimated fluxes of well-understood sources, astronomers can also place a reasonable prior bound on the margin of relative error in a_i at about 20% (Lee et al., 2011; Drake et al., 2006). Additional prior knowledge on the measurement errors in $\{c_{ij}\}$ is available.

How to utilize this prior information, and whether these estimated uncertainties suffice to explain the variations in the data, are among the questions that we investigate in this paper.

1.3   Sample Selection Mechanism

The sample selection mechanism, which involves both the selection of sources and the instrument used to observe each source is important because a biased selection mechanism can lead to misleading results. However, this is not a large concern in our setting for several reasons.

First, the instruments we consider share a broad common energy passband so an object observed with one instrument will likely be seen with another. Indeed, for all the datasets we analyze in this paper, each source is observed by each instrument, although this is not a requirement for our methods. Second, the chance of a source not being observed because it is too faint is low. Since dim sources are not used for calibration, we do not include them in the study. Instead, we include sources with well-understood energy spectra, high intrinsic intensity, and stable spectral-temporal variations. This selection of sources is favorable since our ultimate goal is to calibrate the instruments and each of their Effective Areas is invariant to source fluxes. Furthermore, the vagaries of scheduling introduces large variations in the completeness achievable in a fleet of spacecraft, but this selection bias is negligible. Each spacecraft has several independent intrinsic constraints that are related to the shape of its orbit (Chandra X-ray Center, 2017b); its Sun, Moon, and Earth avoidance angles; and even its thermal environment histories; making scheduling of simultaneous observations difficult (Chandra X-ray Center, 2017a). In other words, the missingness in the observation matrix, if any, is due to factors that are irrelevant to the intrinsic property of the sources or instruments, i.e., the estimands. These considerations permit us to ignore the sample selection mechanism in the sense of Rubin (1976).

The remainder of this paper is organized into 4 sections. Section 2 describes a statistical model for calibration concordance, a log-Normal model, and extends it to a more general log-t model to handle outliers. Using simulated and real data, Sections 3 and 4 assess and verify the empirical performance of our methods. Section 5 briefly discusses a likelihood approach and its connection to our Bayesian approach, and future work. All numerical results are reproducible using the *Python* code and data available on GitHub at https://github.com/astrostat/Concordance.

2 Building and Fitting the Proposed Concordance Models

2.1 Modeling Multiplicative Means

To make distinctions between observed quantities (e.g., estimator) and unknown quantities (e.g., estimand) clear, we adopt the convention that the former is denoted by lowercase (Roman) letters and the latter by uppercase, whenever feasible. We express (1.1) as

$$\log C_{ij} - \log T_{ij} = B_i + G_j, \text{ where } B_i = \log A_i \text{ and } G_j = \log F_j. \tag{2.1}$$

While this is a trivial relationship among the *estimands*, it does not hold for their corresponding *estimators*. In fact, if we let $y_{ij} = \log c_{ij} - \log T_{ij}$ (and ignore the issue of $c_{ij} = 0$ for the moment), $b_i = \log a_i$ and $g_j = \log f_j$, we cannot simultaneously expect that $y_{ij} = b_i + g_j + \epsilon_{ij}$ *and* that ϵ_{ij} is independent of $\{b_i, g_j\}$ with mean zero. If both were true, it would imply (incorrectly) that the expectation of y_{ij} is determined by b_i and g_j, rather than by their respective estimands: B_i and G_j. Table 2.1 gives a summary of the notation used in this section.

The quantity that we observe and aim to model is $y_{ij} = \log c_{ij} - \log T_{ij}$, assuming $c_{ij} > 0$. (The case when $c_{ij} = 0$, which never occurs in our data, is discussed below.) We assume that the measurement error in c_{ij} for C_{ij} is multiplicative (i.e., in terms of a percentage), which results in additive errors on the log-scale. Thus we postulate the regression model

$$y_{ij} = -0.5 \ \sigma_i^2 + B_i + G_j + e_{ij}, \quad e_{ij} \stackrel{\text{indep}}{\sim} \mathcal{N}(0, \sigma_i^2), \tag{2.2}$$

where $-0.5 \ \sigma_i^2$ is a half-variance correction for the multiplicative mean modeling in (1.1). This correction ensures that $\mathrm{E}(c_{ij}) = C_{ij}$ because if $\log x \sim N(\mu, v)$, then $E(x) = e^{0.5v + \mu}$. Consequently, $\mathrm{E}(c_{ij}) = T_{ij}\mathrm{E}(e^{y_{ij}}) = T_{ij}e^{0.5\sigma_i^2 - 0.5\sigma_i^2}e^{B_i}e^{G_j} = C_{ij}$. For convenience, when (and *only when*) σ_i^2 is known, we treat $y_{ij}' = y_{ij} + 0.5\sigma_i^2$ as data. Since $b_i = \log a_i$ is an initial estimate of B_i that is available without access to the calibration data $\{c_{ij}\}$, we formulate it as the prior mean for B_i via $B_i \stackrel{\text{indep}}{\sim} \mathcal{N}(b_i, \tau_i^2)$, where τ_i is provided by astronomers as well.

Given the underlying Poisson nature of the photon counts, c_{ij}, the log-Normal model in (2.2) deserves some explanation. If the expected counts, C_{ij}, are reasonably large, the log-Normal model approximates the Poisson model well. This is what we expect in practice since calibration sources are typically relatively bright, as illustrated in our datasets in Section 4. However, the

	Counts	Effective Area	Flux	Log-Data	Correction
Estimand (Parameter)	C_{ij}	$A_i = \exp(B_i)$	$F_j = \exp(G_j)$		$-0.5\sigma_i^2$
Estimate (Data)	c_{ij}	$a_i = \exp(b_i)$		y_{ij}	T_{ij}
Relationship	$E(c_{ij}) = C_{ij}$	$B_i \sim \mathcal{N}(b_i, \tau_i^2)$	$C_{ij} = A_i F_j$	$y_{ij} = \log(\frac{c_{ij}}{T_{ij}})$	

Table 1: Summary of notations for log-Normal model. The index i ranges from 1 to N and the index j ranges from 1 to M.

primary reason we adopt this approximation is because the log-Normal model permits separate modeling considerations for the mean and variance of the (transformed) counts, whereas the Poisson mean dictates the Poisson variance. This flexibility is especially important when the mean model (1.1) is not perfectly specified, as we expect since the T_{ij} are estimated or approximated in practice (Marshall et al., 2018). The variance of the log-Normal model can (partially) capture imperfections in the mean model; see Section 3 for discussions in our numerical experiments.

Because the log-Normal model works with the log counts, it cannot directly accommodate zero counts; we observe zero counts in some of our simulations studies. Should an observed count of zero be observed we suggest it be replaced by a pseudo count of $c_{ij} = 0.5$. This is a standard strategy (Bilder and Loughin, 2014, p. 42) known as a zero-modified Poisson; the mean and variance of a zero-modified Poisson random variable approximate those of the corresponding Poisson well if the mean of the Poisson is reasonable large.[1] We validate this strategy in our simulation studies.

In (2.2), the variance for the measurement error is assumed to depend only on the instrument. This assumption works reasonably well in our applied examples. More generally, each e_{ij} can have its own variance, σ_{ij}^2, but obviously some constraints are needed in order to ensure identifiability. Other possible constraints include forcing the variances to be source-dependent only or to be additive, i.e., $\sigma_{ij}^2 = \omega_i^2 + \lambda_j^2$. We first consider a *known variance model* because astronomers provide

---

[1]For a Poisson random variable q with mean λ, this replacement leads to a zero-modified Poisson random variable $\tilde{q}$ with mean and variance

$$\mathrm{E}(\tilde{q}) = \lambda + 0.5e^{-\lambda}, \qquad \mathrm{Var}(\tilde{q}) = \lambda(1 - e^{-\lambda}) + 0.25e^{-\lambda}(1 - e^{-\lambda}). \qquad (2.3)$$

With reasonably large λ, $\tilde{q}$ approximates q extremely well.

best guesses of σ_{ij}^2. However, as illustrated in subsequent sections, the *unknown variance model* is more flexible, robust, and hence recommended in practice. This is because the inferred adjustment of Effective Areas could be either overly-optimistic or overly-conservative if the specified σ_{ij}^2 are inaccurate. Unfortunately, this is often the case in practice owing to an incomplete quantification of measurement uncertainties or incomplete understanding of data preprocessing.

2.2 Log-Normal Hierarchical Model and Its Posterior Sampling

Embedding the log-Normal model (2.2) into a Bayesian hierarchical model requires a prior for G_j. Because astronomers do not know enough about the physical processes to place an informative prior on the fluxes, they prefer to use a flat prior for the log-scale flux, i.e., G_j, on the grounds that astronomical source fluxes cover many orders of magnitude in dynamic range (Appenzeller, 2012). When the σ_i^2 are treated as unknown, we adopt independent Inverse-Gamma distributions with shape parameter α and scale parameter β, the values of which are chosen to reflect the astronomers' prior knowledge about the approximate scale of noise levels. Specifically, we assume

$$y_{ij} \mid \boldsymbol{B},\ \boldsymbol{G},\ \boldsymbol{\sigma^2} \overset{\text{indep}}{\sim} \mathcal{N}\left(-0.5\sigma_i^2 + B_i + G_j,\ \sigma_i^2\right), \tag{2.4}$$

$$\sigma_i^2 \overset{\text{indep}}{\sim} \text{Inv-Gamma}(\alpha,\ \beta), \quad B_i \overset{\text{indep}}{\sim} N(b_i,\ \tau_i^2), \quad \text{and} \quad G_j \overset{\text{indep}}{\sim} \text{flat prior},$$

where $\boldsymbol{B} = (B_1, \ldots, B_N)^\top$, $\boldsymbol{G} = (G_1, \ldots, G_M)^\top$, $\boldsymbol{\sigma^2} = (\sigma_1^2, \ldots, \sigma_N^2)^\top$, and $\boldsymbol{\tau^2} = (\tau_1^2, \ldots, \tau_N^2)^\top$, and $\top$ denotes the usual transpose. Under (2.4), we can show that (see Appendix C.1) the posterior distribution is proper with the weakest condition possible: each source is observed by at least one instrument. This theoretical guarantee is especially important because the number of parameters, $2N + M$, varies with the number of observations, NM. Furthermore, the MAP (maximum-a-posterior) estimator of each σ_i^2 is bounded away from zero by a constant which depends only on the hyperparameters and the total number of sources (see Section 2.3). Last, the use of proper conjugate priors for $\boldsymbol{\sigma^2}$ avoids the problem of an unbounded posterior distribution, which can occur when we use uniform prior distributions for $\boldsymbol{\sigma^2}$. We also remark that, because each σ_i^2 enters both the variance and the mean in (2.4), the impact of the choice of prior on the posterior inference is nuanced, as we discuss in the context of astrophysical applications (see Section 4.1).

In general, we let J_i be the set of indexes of the objects observed by detector i and I_j be the set

of indexes of the instruments that observe object j, and hence they accommodate missing data. Under (2.4), the posterior density of $\{\boldsymbol{B}, \boldsymbol{G}, \boldsymbol{\sigma^2}\}$, if it exists, is proportional to

$$\left[\prod_{i=1}^{N} \sigma_i^{-|J_i|-2-2\alpha}\right] \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\sum_{j\in J_i}\frac{(y_{ij}+0.5\sigma_i^2-B_i-G_j)^2}{\sigma_i^2} - \sum_{i=1}^{N}\left[\frac{(b_i-B_i)^2}{2\tau_i^2}+\frac{\beta}{\sigma_i^2}\right]\right\}. \quad (2.5)$$

This implies that the conditional distribution of the column vector $\boldsymbol{\theta} \equiv (\boldsymbol{B}^\top, \boldsymbol{G}^\top)^\top$ given $\boldsymbol{\sigma^2}$ is an $(N+M)$-dimensional Normal density. A simple way of deriving the mean $\boldsymbol{\mu(\sigma^2)}$ and covariance $\boldsymbol{\Sigma(\sigma^2)}$ of this conditional distribution is to use partial derivatives of $L(\boldsymbol{\theta}, \boldsymbol{\sigma^2})$, the logarithm of the joint posterior density given in (2.5). Let

$$\boldsymbol{\gamma(\sigma^2)} = \frac{\partial L(\boldsymbol{\theta}, \boldsymbol{\sigma^2})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{0}} \quad \text{and} \quad \boldsymbol{\Omega(\sigma^2)} = -\frac{\partial^2 L(\boldsymbol{\theta}, \boldsymbol{\sigma^2})}{\partial \boldsymbol{\theta}^2}\bigg|_{\boldsymbol{\theta}=\boldsymbol{0}}. \quad (2.6)$$

Note the second "$|_{\boldsymbol{\theta}=\boldsymbol{0}}$" is cosmetic because for Normal model the Fisher information is free of $\boldsymbol{\theta}$.

By the form of the Normal density, $\boldsymbol{\mu(\sigma^2)} = \boldsymbol{\Omega^{-1}(\sigma^2)\gamma(\sigma^2)}$ and $\boldsymbol{\Sigma(\sigma^2)} = \boldsymbol{\Omega^{-1}(\sigma^2)}$. Evaluating these derivatives yields $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N, \gamma_{N+1}, \ldots, \gamma_{N+M})^\top$ and $\boldsymbol{\Omega}$ as functions of $\boldsymbol{\sigma^2}$:

$$\gamma_i = \frac{\sum_{j\in J_i} y_{ij}}{\sigma_i^2} + \frac{b_i}{\tau_i^2} + 0.5|J_i|, \ i=1,\ldots N, \qquad \gamma_{j+N} = \sum_{i\in I_j}\frac{y_{ij}}{\sigma_i^2} + 0.5|I_j|, \ j=1,\ldots,M, \quad (2.7)$$

$$\boldsymbol{\Omega(\sigma^2)} = \begin{pmatrix} \boldsymbol{D_N} & \boldsymbol{DR} \\ \boldsymbol{R^\top D} & \boldsymbol{D_M} \end{pmatrix}, \quad \text{where} \quad \begin{aligned} \boldsymbol{D_N} &= \text{Diag}\{|J_i|\sigma_i^{-2} + \tau_i^{-2}, \ i=1,\ldots,N\}, \\ \boldsymbol{D_M} &= \text{Diag}\{\sum_{i\in I_j}\sigma_i^{-2}, \ j=1,\ldots,M\}, \end{aligned} \quad (2.8)$$

$\boldsymbol{D} = \text{Diag}\{\sigma_i^{-2}, \ i=1,\ldots,N\}$, and $\boldsymbol{R} = \{r_{ij}\}$, with $\boldsymbol{R}$ the $N \times M$ data "recoding matrix", i.e., $r_{ij} = 1$ if source j is observed with instrument i, and $r_{ij} = 0$ otherwise. When all the instruments measure all the sources, as in all our applications, $\boldsymbol{R}$ and $\boldsymbol{DR}$ are rank-one. In such cases, the inverse of $\boldsymbol{\Omega(\sigma^2)}$ can be calculated analytically, as seen in Appendix D.

When $\boldsymbol{\sigma^2}$ is unknown, its marginal posterior density can be obtained by evaluating the identity $P(\boldsymbol{\sigma^2}) = P(\boldsymbol{\theta}, \boldsymbol{\sigma^2})/P(\boldsymbol{\theta}|\boldsymbol{\sigma^2})$ at $\boldsymbol{\theta} = \boldsymbol{0}$, where the numerator is given in (2.5) and the denominator is given by the conditional Normal distribution with mean and variance given in (2.6). For ease of notation, we use $P(\cdot)$ as simplified notation for the posterior density $P(\cdot \mid \{y_{ij}\})$. In particular, noting that $P(\boldsymbol{\theta} = \boldsymbol{0}|\boldsymbol{\sigma^2}) \propto \sqrt{|\boldsymbol{\Omega(\sigma^2)}|}\, e^{-\boldsymbol{\mu}^\top(\boldsymbol{\sigma^2})\boldsymbol{\Omega(\sigma^2)}\boldsymbol{\mu(\sigma^2)}/2}$, $P(\boldsymbol{\sigma^2})$ is proportional to

$$\prod_{i=1}^{N} \sigma_i^{-|J_i|-2-2\alpha}\frac{1}{\sqrt{|\boldsymbol{\Omega(\sigma^2)}|}} \exp\left\{\frac{1}{2}\boldsymbol{\mu}^\top(\boldsymbol{\sigma^2})\boldsymbol{\Omega(\sigma^2)}\boldsymbol{\mu(\sigma^2)} - \sum_{i=1}^{N}\left[\frac{\sum_{j\in J_i} y_{ij}^2 + 2\beta}{2\sigma_i^2} + \frac{|J_i|\sigma_i^2}{8}\right]\right\}. \quad (2.9)$$

Because this is not a standard distribution, numerical methods are required. We can obtain a Monte Carlo sample from the joint posterior in one of several ways, including applying an MCMC algorithm to sample $\{\boldsymbol{B}, \boldsymbol{G}, \boldsymbol{\sigma}^2\}$ jointly, or sampling $\boldsymbol{\sigma}^2$ from (2.9) via rejection sampling and $\{\boldsymbol{B}, \boldsymbol{G}\}$ from its conditional Normal distribution given $\boldsymbol{\sigma}^2$. Incidentally, a good rejection proposal density is a convenient independent-component inverse Gamma distribution found in the proof of the posterior propriety; see Appendix C.1. The latter strategy is very efficient, especially as it provides independent draws. However, it is less flexible when we extend the model (e.g., the log-t extension of Section 2.4). Consequently, we adopt the more flexible MCMC approach.

Since the dimension of the parameter space, $2N + M$, is typically large for calibration purposes and the parameters are highly correlated, we use a Hamiltonian Monte Carlo (HMC) algorithm (Neal, 2011), which delivers a less correlated sample than do other MCMC techniques (e.g. Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984). We implement HMC using the STAN package in *Python* (Hoffman and Gelman, 2014; Stan Development Team, 2015, 2016), along with a blocked Gibbs sampler as an independent cross-check of STAN. In the blocked Gibbs sampling, we sample $\{\boldsymbol{B}, \boldsymbol{G}\}$ jointly to improve mixing as opposed to one-at-a-time Gibbs sampling, thanks to the joint normality of $\{\boldsymbol{B}, \boldsymbol{G}\}$ conditioning on $\boldsymbol{\sigma}^2$; see Appendix B for details.

As is well known, computational efficiency for posterior sampling often is affected by modeling defects, such as near non-identifiability (e.g., Meng, 2018). Model (2.4) does not suffer from this as long as τ_i is not too large compared to the magnitude of σ_i; see Appendix C.2.

2.3 Building Intuition: Power Shrinkage and Variance Shrinkage

To communicate our model aims clearly, when $\boldsymbol{\sigma}^2$ is known, we express the MAP estimators of $\boldsymbol{\theta}$ in terms of the usual linear shrinkage estimators (Efron and Morris, 1975; Morris, 1983) of $\boldsymbol{B}$ and $\boldsymbol{G}$. Intuitively, shrinkage estimators combine information from all the instruments and sources as well as experts' prior information through weighted averages, which serve the purpose of calibration concordance across instruments and sources well. Specifically, by setting the derivative of the log posterior in (2.5) with respect to $\boldsymbol{B}$ and $\boldsymbol{G}$ to be zero, we find that the MAP estimators, conditional on $(\boldsymbol{\tau^2}, \boldsymbol{\sigma}^2)$ and denoted by $\widehat{B}_i = \widehat{B}_i(\boldsymbol{\tau^2}, \boldsymbol{\sigma}^2)$ and $\widehat{G}_j = \widehat{G}_j(\boldsymbol{\tau^2}, \boldsymbol{\sigma}^2)$, must satisfy

$$\widehat{B}_i(\boldsymbol{\tau^2}, \boldsymbol{\sigma}^2) = W_i(\bar{y}'_{i.} - \bar{G}_i) + (1 - W_i)b_i \quad \text{and} \quad \widehat{G}_j(\boldsymbol{\tau^2}, \boldsymbol{\sigma}^2) = \bar{y}'_{.j} - \bar{B}_j, \qquad (2.10)$$

where for notational simplicity we write $y'_{ij} = y_{ij} + 0.5\sigma_i^2$, suppressing the dependence on σ_i^2; $\bar{y}'_{i.}$ is the precision (i.e., the reciprocal variance) weighted average of the y'_{ij} over $j \in J_i$, and $\bar{y}'_{.j}$ is the precision weighted average of the y'_{ij} over $i \in I_j$, i.e., $\bar{y}'_{i.} = \frac{\sum_{j \in J_i} y'_{ij} \sigma_i^{-2}}{\sum_{j \in J_i} \sigma_i^{-2}}$ and $\bar{y}'_{.j} = \frac{\sum_{i \in I_j} y'_{ij} \sigma_i^{-2}}{\sum_{i \in I_j} \sigma_i^{-2}}$.

In (2.10), $\bar{G}_i$ is the precision weighted average of the $\widehat{G}_j(\boldsymbol{\tau^2}, \boldsymbol{\sigma^2})$ over $j \in J_i$ and $\bar{B}_j$ is the precision weighted average of the $\widehat{B}_i(\boldsymbol{\tau^2}, \boldsymbol{\sigma^2})$ over $j \in J_i$, i.e.,

$$\bar{G}_i = \frac{\sum_{j \in J_i} \widehat{G}_j(\boldsymbol{\tau^2}, \boldsymbol{\sigma^2}) \sigma_i^{-2}}{\sum_{j \in J_i} \sigma_i^{-2}} \quad \text{and} \quad \bar{B}_j = \frac{\sum_{i \in I_j} \widehat{B}_i(\boldsymbol{\tau^2}, \boldsymbol{\sigma^2}) \sigma_i^{-2}}{\sum_{i \in I_j} \sigma_i^{-2}}.$$

Note that the expressions for $\widehat{B}_i(\boldsymbol{\tau^2}, \boldsymbol{\sigma^2})$ and $\widehat{G}_j(\boldsymbol{\tau^2}, \boldsymbol{\sigma^2})$ involve $\bar{G}_i$ and $\bar{B}_j$, which are linear combinations of $\widehat{B}_i(\boldsymbol{\tau^2}, \boldsymbol{\sigma^2})$ and $\widehat{G}_j(\boldsymbol{\tau^2}, \boldsymbol{\sigma^2})$. Therefore, when the variance parameters $\boldsymbol{\sigma^2}$ and $\boldsymbol{\tau^2}$ are known, (2.10) and the expressions for $\bar{G}_i$ and $\bar{B}_j$ form a system of linear equations, the solutions of which are the MAP estimators for the B_i and G_j. Finally, the weights

$$W_i = \frac{|J_i| \sigma_i^{-2}}{\tau_i^{-2} + |J_i| \sigma_i^{-2}} \tag{2.11}$$

serve as the shrinkage factor for estimating B_i. The form of (2.11) is intuitive because it measures the relative precision provided by the likelihood with respect to the total posterior precision. Hence $1 - W_i$ is the proportion of information from the prior distribution. This metric permits us to make judicious choices of the prior variances, $\boldsymbol{\tau^2}$, when they are not given by experts, so that our results are not unduly prior-driven. See Sections 4.1, 4.2 and 4.3 for in-context discussions.

The linear shrinkage corresponds to "power shrinkage" on the original scale. Consider the case where $\boldsymbol{\tau^2}$ and all $G_j = g_j$ are known. In this case, (2.10) yields $\widehat{B}_i(\boldsymbol{\tau^2}, \boldsymbol{\sigma^2}) = W_i(\bar{y}'_{i.} - \bar{g}_i) + (1 - W_i)b_i$. Consequently, the Effective Area is estimated by

$$\widehat{A}_i = \widehat{A}_i(\boldsymbol{\tau^2}, \boldsymbol{\sigma^2}) = \exp[\widehat{B}_i(\boldsymbol{\tau^2}, \boldsymbol{\sigma^2})] = a_i^{1-W_i} \left[(\tilde{c}_{i.} \tilde{f}_i^{-1}) e^{\sigma_i^2/2}\right]^{W_i},$$

where $\tilde{c}_{i.}$ and $\tilde{f}_i$ are the geometric means: $\tilde{c}_{i.} = \left[\prod_{j \in J_i} c_{ij}\right]^{1/|J_i|}$ and $\tilde{f}_i = \left[\prod_{j \in J_i} f_j\right]^{1/|J_i|}$. This adjustment depends on the relative precision $1 - W_i$ for the b_i. If $W_i = 1$, that is, if b_i is not trustworthy at all, we ignore a_i and estimate A_i by $\widehat{A}_i = \left[\tilde{c}_{i.} \tilde{f}_i^{-1}\right] e^{\sigma_i^2/2}$. Note that the bias correction $e^{\sigma_i^2/2}$ is needed because otherwise $\tilde{c}_{i.} \tilde{f}_i^{-1}$ converges to $A_i e^{-\sigma_i^2/2}$ as $|J_i| \to \infty$. In contrast, if $W_i = 0$, i.e., b_i possesses no error, then we ignore any data and just use $\widehat{A}_i = a_i$ to estimate A_i.

Because W_i grows with $|J_i|\sigma_i^{-2}$, for fixed σ_i^2, the more calibration data we have, the larger the adjustment we make. However, the precision is not determined by the *data size* $|J_i|$ alone, but also by the *quality* of the data, as reflected in σ_i^2. Hence if both $|J_i|$ and σ_i^2 are large, W_i may not be near 1 because the indirect information $|J_i|\sigma_i^{-2}$ may not be large compared to τ_i^{-2}.

When $\boldsymbol{\sigma}^2$ is unknown, we use the conjugate prior distributions for $\boldsymbol{\sigma}^2$ as in Section 2.2. Taking the derivative of the log of (2.5) with respect to σ_i^2 reveals that the MAP estimators also satisfy

$$\widehat{\sigma}_i^2 = 2\left[\sqrt{1 + S_{y,i}^2} - 1\right], \quad S_{y,i}^2 = \frac{1}{|J_i| + \alpha}\left[\sum_{j \in J_i}(y_{ij} - \widehat{B}_i - \widehat{G}_j)^2 + \beta\right], \tag{2.12}$$

where β is the shape parameter for the inverse Gamma prior distribution for σ_i^2 as given in (2.4). We then solve (2.10) and (2.12) to obtain the MAP estimators $\{\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{G}}, \widehat{\boldsymbol{\sigma}^2}\}$. For finite $|J_i|$, because $S_{y,i}^2 \geq \beta/(|J_i|+\alpha) \geq \beta/(M+\alpha)$, all $\widehat{\sigma}_i^2$ are bounded below by $2\sqrt{1 + \beta/(M+\alpha)} - 2 > 0$. Hence our model, including its prior specifications, avoids the problem of an unbounded likelihood at $\sigma_i^2 = 0$, which is a known problem of hierarchical modeling with weak likelihood or prior information.

Intriguingly, the MAP estimator for the variance is also a shrinkage estimator,

$$\widehat{\sigma}_i^2 = 2\left[\sqrt{1 + S_{y,i}^2} - 1\right] = \frac{2}{1 + \sqrt{1 + S_{y,i}^2}}\, S_{y,i}^2 \equiv R_i S_{y,i}^2,$$

where $S_{y,i}^2$ of (2.12) is a natural extension of residual variance estimator, incorporating prior information through $\{\alpha, \beta\}$. The half-variance correction leads to a shrinkage of $S_{y,i}^2$ because $R_i \leq 1$. The degree of shrinkage depends on $S_{y,i}^2$ itself: the larger $S_{y,i}^2$ is, the more shrinkage. Such a self-weighted non-linear shrinkage phenomenon appears to be new.

2.4   Extensions to Handling Outliers: Log-t Model

Outliers are not uncommon in astronomical observations because the harsh environments in which the detectors operate can be subject to large variations in background intensities, potentially leading to large errors in flux estimates. In addition, astronomical sources have intrinsic variabilities covering many orders of magnitude, and some measurements could be performed in regimes where the detectors do not respond linearly to the incoming signal. For these reasons, we propose a robust log-t model as a generalization of the log-Normal model to better handle outliers (see,

e.g. Lange et al. (1989)). Specifically, we introduce a latent variable ξ_{ij} for each observation y_{ij} that is used to down-weight outliers. Formally, for each observation y_{ij}, we assume

$$y_{ij} \mid \boldsymbol{B}, \boldsymbol{G}, \boldsymbol{\xi} \quad = \quad -\frac{\kappa^2}{2\xi_{ij}} + B_i + G_j + \frac{Z_{ij}}{\sqrt{\xi_{ij}}}, \tag{2.13}$$

$$Z_{ij}|\boldsymbol{\xi} \overset{\text{indep}}{\sim} \mathcal{N}(0, \kappa^2), \quad \text{and} \quad B_i \overset{\text{indep}}{\sim} \mathcal{N}(b_i, \tau_i^2),$$

where $\boldsymbol{\xi} = \{\xi_{ij}\}$. Because $\mathrm{E}(e^{y_{ij}} \mid \boldsymbol{B}, \boldsymbol{G}) = \mathrm{E}\left[\mathrm{E}(e^{y_{ij}} \mid \boldsymbol{B}, \boldsymbol{G}, \boldsymbol{\xi})|\boldsymbol{B}, \boldsymbol{G}\right] = A_i F_j$, the multiplicative model in (1.1) is maintained. Depending on the assumptions made for ξ_{ij}, (2.13) includes:

Case 1: *log-Normal model with known variances.* If the ξ_{ij} are known constants, the noise terms $e_{ij} = Z_{ij}/\sqrt{\xi_{ij}}$ are independent Normals with mean 0 and variance $\sigma_{ij}^2 = \kappa^2/\xi_{ij}$. Thus the model in (2.13) is equivalent to (2.4) with known variances.

Case 2: *log-Normal model with unknown variances.* If $\xi_{ij} = \xi_i \overset{\text{indep}}{\sim} \chi_u^2$ for all j, then the variances of $Z_{ij}/\sqrt{\xi_i}$ conditional on ξ_i are κ^2/ξ_i, which are distributed as independent scaled inverse χ^2 with degree of freedom u and scale $u\kappa^2$. Thus (2.13) is equivalent to (2.4) with $\alpha = u/2$ and $\beta = \kappa^2/2$, noting the equivalence between χ^2 and Gamma distributions.

Case 3: *log-t model.* If $\xi_{ij} \overset{\text{indep}}{\sim} \chi_\nu^2$, i.e., mutually independent χ^2 random variables, which are also independent of the Z_{ij}, then the error terms $Z_{ij}/\sqrt{\xi_{ij}}$ follows independent (scaled) student-t distributions: $Z_{ij}/\sqrt{\xi_{ij}} \overset{\text{indep}}{\sim} (\kappa/\sqrt{\nu}) \ t_\nu$.

All of these models can be fit using HMC via `STAN` in the package *CalConcordance*.

Besides down-weighting outliers, the latent $\boldsymbol{\xi}$ also permits a unique variance κ^2/ξ_{ij} for each instrument-source combination. The log-t model is thus more flexible, but computationally more demanding, than the log-Normal model: convergence of HMC is harder to achieve and the sampling is more costly. When the ξ_{ij} are small, the half variance corrections $\kappa^2/2\xi_{ij}$ are also likely to dominate the error terms $Z_{ij}/\sqrt{\xi_{ij}}$, because of their (much) smaller denominator. This results in small y_{ij}, and hence the model tends to generate heavier left tails than right tails.

Despite these challenges, which are topics for further study, we demonstrate the effectiveness of log-t compared with log-Normal model for simulated and real data in the presence of serious outliers. Among heavy-tailed distributions, we choose log-t because it is a natural extension of

the log-Normal model and it carries the intuitive interpretation of "down-weighting" outliers: ξ_{ij} serves as the "weight" for y_{ij}. The last point is confirmed in Simulation III in Section 3.2: outliers have much smaller estimated ξ_{ij} relative to other observations. Without outliers, however, we recommend the log-Normal model for its adequacy and computational simplicity.

3 Testing the Concordance Models with Simulated Data

Our simulation studies aim to demonstrate that (1) the log-Normal model is reasonably robust to the type of model specifications likely to occur in practice; (2) a commonly adopted plug-in method treating guesstimated variances as known can lead to very poor adjustments; and (3) the log-t model is preferred in the presence of serious outliers. We choose the simulation sample sizes to be on the same order as those in our applied examples to make the results more interpretable.

3.1 Checking Robustness to Likely Misspecification

As seen in Section 2, we approximate the Poisson counts via a log-Normal distribution and treat the T_{ij} as known quantities in (1.1). These assumptions have reasonable justifications (Marshall et al., 2018), but nevertheless we should exercise due diligence. Here we study the adequacy of the log-Normal approximation. Further simulations (IV-VI) are presented in Appendix F to investigate the effect of treating T_{ij} fixed and how it interacts with the log-Normal approximation.

In Simulations I and II, there are $N = 10$ instruments and $M = 40$ sources. The data are generated as $y_{ij} = \log \tilde{c}_{ij}$, where $\{\tilde{c}_{ij}\}$ are independent zero-modified Poisson counts with $\lambda_{ij} = A_i F_j = \exp(B_i + G_j)$. We set $\{B_i = 1, G_j = 1\}$ in Simulation I and $\{B_i = 5, G_j = 3\}$ in Simulation II. We independently sample $b_i = \log a_i$ from $\mathcal{N}(B_i, 0.05^2)$. Thus Simulation I represents a low count scenario where the log-Normal approximation may not be appropriate. When $\boldsymbol{\sigma}^2 = 0.1^2 \mathbf{1}_N$, where $\mathbf{1}_N$ denotes an $N \times 1$ column vector of 1s, we use the posterior for B_i and G_j given in Section 2.2. Otherwise, we specify the prior of each σ_i^2 as independent inverse Gammas with degree of freedom $\alpha = 2$ and scale $\beta = 0.01$, and use HMC to obtain draws from the joint posterior distribution. Our choice of hyperparameters are set to match astronomers' prior knowledge; for example, 0.1^2 is the maximum of their guesstimates of $\boldsymbol{\sigma}^2$, reflecting the
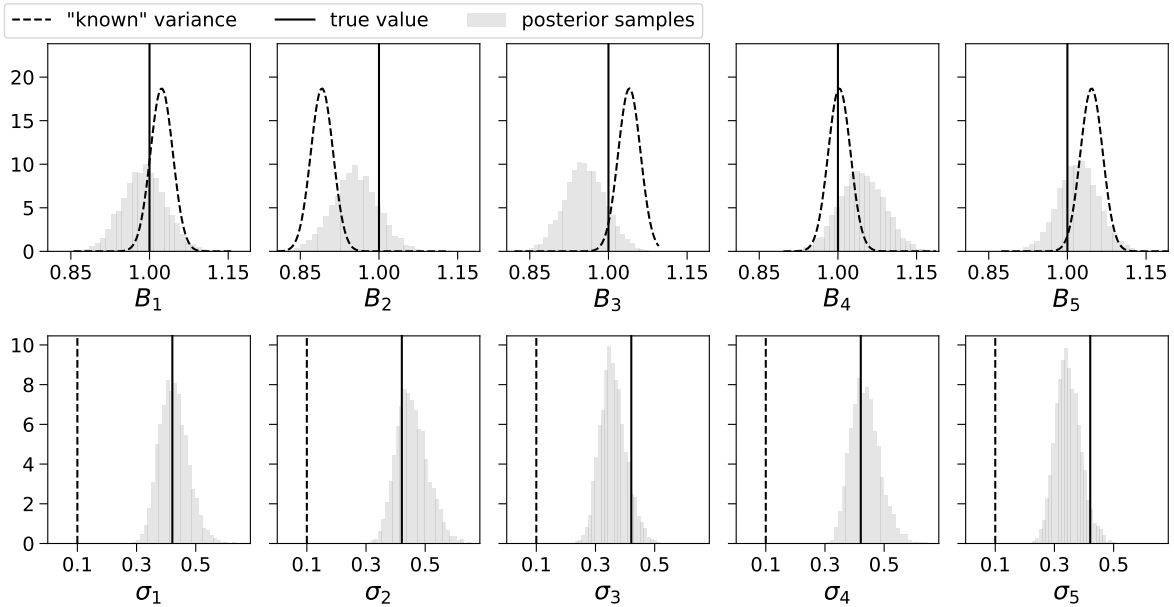
Figure 2: Simulation I. Posterior histograms of $\{B_i, \sigma_i\}_{i=1}^5$. The solid vertical black lines denote the true/theoretical values of $B_i = 1$ (top row) and of $\sigma_i = 0.421$ (second row). The dashed vertical lines denote $\sigma_i = 0.1$ (second row). The black dashed density curves denote the exact posterior densities of B_i when we set the variances equal to their guesstimated value $\sigma_i^2 = 0.1^2$.

general consensus that 10% or less relative error (recall y_{ij} is on log scale) does not alter physical interpretations in important ways. As expected, the fitted values of B_i and G_j are much closer to their targets in Simulation II since it has more Poisson counts resulting from larger values of B_i and G_j; Figures 2 and 3 give detailed results (for B_i, σ_i) under Simulations I and II.

Given B_i and G_j, since $\tilde{c}_{ij}$ is discrete, the theoretical variance $\mathrm{Var}[\log(\tilde{c}_{ij})]$, which is a function of $\lambda = \exp(B_i + G_j)$, can be calculated numerically to any desired accuracy. For reasonably large $B_i + G_j$, we can also approximate σ_{ij}^2 by the δ-method: $\sigma_{ij}^2 \approx \mathrm{Var}(\tilde{c}_{ij})/\mathrm{E}^2(\tilde{c}_{ij})$, where $\mathrm{E}(\tilde{c}_{ij})$ and $\mathrm{Var}(\tilde{c}_{ij})$ are obtained from (2.3) with $\lambda = \exp(B_i + G_j)$. For Simulations I, $\lambda = e^2 = 7.4$, which leads to $\sigma_{ij} = 0.421$ numerically; in contrast, the δ-method gives $\sigma_{ij} \approx 0.367$, a poor approximation due to the smallness of λ. This is a warning as to the inadequacy of using the log-Normal approximation. In contrast, for Simulation II, $\lambda = e^8 = 2981$, and hence $\sigma_{ij} = 0.018$; the δ-method gives the same figure (to four significant digits). Note that the priors for $\boldsymbol{\sigma}^2$ are inverse Gammas with degrees of freedom 2 and scale β for both simulations; we use $\beta = 0.01$ for
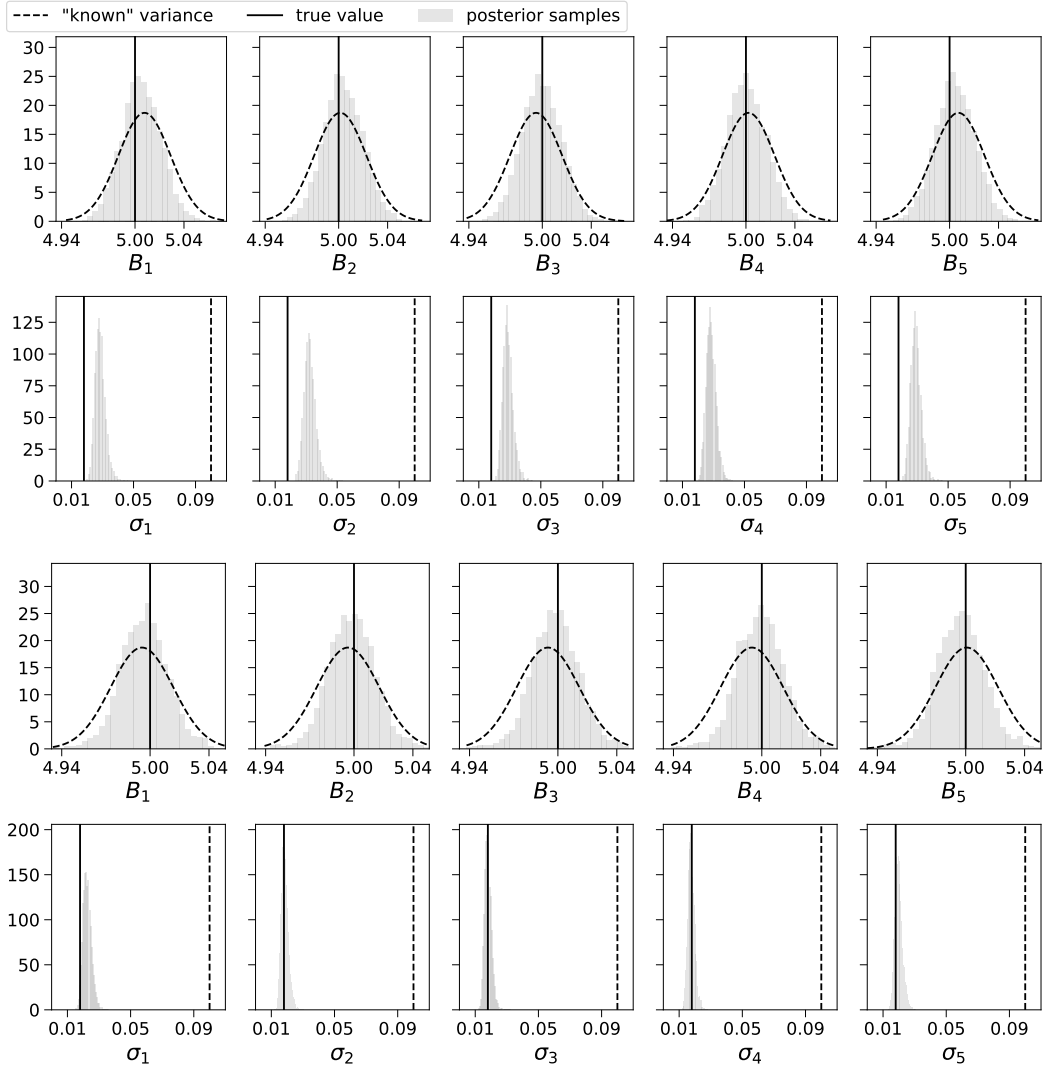
16

Figure 3: Simulation II. Posterior histograms of $\{B_i, \sigma_i\}_{i=1}^{5}$ corresponding to $\beta = 0.01$ (rows 1 & 2) and $\beta = 0.001$ (rows 3 & 4). The solid vertical black lines denote the true values of $B_i = 5$ (rows 1 & 3) and of $\sigma_i = 0.018$ (rows 2 & 4). The dashed vertical lines denote the guesstimated value $\sigma_i = 0.1$ (rows 2 & 4). The black dashed density curves in rows 1 & 3 denote the exact posterior densities of B_i when we set the variances equal to their guesstimated value $\sigma_i^2 = 0.1^2$.

Simulation I, and $\beta = 0.01$ as well as $\beta = 0.001$ for Simulation II.

Now suppose we set each $\sigma_i^2 = 0.1^2$ as guesstimates of the variances. Comparing the histograms and the overlaying curves from Figures 2 and 3, we see in Simulation I that the posterior distributions largely miss their targets, because $\sigma_i^2 = 0.1^2$ is significantly smaller than the vari-

ances estimated under the unknown variance model. In contrast, when the true value of $\boldsymbol{\sigma}^2$ is larger than its guesstimate, as in Simulation II, the posterior distributions of the B_i do capture the target, but exhibit longer tails compared to those resulting from estimated variances. Both phenomena are expected and confirm that if one must guesstimate the variances, it is better to err on the conservative side. Of course, larger variances imply less precision, which leads to less informative results, an inevitable but small price for overestimating the variances.

3.2 Dealing with Outliers via log-t Model

Simulation III, which is the same as Simulation II except we set $G_1 = -2$ to induce outliers, demonstrates the effectiveness of the log-t model in dealing with outliers. Setting $G_1 = -2$ leads to more outliers in the first source as compared to other sources with $G_j = 3$ because the data generating model is Poisson with count rate $\exp(B_i + G_j)$. Under this model the variance of the logarithm of counts is approximately $e^{-B_i-G_j}$ when $B_i + G_j$ is large. Thus, a very small G_1 yields a much larger variance relative to the other sources (by a magnitude of $e^{3+2} \approx 150$) and more extreme observed values. This is a realistic mechanism for generating outliers because it represents the case where one of the sources is much fainter than the others. Following the notation in Case 3 of Section 2.4, setting the shape parameter for the inverse Gamma prior to $\alpha = 2$ is the same as setting $\nu = 2\alpha = 4$ in a χ^2_ν prior; the scale parameter $\beta = 0.01$ corresponds to $\kappa = \sqrt{2\beta} \approx 0.141$. In this section, we use σ^2_{ij} to denote the residual variance for y_{ij}. From Simulation II, $\sigma_{ij} = 0.018$ for $j > 1$. For $j = 1$, $\lambda = e^{-2+5} = 20.1$, and hence the exact numerical calculation gives $\sigma_{i1} = 0.232$, and the δ-method yields $\sigma_{i1} \approx 0.223$, which is quite a reasonable approximation.

Using the first three sources as an example, the upper panel in Figure 4 compares the results from the log-Normal model and the log-t model through the (fitted) standardized residuals, which are given respectively by (using the notation in Section 2.2 and Section 2.4):

$$\widehat{\mathcal{R}}_{ij} = \frac{y_{ij} - \widehat{B}_i - \widehat{G}_j + 0.5 \times \widehat{\sigma}^2_i}{\widehat{\sigma}_i} \quad \text{and} \quad \widehat{\mathcal{R}}_{ij} = \frac{y_{ij} - \widehat{B}_i - \widehat{G}_j + 0.5 \times \kappa^2/\widehat{\xi}_{ij}}{\kappa/\widehat{\xi}^{1/2}_{ij}}, \tag{3.1}$$

where $\widehat{B}_i, \widehat{G}_j, \widehat{\xi}_{ij}$ and $\widehat{\sigma}_i$ are the posterior means. We see some observations from the first source (black circles) are outliers with standardized residuals outside $[-2, 2]$, in the log-Normal model (upper panel) but not for the log-t model (lower panel). In the log-Normal model, setting $\sigma^2_{ij} = \sigma^2_i$
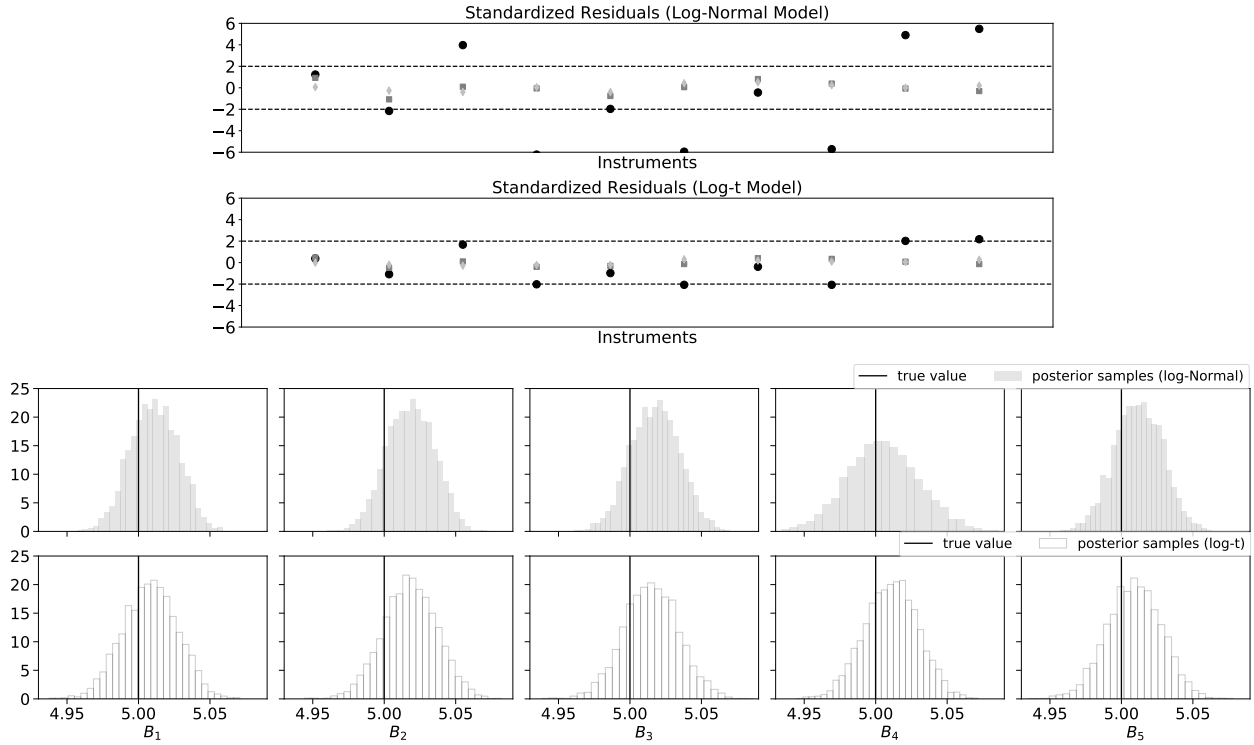
Figure 4: Simulation III. Rows 1 and 2 show standardized residuals of the log-Normal model and the log-t model. The black circles, gray squares and silver rhombi represent the first three sources respectively. The instruments are plotted on the x-axes. The dashed horizontal lines denote the $[-2, 2]$ intervals. Rows 3 and 4 show posterior histograms of $\{B_i\}_{i=1}^5$ from the log-Normal model and the log-t model, where the black vertical bars indicate true values.

causes large standardized residuals due to some source-dependent large variances: $\sigma_{i1}^2 >> \sigma_{ij}^2$, $j \geq 2$. In the log-t model, the outliers are down weighted by ξ_{ij} and each observation is assigned a unique conditional variance $\sigma_{ij}^2 = \kappa^2/\xi_{ij}$, illustrating the benefit of using the log-t model.

The lower panel in Figure 4 shows the posterior distributions of B_i under the log-Normal and log-t models, both capturing the true value. The log-t model exhibits slightly larger variances between the two. This is comforting, especially considering the flexibility of the log-t model, permitting individual σ_{ij} rather than the hard-constraint $\sigma_{ij} = \sigma_i$ of the log-Normal model.

The first three rows of Table 2 give the (average) coverage of nominal equal-tailed 95% posterior intervals for B_i and G_j obtained from 2000 simulations with the same configurations as in

19

Data Generating Model (Poisson)	Parameter	Coverage Probability		Length of Interval	
		log-Normal	log-t	log-Normal	log-t
$N = 10, M = 40$	$\boldsymbol{B}$	[0.941, 0.959]	[0.971, 0.975]	0.067±0.005	0.073 ± 0.002
$N = 10, M = 40$	G_1	*0.399*	*0.700*	*0.090± 0.015*	*0.182±0.045*
$N = 10, M = 40$	$G_2, \ldots, G_M$	[0.967, 0.977]	[0.996, 0.999]	0.077±0.003	0.104±0.002
$N = 40, M = 40$	$\boldsymbol{B}$	[0.953, 0.969]	[0.993, 0.998]	0.041±0.007	0.050±0.001
$N = 40, M = 40$	G_1	*0.398*	*0.686*	*0.045±0.003*	*0.093±0.013*
$N = 40, M = 40$	$G_2, \ldots, G_M$	[0.965,0.977]	[0.996,0.999]	0.038±0.001	0.051±0.001

Table 2: Coverage of nominal 95% posterior intervals calculated from 2000 datasets simulated under a Poisson model using the same configurations as in Simulation III. The intervals in columns 3 and 4 give the smallest and largest coverage observed for the corresponding parameter. The last two columns give the lengths of nominal 95% intervals in the format: mean ± standard deviation.

Simulation III. The log-t model is more robust to outliers than the log-Normal model, exhibiting significantly better coverage for G_1. (Though coverage under the log-t model is still poor relative to the nominal level.) Table 2 also indicates that outliers are not as problematic for estimating B_i, our primary interest, as they are for G_1. This is because we are more "informed" about the B_i than the G_j since (1) in this experiment $N = 10, M = 40$, and hence there are more sources than instruments, and (2) each B_i has an informative prior whereas each G_j only has a flat prior.

As illustrated in Table 2, when N is increased from 10 to 40, the coverage of $\boldsymbol{G}$ changes little. The coverage of $\boldsymbol{B}$ on the other hand increases noticeably even with narrower intervals, especially under the log-t model. The narrowing of the intervals for $\boldsymbol{G}$ is expected with more instruments per source. The simultaneous increase of coverage and decrease of interval widths for $\boldsymbol{B}$ is intriguing. It is a welcome finding from the astrophysics application perspective. But it also indicates potential defects in the log-Normal or the log-t approximation because over-coverage suggests a non-optimal posterior uncertainty calibration. The half-variance correction likely plays a role here because it permits uncertainty in variance estimation to directly affect inference for the mean. Overall, we recommend the log-t model when one suspects serious outliers. This may lead to unnecessarily larger error bars for flux estimates that are not (directly) affected by the outliers,

a worthwhile premium against disastrous loss of coverages for estimands that are affected.

4   Applying the Proposed Methods to IACHEC Data

In this section, we fit the log-Normal model to three datasets (given in Appendix A) compiled by researchers from IACHEC (2017), with the aim of increasing understanding of the calibration properties of various X-ray telescopes (a.k.a. instruments) such as *Chandra*, *XMM-Newton*, *Suzaku*, *Swift*, etc. See Marshall et al. (2018) for details on data collection and preprocessing.

4.1   E0102 Data

SNR 1E 0102.2-7219 (abbreviated as E0102) is the remnant of a supernova that exploded in a neighboring galaxy known as the Small Magellanic Cloud (Chandra X ray observatory, 2009) and is a calibration target for a variety of X-ray missions. We consider four photon sources associated with E0102. Each is a local peak or "line" in the E0102 spectrum, which can be thought of as a high-resolution histogram of the energies of photons originating from E0102. Our "sources" corresponds to the photon counts in four bins of this histogram. Two of the lines are associated with highly ionized Oxygen (Hydrogen Lyman-α like O VIII at 18.969Å and the resonance line of O VII from the He-like triplet at 21.805Å) and the other two are associated with Neon (H-like Ne X at 12.135Å and He-like resonance line Ne IX at 13.447Å). We consider replicate data obtained with 13 different detector configurations respectively over 4 separate telescopes, *Chandra* (HETG and ACIS-S), XMM-*Newton* (RGS, EPIC-MOS, EPIC-pn), *Suzaku* (XIS), and *Swift* (XRT).

Because the energies of the two Oxygen lines are similar, it is reasonable to assume that their associated Effective Areas are also similar; likewise for the Neon lines and their Effective Areas. Thus, we consider two separate datasets, one with O VII and O VIII and the other with Ne IX and Ne X, each with $M = 2$ and $N = 13$. In this way, we have more confidence in the multiplicative model (1.1) than if we were to combine the two into a single dataset with $M = 4$ and $N = 13$. In addition, standard astronomical practice is to work with dimensionless measurements in log space (e.g., optical magnitudes in a given passband are defined as $-2.5 \log_{10} \frac{\text{flux}}{\text{flux from Vega}}$). Since our log transformation of brightness mimics this process, we also normalize the measured line fluxes by
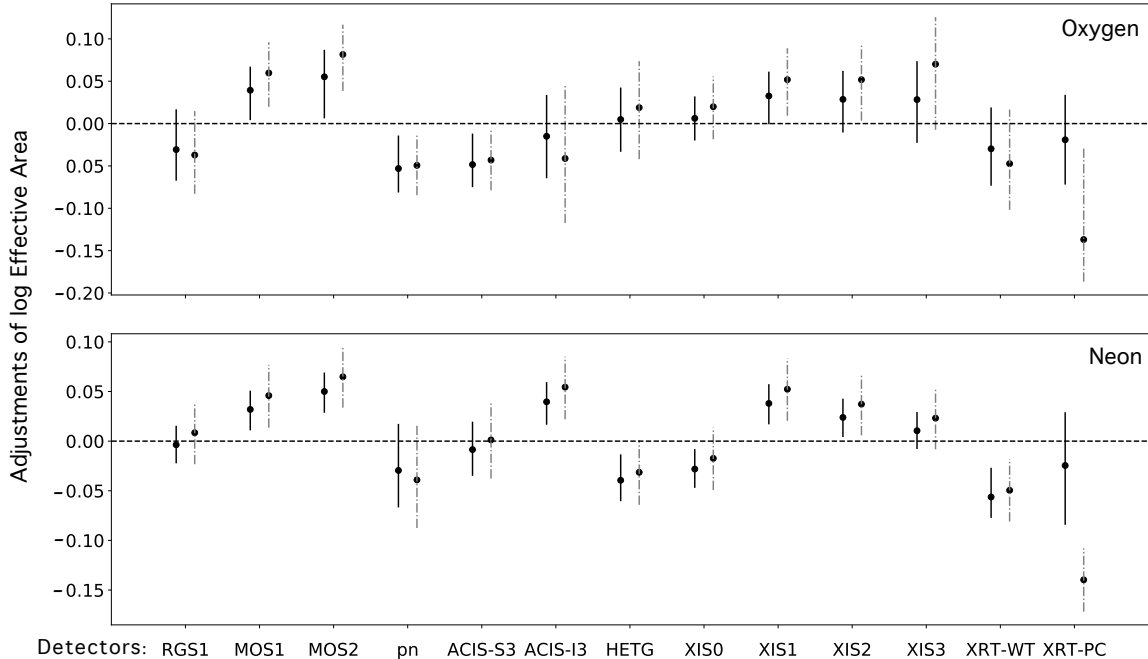
Figure 5: Adjustments of the logarithm of the Effective Areas for Oxygen (row 1) and Neon (row 2) in the E0102 dataset. The x-axis labels the detectors (instruments) and the y-axis is $\boldsymbol{B}$. The horizontal dashed lines represent zero, which indicates no adjustments for the Effective Areas. The vertical bars denote 95% posterior interval for each B_i, and the dots denote the posterior means. The black and gray bars correspond to $\tau_i = 0.025$ and 0.05, respectively.

those from an arbitrarily chosen detector, as done in Plucinsky et al. (2017).

To apply the log-Normal model to the two datasets, we choose priors with hyperparameters $\alpha = 1.5$, $\beta = 2 \times 10^{-4}$ for O VII, O VIII and $\beta = 8 \times 10^{-5}$ for Ne IX, Ne X. We set each $b_i = 0$, i.e., a priori we expect no adjustment is needed, with confidence τ_i, taking two possible values $\tau_i = 0.025$ and $\tau_i = 0.05$. These (and subsequent) choices are based on astronomers' knowledge.

Figure 5 shows the adjustments of the log-scale Effective Areas for Oxygen (row 1) and Neon (row 2) data. We see that the estimated values of B_i are not sensitive to the choices of τ_i except for detector XRT-PC. For XRT-PC, with the Neon data, the estimated shrinkage factor towards the prior, $1 - W_i$, as given in (2.11), is 0.91 with $\tau_i = 0.025$ and 0.02 with $\tau_i = 0.05$. This indicates that, if the prior variance of B_i is too small ($\tau_i = 0.025$ here), the model treats the observations

as being less accurate (by fitting a large σ_i) instead of further adjusting the Effective Area of the corresponding instrument (a larger deviation from b_i). Numerical results presented in Table 5 in Appendix A reveal that the estimated shrinkage factors can vary slightly or drastically with τ_i. Since $M = 2$, sensitivity to the choice of hyperparameters is expected. A feature of the log-Normal model is the direct link between its mean and variance stemming from the half-variance correction. This link indicates additional sensitivities that are neither commonly observed nor well studied.

From Figure 5 and Table 5 (Appendix A), XRT-PC has a much lower Effective Area than the other instruments: about -0.15 versus between $[-0.05, 0.05]$ on the log scale. The corresponding estimated shrinkage factor is more sensitive to the choice of τ_i, for both the Oxygen and Neon data: when τ_i is small ($= 0.025$), the posterior mean of B_i is constrained too much by its zero-centered prior. Thus we need a larger σ_i to compensate for the the influence of the prior. In contrast, when τ_i is larger (we also tried $0.05, 0.075, 0.1$), the estimated shrinkage factor is not as sensitive to τ_i. Overall, Figure 5 suggests that the Effective Areas of MOS1, MOS2, XIS1, XIS2, XIS3 need to be adjusted upward and those of pn, XRT-WT, XRT-PC need to be adjusted downward.

4.2   2XMM Data

The 2XMM catalog (Watson et al., 2009) can be used to generate large, well-defined samples of various types of astrophysical objects, notably active galaxies (AGN), clusters of galaxies, interacting compact binaries, and active stellar coronae, using the power of X-ray selection (XMM Catalogue public pages, 2008). The 2XMM catalog data are collected with the XMM-Newton European Photon Imaging Cameras (EPIC). Briefly, there are three EPIC instruments: the EPIC-pn (hereafter referenced as "pn") and the two EPIC-MOS detectors (hereafter referenced as "MOS1" and "MOS2"). These detectors have separate X-ray focusing optics but are co-aligned so that the sources in our samples are observed simultaneously in the pn, MOS1, and MOS2 detectors.

Our 2XMM data contain three datasets, corresponding to the hard (2.5 - 10.0 keV), medium (1.5 - 2.5 keV) and soft (0.5 - 1.5 keV) energy bands. The three instruments (pn, MOS1 and MOS2) measured 41, 41, and 42 sources respectively in hard, medium, and soft bands. The sources are from the 2XMM EPIC Serendipitous Source Catalog (Watson et al., 2009), selected to be sufficiently faint that the thorny issue of "pileup", which occurs when several photons hit
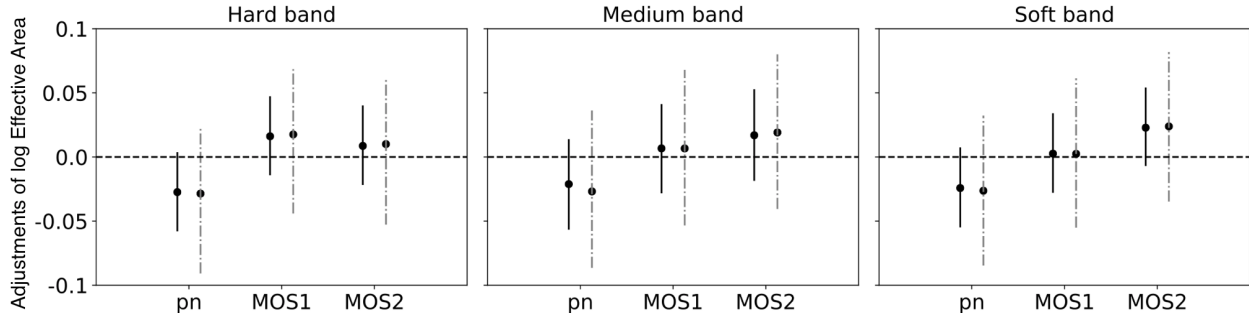
Figure 6: Adjustments of the log-scale Effective Areas for hard band (left), medium band (middle) and soft band (right) of the 2XMM datasets. The legend is the same as in Figure 5.

the detector at the same time, can be ignored. With sufficient exposure, on average 1,500 counts are collected from the faint sources in each band of the detector (Marshall et al., 2018).

The log-Normal model is fit to the three datasets separately, with $\beta = 0.014, 0.083$ and 0.022 respectively for the hard, median, and soft bands, but with $\alpha = 1.5$ for all three. We again use $b_i = 0$ and try $\tau_i = 0.025$ and $\tau_i = 0.05$. Figure 6 shows the resulting adjustments of the log-scale Effective Area, and confirms the astronomers' expectation that no adjustment is needed for 2XMM, regardless of the choice of the τ_i. In contrast to Table 5, Table 6 (also in Appendix A) shows a much more stable patterns of proportion of prior information for 2XMM data.

4.3   XCAL Data

XCAL consists of bright AGN from the XMM-Newton cross-calibration sample[2]. The image data are clipped, using a standard XMM software task (called `epatplot`), to eliminate the regions affected non-trivially by pileup. The amount of clipping depends on the observed source intensity: unused regions are larger for brighter sources (Marshall et al., 2018). The initial estimate of the Effective Area is then adjusted according to lookup tables (from other in-flight data) to account for the unused regions. Like the 2XMM data, XCAL data are composed of three datasets: the hard (94 sources), medium (103), and soft (108) bands, all measured by three instruments, pn, MOS1 and MOS2. We use the same procedure and hyperparameters as in Section 4.2, except we set $\beta = 8.0 \times 10^{-4}, 8.6 \times 10^{-3}$ and 6.8×10^{-4} respectively for hard, median, and soft bands.

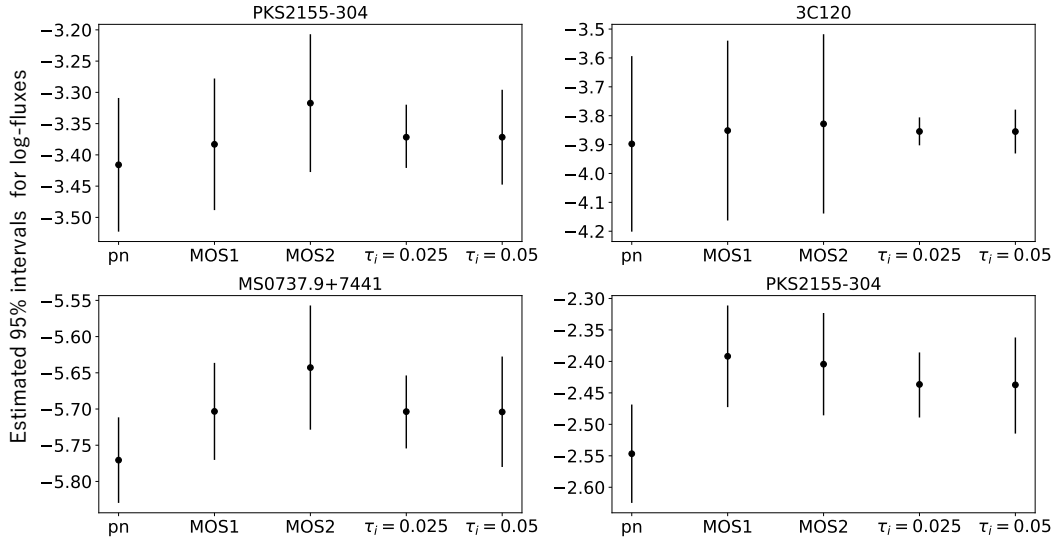[2]See Section 4 in `http://xmm2.esac.esa.int/docs/documents/CAL-TN-0052.ps.gz`

Figure 7: Comparison of estimated 95% intervals for log-fluxes using a standard astronomical method (left three bars) and those from the log-Normal model (right two bars) for four representative medium-band sources in XCAL data, as indicated by the panel titles.

Figure 7 demonstrates that adjustment of the Effective Areas is needed to align the measured fluxes across the detectors. Results are presented for four sources from the medium band data, where the left three bars—corresponding to three instruments—depict the 95% intervals (mean $\pm$ 2 given standard deviations) for the log-fluxes obtained by a standard astronomical method. The right two bars—corresponding to two choices of the prior variance τ_i—represent the 95% posterior intervals of log-fluxes after adjustment using our log-Normal model. This visualization illustrates the reliability of our calibration of Effective Areas, as it helps to bring together the varied flux estimates from individual detectors in a statistically principled way. In particular, we see that the posterior mean of the log-flux is rather robust to the choice of τ_i, yet the corresponding posterior variance respects astronomers' a priori knowledge as coded into τ_i.

Finally, we show how to adjust the Effective Areas of each instrument to obtain the results illustrated in the rightmost interval in each panel of Figure 7. Figure 8 shows the necessary adjustment of $\boldsymbol{B}$ for hard band (left), medium band (middle) and soft band (right). For all these bands, we adjust the Effective Area of pn downward and that of MOS2 upward.
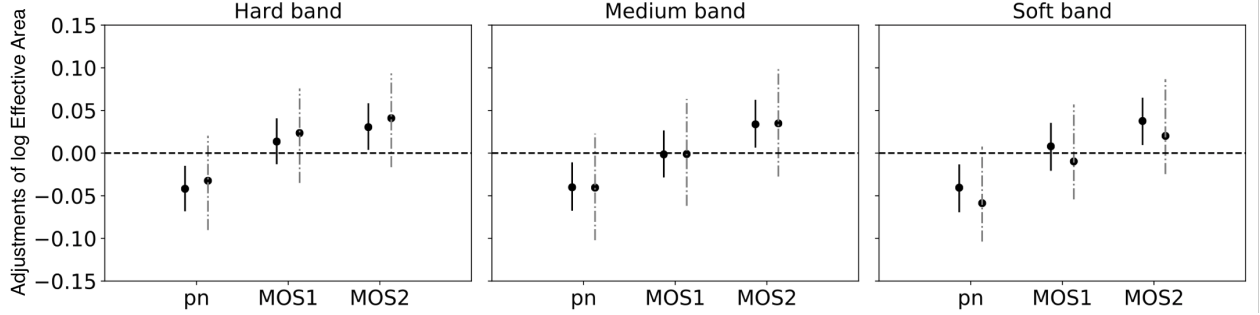
25

Figure 8: Adjustments of the log-scale Effective Areas for hard band (left), medium band (middle) and soft band (right) based on XCAL data. The legend is the same as in Figure 5.

4.4 Model Checking

To check how well the log-Normal model captures the observed variability in the data, we use residual plots to visualize the goodness-of-fit. Figure 9 plots the standardized residuals $\widehat{\mathcal{R}}_{ij}$ for the data analyzed in Section 4.3 with $\tau_i = 0.05$, with the left panels denoting residuals from the log-Normal model and the right panels from the log-t model (see the two expressions of $\widehat{\mathcal{R}}_{ij}$ in (3.1)). Nearly all residuals fall in $[-3, 3]$ under the log-Normal model and $[-2, 2]$ under the log-t model. The observations of 3C111 in all three energy bands are the only outliers under the log-Normal model but are not outliers under the log-t model, confirming the latter's ability to handle outliers. The adjusted Effective Areas and the estimated fluxes are not too sensitive to whether or not the outliers are excluded. Thus the log-Normal model is acceptable for the data in Section 4.3.

We also employ a posterior predictive check (Meng, 1994; Gelman et al., 1996) for the log-Normal model. In a posterior predictive check, one chooses test statistics and computes the posterior predictive p-value. The test statistics we choose are

$$\left\{ T_i = \overline{y}_{i\cdot} - \overline{y} = \frac{\sum_{j=1}^{M} y_{ij}}{M} - \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}}{NM} \right\}_{i=1}^{N},$$

which reflect the relative magnitudes of the log scale Effective Areas. None of the posterior predictive p-values for any of our datasets are significant, i.e., we never fail the posterior predictive check. This does not, however, imply that no serious model defects exist. Below we discuss directions for improving our models and ultimately the reliability of the proposed concordance adjustments.
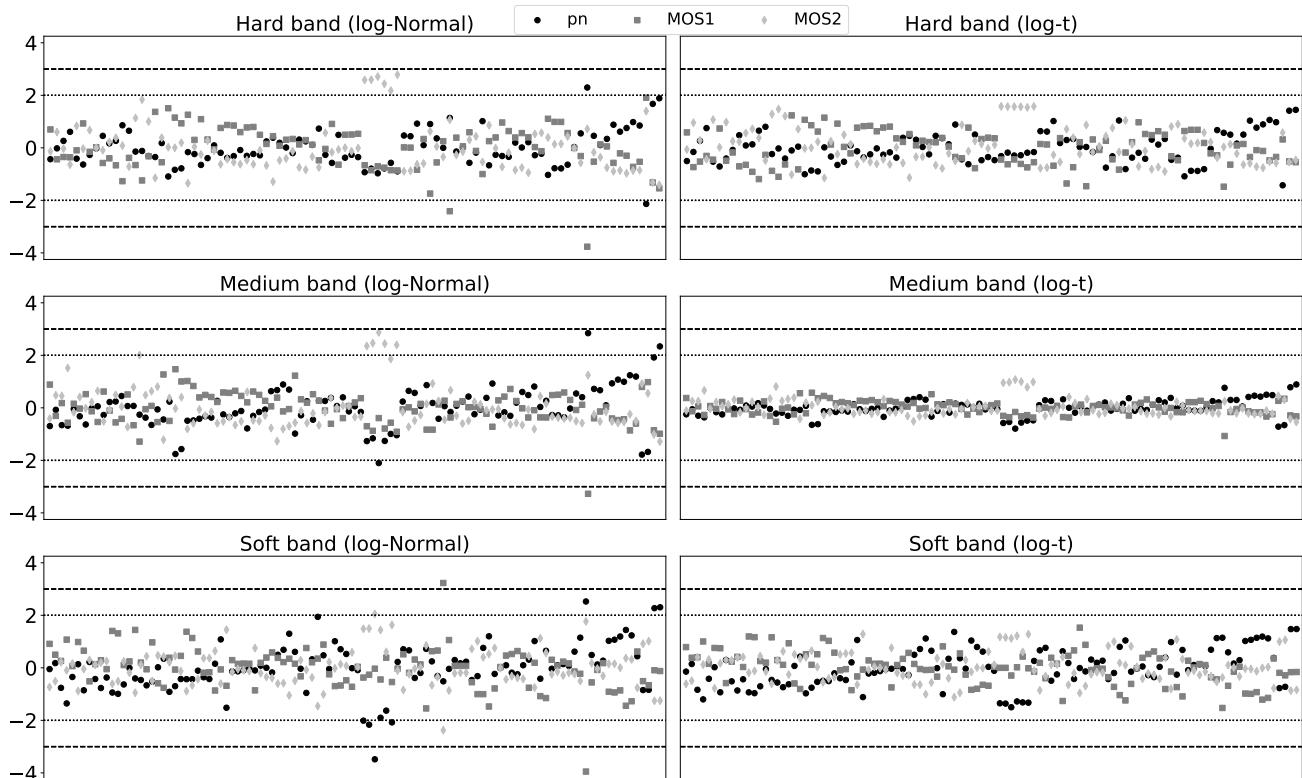
Figure 9: Standardized residuals for the XCAL data in Section 4.3 with $\tau_i = 0.05$. Left panel for the log-Normal model and right panel for the log-t model. The black circles, gray squares and silver rhombi denote the instruments pn, MOS1 and MOS2 respectively. The dashed and dotted lines depict respectively the $[-3, 3]$ and $[-2, 2]$ intervals.

5   Alternative Methods and Future Work

5.1   Comparing Likelihood and Bayesian Estimations

In Section 2, we adopt a Bayesian perspective, which leads to the log-Normal model. Alternatively, we could view b_i as a *noisy observation* with known variance τ_i^2: $b_i \overset{\text{indep}}{\sim} \mathcal{N}(B_i, \tau_i^2)$. Together with (2.2), this gives a multivariate Normal regression model and can be fit via maximum likelihood estimation (MLE). In particular, when $\boldsymbol{\sigma^2}$ is known, Proposition 1 in Appendix E.1 gives closed-form expressions when all the instruments measure all sources, which implies the asymptotic properties of the MLEs (Corollary 1). Furthermore, the standardized residual sum of squares follows a χ^2 distribution, which enables testing of the goodness-of-fit; see Appendix E.2 for details.

27

When the variances $\boldsymbol{\sigma^2}$ are unknown, in principle, we can still obtain the (asymptotic) variance of the MLEs by calculating the observed/expected Fisher information. However, the number of unknown parameters we consider, $2N + M$, grows with the number of observations $NM + N$. Conditions on the relationship between N and M that would ensure the classical asymptotic theory for MLEs would be of interest to those who prefer likelihood inference. Furthermore, under such conditions, these estimators should be *approximately* valid even if the Normal assumptions made in (2.2) for the e_{ij} fail. In this case, the variance of the estimator requires a more complicated "sandwich" formula (Freedman, 2006), involving both the Fisher information and the variance of the score function. We say *approximately* valid because the half-variance correction of Section 2 would depend on the log-Normal assumption. Consequently, when the variance is large, the half-variance correction can be misleading if the log-Normal assumption is severely violated.

As usual, the likelihood method is closely related to the Bayesian approach. For example, when $\boldsymbol{\sigma^2}$ is known, the MLEs of $\boldsymbol{B}$ and $\boldsymbol{G}$ correspond to the MAP estimates defined in (2.10), which also have the intuitive interpretation as shrinkage estimators. When the variances are unknown, the likelihood function is unbounded at the boundary of the parameter space ($\sigma_i^2 = 0$). The conjugate priors for variance parameters in the Bayesian model regularize the likelihood and give a proper posterior distribution. This is another reason we adopt the Bayesian approach.

5.2   Future and Related Work

The log-Normal model works reasonably well in our applied illustrations, and it yields important findings that are welcomed by astronomers — concrete guidance about systematic adjustments of the Effective Areas are given and thus concordance of an intrinsic characteristic of each astronomical object across different instruments can be achieved. Calibration scientists are thus able to make *absolute* measurements of characteristics of astronomical objects using different instruments. The posterior distribution of the Effective Area of each instrument can and should be used for downstream analysis of measured fluxes to obtain principled estimates of absolute flux and to properly quantify their uncertainty. Furthermore, we highlight the danger of incorrectly fixing the observation noise through simulation experiments that mimic possible realistic uncertainties.

There are several directions of future work in order to improve the current model. First, so far

we have assumed that the Effective Areas are a priori independent, which is not always true in practice. Sometimes the Effective Areas across different energy bands are noticeably correlated. This correlation structure should be taken into account in future modeling to gain more efficiency in estimation. Second, the log-Normal model gives conservative results under realistic model misspecification, as revealed by our simulation studies. Unfortunately, the scope of simulation studies is always limited. Hence theoretical properties of both the log-Normal and the log-t approaches under model misspecification need to be further investigated. Third, the asymptotic (sampling) properties of the proposed models need to be established, as both the number of parameters and the number of observations approach infinity. Fourth, the robustness of the model with respect to possible sample selection bias and non-ignorable missing data needs to be studied more thoroughly. Although this is not of concern for our current analyses, it could become a more severe problem if we include more instruments and more sources in the calibration data. Last, possible hierarchical extensions of the model that addresses population characteristics, in which we are interested in the flux distribution of a certain type or population of objects instead of each individual object, as in the case of supernovae, could be considered. Of course, in such cases, a representative sample of the population of objects is critical for a meaningful analysis.

Moving forward, to increase the impact of the proposed method, we need to involve more IACHEC members and datasets. Cooperation among IACHEC member projects can lead to enacting adjustments as recommended from the concordance analysis, which will result in closer agreement between different instruments that make similar measurements, to achieve a main goal of IACHEC. Experts from the projects that comprise the IACHEC are needed to examine possible bias in sample selection and to set the values of τ that are needed for the concordance analysis. In our follow-up paper (Marshall et al., 2018), we apply this concordance analysis more broadly and allow the values of τ to be instrument-dependent.

Finally, calibration is a well-known problem in several areas of applications. For example, inter-laboratory calibration (motivated from analytical chemistry) is studied in Gibbons and Bhaumik (2001) and Bhaumik and Gibbons (2005), where they also address simultaneously the issue of multiplicative signals and additive noises, but with a different modeling strategy. More recently, a fiducial approach is used in Hannig et al. (2017) to tackle similar problems. We therefore hope our

modeling strategies add to the toolkits to conduct similar calibration and concordance analysis, such as for environmental monitoring (e.g., Weatherhead et al., 1998). Much more work is needed and can be done, and hence we invite and encourage interested researchers to join us to address these theoretically challenging and practically impactful problems.

Acknowledgement

References

Appenzeller, I. (2012). Introduction to astronomical spectroscopy, volume 9. Cambridge University Press.

Bhaumik, D. K. and Gibbons, R. D. (2005). Confidence regions for random-effects calibration curves with heteroscedastic errors. Technometrics, 47(2):223–231.

Bilder, C. R. and Loughin, T. M. (2014). Analysis of categorical data with R. CRC Press.

Chandra X-ray Center (2017a). The Chandra proposers' observatory guide, chapter 2. Accessed: 2018-02-28.

Chandra X-ray Center (2017b). The Chandra proposers' observatory guide, chapter 3. Accessed: 2018-02-28.

Chandra X ray observatory (2009). E0102-72.3: Adding a new dimension to an old explosion. http://chandra.harvard.edu/photo/2009/e0102/.

Drake, J. J., Ratzlaffa, P., Kashyapa, V., Edgara, R., Izemb, R., Jeriusa, D., Siemiginowskaa, A., and Vikhlinina, A. (2006). Monte Carlo processes for including Chandra instrument response uncertainties in parameter estimation studies. In SPIE Astronomical Telescopes + Instrumentation, pages 62701I–62701I. International Society for Optics and Photonics.

Efron, B. and Morris, C. N. (1975). Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association, 70(350):311–319.

Freedman, D. A. (2006). On the so-called Huber sandwich estimator and robust standard errors. The American Statistician, 60(4):299–302.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6(4):733–760.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6:721–741.

George, I. M., Arnaud, K. A., Pence, B., and Ruamsuwan, L. (1992). Calibration sources for spectral analysis. Legacy, 2.

Gibbons, R. D. and Bhaumik, D. K. (2001). Weighted random-effects regression models with application to interlaboratory calibration. Technometrics, 43(2):192–198.

Graessle, D. E., Evans, I. N., Glotfelty, K., He, X. H., Evans, J. D., Rots, A. H., Fabbiano, G., and Brissenden, R. J. (2006). The Chandra X-ray observatory calibration database (CalDB): building, planning, and improving. Proc. SPIE, 6270:62701X–62701X–12.

Guainazzi, M., David, L., Grant, C. E., Miller, E., Natalucci, L., Nevalainen, J., Petre, R., and Audard, M. (2015). On the in-flight calibration plans of modern X-ray observatories. Journal of Astronomical Telescopes, Instruments, and Systems, 1(4):047001.

Hannig, J., Feng, Q., Iyer, H., Wang, C., and Liu, X. (2017). Fusion learning for inter-laboratory comparisons. Journal of Statistical Planning and Inference.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1):97–109.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research, 15(1):1593–1623.

IACHEC (2017). International Astronomical Consortium for High Energy Calibration. http://web.mit.edu/iachec/.

Kashyap, V., Lee, H., Siemiginowska, A., McDowell, J., Rots, A., Drake, J., Ratzlaff, P., Zezas, A., Izem, R., Connors, A., and van Dyk, D. (2008). How to handle calibration uncertainties in high-energy astrophysics. Proceedings of SPIE (Observatory Operation: Strategies, Processes, and System II), 7016:23 (8pp).

Lange, K. L., Little, R. J., and Taylor, J. M. (1989). Robust statistical modeling using the t distribution. Journal of the American Statistical Association, 84(408):881–896.

Lee, H., Kashyap, V. L., Van Dyk, D. A., Connors, A., Drake, J. J., Izem, R., Meng, X.-L., Min, S., Park, T., Ratzlaff, P., and Siemiginowska, A. (2011). Accounting for calibration uncertainties in X-ray analysis: effective areas in spectral fitting. The Astrophysical Journal, 731(2):126 (19pp).

Madsen, K. K., Beardmore, A. P., Forster, K., Guainazzi, M., Marshall, H. L., Miller, E. D., Page, K. L., and Stuhlinger, M. (2016). IACHEC cross-calibration of Chandra, Nustar, Swift, Suzaku, XMM-Newton with 3C 273 ANDPKS 2155-304. The Astronomical Journal, 153(1):2(17pp).

Marshall, H., Kashyap, V., Chen, Y., Meng, X.-L., and Wang, X. (2018). Concordance: In-flight calibration of X-ray telescopes without absolute references. In Preparation.

Matthews, G. and Havey, Jr., K. (2010). Ten years of Chandra: reflecting back on engineering

lessons learned during the design, fabrication, integration, test, and verification of NASA's great X-ray observatory. Proc. SPIE, 7738:77380Y–77380Y–13.

Meng, X.-L. (1994). Posterior predictive p-values. The Annals of Statistics, 22(3):1142–1160.

Meng, X.-L. (2018). Conducting highly-principled data science: A statistician's job and joy. Statistics and Probability Letters. To appear.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. Journal of Chemical Physics, 21(6):1087–1092.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. Journal of the American Statistical Association, 78(381):47–55.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, Handbook of Markov chain Monte Carlo, volume 2. CRC Press.

Nevalainen, J., David, L., and Guainazzi, M. (2010). Cross-calibrating X-ray detectors with clusters of galaxies: An IACHEC study. Astronomy & Astrophysics, 523:A22.

Plucinsky, P. P., Beardmore, A. P., Foster, A., Haberl, F., Miller, E. D., Pollock, A. M., and Sembay, S. (2017). SNR 1E 0102.2-7219 as an X-ray calibration standard in the 0.5-1.0 keV bandpass and its application to the CCD instruments aboard Chandra, Suzaku, Swift and XMM-Newton. Astronomy & Astrophysics, 597(A35).

Read, A. M., Guainazzi, M., and Sembay, S. (2014). Cross-calibration of the XMM-Newton EPIC pn and MOS on-axis effective areas using 2XMM sources. Astronomy & Astrophysics, 564:A75.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3):581–592.

Schellenberger, G., Reiprich, T., Lovisari, L., Nevalainen, J., and David, L. (2015). XMM-Newton and Chandra cross-calibration using HIFLUGCS galaxy clusters-systematic temperature differences and cosmological impact. Astronomy & Astrophysics, 575:A30.

Schwartz, D. A. (2014). Invited review article: The Chandra X-ray observatory. Review of Scientific Instruments, 85(6):061101.

Sembay, S., Guainazzi, M., Plucinsky, P., and Nevalainen, J. (2010). Defining high-energy calibration standards: IACHEC (International Astronomical Consortium for High-Energy Calibration). In AIP Conf. Proceedings, volume 1248, pages 593–594. AIP.

Seward, F. (1992). Calibration sources for spectral analysis.

Stan Development Team (2015). Stan Modeling Language User's Guide and Reference Manual, Version 2.10.0.

Stan Development Team (2016). Pystan: the Python interface to Stan.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association, 82(398):528–540.

Tsujimoto, M., Guainazzi, M., Plucinsky, P. P., Beardmore, A. P., Ishida, M., Natalucci, L., Posson-Brown, J. L., Read, A. M., Saxton, R. D., and Shaposhnikov, N. V. (2011). Cross-calibration of the X-ray instruments onboard the Chandra, Integral, RXTE, Suzaku, Swift, and XMM-Newton observatories using G21. 5–0.9. Astronomy & Astrophysics, 525:A25.

Watson, M. G., Schröder, A. C., Fyfe, D., Page, C. G., Lamer, G., Mateos, S., Pye, J., Sakano, M., Rosen, S., Ballet, J., Barcons, X., Barret, D., Boller, T., Brunner, H., Brusa, M., Caccianiga, A., Carrera, F. J., Ceballos, M., Della Ceca, R., Denby, M., Denkinson, G., Dupuy, S., Farrell, S., Fraschetti, F., Freyberg, M. J., Guillout, P., Hambaryan, V., Maccacaro, T., Mathiesen, B., McMahon, R., Michel, L., Motch, C., Osborne, J. P., Page, M., Pakull, M. W., Pietsch, W., Saxton, R., Schwope, A., Severgnini, P., Simpson, M., Sironi, G., Stewart, G., Stewart, I. M., Stobbart, A.-M., Tedds, J., Warwick, R., Webb, N., West, R., Worrall, D., and Yuan, W. (2009). The XMM-Newton serendipitous survey. V. The Second XMM-Newton serendipitous source catalogue. Astronomy and Astrophysics, 493:339–373.

Weatherhead, E. C., Reinsel, G. C., Tiao, G. C., Meng, X.-L., Choi, D., Cheang, W.-K., Keller, T., DeLuisi, J., Wuebbles, D. J., Kerr, J. B., and Miller, A. (1998). Factors affecting the detection of trends: Statistical considerations and applications to environmental data. Journal of Geophysical Research: Atmospheres, 103(D14):17149–17161.

XMM Catalogue public pages (2008). XMM-Newton serendipitous source catalogue: 2XMM. `http://xmmssc-www.star.le.ac.uk/Catalogue/xcat_public_2XMM.html`.

Xu, J., Van Dyk, D. A., Kashyap, V. L., Siemiginowska, A., Connors, A., Drake, J., Meng, X.-L., Ratzlaff, P., and Yu, Y. (2014). A fully Bayesian method for jointly fitting instrumental calibration and X-ray spectral models. The Astrophysical Journal, 794(2):97.

A   Tables of Data Description and Prior Influence

Tables 3 and 4 give summaries of the data used in Sections 4.1, 4.2 & 4.3.

Lines (Sources)	He-like OVII	H-like OVIII	He-like NeIX	H-like NeX
Spectrum	21.805Å	18.969Å	13.447Å	12.135Å

Telescopes	*Chandra*	XMM-*Newton*	*Suzaku*	*Swift*
Detectors (Instruments)	HETG, ACIS-S	RGS, EPIC-MOS, EPIC-pn	XIS	XRT

Table 3: Summary of E0102 data. The first table gives the sources for two data sets, Highly ionized Oxygen and Neon. The second table gives instruments for both data sets.

Observatory	XMM-*Newton* European Photon Imaging Cameras (EPIC)					
Detectors (Instruments)	EPIC-pn (pn), EPIC-MOS (MOS1 & MOS2)					
Data Acronym	2XMM			XCAL		
Energy Band	Hard	Medium	Soft	Hard	Medium	Soft
Energy (keV)	2.5-10.0	1.5-2.5	0.5-1.5	2.5-10.0	1.5-2.5	0.5-1.5
No. Sources	41	41	42	94	103	108

Table 4: Summary of 2XMM data and XCAL data. The number of instruments is $N = 3$ (pn, MOS1, MOS2) and the number of sources (M) is given in the last row for the six data sets, three from different energy bands of 2XMM data and XCAL data respectively.

B   Details of Fitting the Log-Normal Model

The following three MCMC algorithms are used for our posterior sampling.

1. **Standard Gibbs Sampler:** iterates the following three sets of conditional distributions, all easily derived from (2.5):

   (a) Conditioning on $\boldsymbol{G}$ and $\boldsymbol{\sigma^2}$, sample B_i independently for $i = 1, \ldots, N$ from
   $$\mathcal{N}\left(\frac{b_i/\tau_i^2 + \sum_{j \in J_i}(y_{ij} + 0.5\sigma_i^2 - G_j)/\sigma_i^2}{1/\tau_i^2 + \sum_{j \in J_i} 1/\sigma_i^2}, \ \frac{1}{1/\tau_i^2 + \sum_{j \in J_i} 1/\sigma_i^2}\right).$$

   (b) Conditioning on $\boldsymbol{B}$ and $\boldsymbol{\sigma^2}$, sample G_j independently for $1 \leq j \leq M$ from
   $$\mathcal{N}\left(\frac{\sum_{i \in I_j}(y_{ij} + 0.5\sigma_i^2 - B_i)/\sigma_i^2}{\sum_{i \in I_j} 1/\sigma_i^2}, \ \frac{1}{\sum_{i \in I_j} 1/\sigma_i^2}\right).$$

   (c) Conditioning on $\boldsymbol{B}$ and $\boldsymbol{G}$, sample σ_i^2 independently for $i = 1, \ldots, N$ from
   $$\sigma_i^{-|J_i|-2-2\alpha} \exp\left\{-\frac{1}{2}\frac{\sum_{j \in J_i}(y_{ij} - B_i - G_j)^2 + 2\beta}{\sigma_i^2} - \frac{|J_i|\sigma_i^2}{8}\right\}$$

   via the Metropolis-Hastings algorithm using a simple random walk proposal (Gaussian proposal) on the log-scale, i.e., $\log(\sigma_i^2)$.

2. **Block Gibbs Sampler**: same as above except replace the two conditional steps (1a) and (1b) by a joint draw of $\{\boldsymbol{B}, \boldsymbol{G}\}$ from $(N+M)$-dimensional Gaussian with mean $\boldsymbol{\Omega(\sigma^2)}^{-1}\boldsymbol{\gamma(\sigma^2)}$ and covariance matrix $\boldsymbol{\Omega(\sigma^2)}^{-1}$; see (2.7) and (2.8) in Section 2.2.

Instrument	Oxygen		Neon	
	$\tau = 0.025$	$\tau = 0.05$	$\tau = 0.025$	$\tau = 0.05$
RGS1	0.570	0.205	0.063	0.016
MOS1	0.279	0.077	0.075	0.019
MOS2	0.355	0.065	0.077	0.017
pn	0.250	0.041	0.620	0.218
ACIS-S3	0.218	0.040	0.270	0.088
ACIS-I3	0.906	0.640	0.099	0.026
HETG	0.648	0.341	0.129	0.034
XIS0	0.180	0.051	0.069	0.018
XIS1	0.298	0.078	0.071	0.019
XIS2	0.463	0.140	0.063	0.016
XIS3	0.772	0.364	0.062	0.018
XRT-WT	0.726	0.278	0.154	0.026
XRT-PC	0.934	0.235	0.906	0.017

Table 5: Proportion of prior influence, as defined by $1 - W_i$ (of (2.11)), for E0102 data in Section 4.1.

3. **Hamiltonian Monte Carlo (HMC)**: samples the entire vector $\boldsymbol{\theta} = \{B_i, G_j, \sigma_i^2\}$ through the non-U-turn HMC sampler (Hoffman and Gelman, 2014), implemented with the `STAN` package. Here we give a brief description of HMC; see Neal (2011) for more details. Let $\pi(\boldsymbol{\theta})$ denote the (unnormalized) joint posterior $\boldsymbol{\theta}$, as given by (2.5). Define potential energy as $U(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta})$ and kinetic energy as $k(\boldsymbol{p}) = \boldsymbol{p}^\top \mathcal{M}^{-1} \boldsymbol{p}$, where $\mathcal{M}$ is a symmetric positive-definite matrix, thus the total energy is $H(\boldsymbol{\theta}, \boldsymbol{p}) = U(\boldsymbol{\theta}) + k(\boldsymbol{p})$. We can obtain samples of $\pi(\boldsymbol{\theta})$ by sampling from the target density $\exp[-H(\boldsymbol{\theta}, \boldsymbol{p})] \propto \pi(\boldsymbol{\theta}) \exp(-\boldsymbol{p}^\top \mathcal{M}^{-1} \boldsymbol{p})$, which is essentially a data-augmentation technique (Tanner and Wong, 1987). By defining the potential energy and kinetic energy, we can propose MCMC moves according to the Hamiltonian dynamics, which explores the parameter space more efficiently by taking bigger and less correlated moves, as opposed to random walk Metropolis-Hastings or a Gibbs sampler. In practice, we use the leapfrog move to approximate the Hamiltonian dynamics. Due

Data Name	$\tau_i = 0.025$			$\tau_i = 0.05$		
	pn	mos1	mos2	pn	mos1	mos2
hard band 2XMM	0.093	0.075	0.082	0.025	0.020	0.022
medium band 2XMM	0.250	0.216	0.222	0.076	0.065	0.067
soft band 2XMM	0.093	0.075	0.069	0.025	0.020	0.018
hard band XCAL	0.010	0.019	0.031	0.003	0.005	0.008
medium band XCAL	0.023	0.016	0.028	0.006	0.004	0.007
soft band XCAL	0.021	0.011	0.007	0.005	0.003	0.002

Table 6: Proportion of prior influence for data used in the analysis in Sections 4.2 and 4.3.

to the energy-preserving property of Hamiltonian dynamics, the acceptance rate of the resulting HMC is approximately 1. It is not exactly 1 because we use the (discretized) leapfrog moves to approximate (continuous) Hamiltonian dynamics. The tuning parameters of the HMC algorithm include the covariance matrix $\mathcal{M}$, the leapfrog step size ϵ, and the number of leapfrog steps L. These are all self-tuned in the STAN package.

We compare the performance of these three algorithms using auto-correlation plots of the posterior samples and the effective sample size, in both the simulated and real data examples. Not surprisingly, the Gibbs sampler converges very slowly relative to the other two algorithms. We are able to cross check our results by comparing the samples obtained with the block Gibbs sampler and HMC – they give practically the same posterior distributions.

C   Proprieties of the Posterior Distribution

C.1   Propriety of Posterior

Theorem 1. *Under the prior specifications for $\{B_i, G_j, \sigma_i^2 : 1 \leq i \leq N, 1 \leq j \leq M\}$ given in (2.4), the posterior is proper if each source is measured by at least one instrument, i.e., $|I_j| \geq 1$ for all $1 \leq j \leq M$.*

Proof. We prove the propriety of the posterior by first integrating out the G_j first, then the B_i,

and finally the σ_i^2. By (2.5), $p(\boldsymbol{B}, \boldsymbol{G}, \boldsymbol{\sigma}^2 | \boldsymbol{D}, \boldsymbol{\tau}^2)$ is proportional to

$$\prod_{i=1}^{N} \sigma_i^{-|J_i|-2-2\alpha} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{M} \sum_{i \in I_j} \sigma_i^{-2} (y'_{ij} - B_i - G_j)^2 - \sum_{i=1}^{N} \left[\frac{(b_i - B_i)^2}{2\tau_i^2} + \frac{\beta}{\sigma_i^2} \right] \right\}. \qquad (C.1)$$

Now for each $1 \leq j \leq M$, if we define a random index $\mathcal{I}$ on I_j such that $\Pr(\mathcal{I} = i) \propto \sigma_i^{-2}$, then

$$\frac{\sum_{i \in I_j} \sigma_i^{-2} (y'_{ij} - B_i - G_j)^2}{\sum_{i \in I_j} \sigma_i^{-2}} = \mathrm{E}\left[y'_{\mathcal{I}j} - B_{\mathcal{I}} - G_j \right]^2 \geq \left[\mathrm{E}(y'_{\mathcal{I}j} - B_{\mathcal{I}}) - G_j \right]^2. \qquad (C.2)$$

Therefore, the first term in the exponential part of (C.1) is less than $-0.5 \left(\sum_{i \in I_j} \sigma_i^{-2} \right) (G_j - C_j)^2$, where $C_j = \mathrm{E}(y'_{\mathcal{I}j} - B_{\mathcal{I}})$ is free of G_j. The property of Normal density (for G_j) then yields

$$\int p(\boldsymbol{B}, \boldsymbol{G}, \boldsymbol{\sigma}^2 | \boldsymbol{D}, \boldsymbol{\tau}^2) \, d\boldsymbol{G} \leq C^* \prod_{i=1}^{N} \sigma_i^{-|J_i|-2-2\alpha} \prod_{j=1}^{M} \left[\sum_{i \in I_j} \sigma_i^{-2} \right]^{-1/2} \exp\left\{ -\sum_{i=1}^{N} \left[\frac{(b_i - B_i)^2}{2\tau_i^2} + \frac{\beta}{\sigma_i^2} \right] \right\}$$

where C^* is a constant that depends only on $\boldsymbol{D}, \boldsymbol{\tau}^2$. Integrating out $\boldsymbol{B}$ then gives

$$\int \int p(\boldsymbol{B}, \boldsymbol{G}, \boldsymbol{\sigma}^2 | \boldsymbol{D}, \boldsymbol{\tau}^2) \, d\boldsymbol{G} \, d\boldsymbol{B} \leq C^{**} \prod_{i=1}^{N} \sigma_i^{-|J_i|-2-2\alpha} \prod_{j=1}^{M} \left[\sum_{i \in I_j} \sigma_i^{-2} \right]^{-1/2} \exp\left\{ -\sum_{i=1}^{N} \frac{\beta}{\sigma_i^2} \right\} \qquad (C.3)$$

where C^{**} is a constant that depends only on $\boldsymbol{D}, \boldsymbol{\tau}^2$. Since I_j is non-empty, it is meaningful to invoke the well-known harmonic-geometric mean inequality to obtain that

$$\prod_{j=1}^{M} \left[\sum_{i \in I_j} \sigma_i^{-2} \right]^{-1/2} \leq \prod_{j=1}^{M} |I_j|^{-1/2} \left[\prod_{i \in I_j} \sigma_i \right]^{1/|I_j|} \leq \prod_{i=1}^{N} \sigma_i^{\sum_{j \in J_i} |I_j|^{-1}}. \qquad (C.4)$$

Inequalities (C.3) and (C.4) together imply that the unnormalized $p(\boldsymbol{\sigma}^2 | \boldsymbol{D}, \boldsymbol{\tau}^2)$ is dominated above by a constant times $\prod_{i=1}^{N} p_i(\sigma_i^2)$, where $p_i(x)$ is the density of the inverse Gamma distribution with shape parameter $\alpha_i = \alpha + [|J_i| - \sum_{j \in J_i} |I_j|^{-1}]/2$ and scale parameter β. Because $|I_j| \geq 1$, we have $\alpha_i \geq \alpha$. Hence as long as the hyperparameter $\alpha > 0$, which is always chosen to be so, p_i is a proper density. Consequently, $p(\boldsymbol{\sigma}^2 | \boldsymbol{D}, \boldsymbol{\tau}^2)$ is a proper density after renormalization.

$\square$

C.2 Identifiability

When τ_i^2 is large, the likelihood information for estimating B_i (i.e., from c_{ij}) dominates the prior information (i.e., from b_i). In the extreme case of $\tau_i^2 = \infty$, the model is not identifiable because for

37

fixed variances, $\{B_i, G_j\}$ and $\{B_i + \delta, G_j - \delta\}$ yield the same posterior densities for $\{\boldsymbol{B}, \boldsymbol{G}\}$ for any constant δ. Let $\lambda_{\max}$ and $\lambda_{\min}$ be the maximum and minimum eigenvalues of $\boldsymbol{\Omega}(\boldsymbol{\sigma^2})$, as defined in Section 2.2. Taking $u = (\mathbf{1}_N, \mathbf{1}_M)^\top$ and $v = (\mathbf{1}_N, -\mathbf{1}_M)^\top$, the *condition number* of $\boldsymbol{\Omega}(\boldsymbol{\sigma^2})$ is

$$\frac{\lambda_{\max}}{\lambda_{\min}} \geq \frac{u^\top \boldsymbol{\Omega}(\boldsymbol{\sigma^2})\, u}{v^\top \boldsymbol{\Omega}(\boldsymbol{\sigma^2})\, v} = 1 + \frac{4\sum_{i=1}^{N} |J_i| \sigma_i^{-2}}{\sum_{i=1}^{N} \tau_i^{-2}}, \tag{C.5}$$

where $\mathbf{1}_n$ denotes an $n \times 1$ vector of ones. As a consequence, when $\{\tau_i^2\}$ are generally larger than $\{\sigma_i^2\}$, the ratio in (C.5) can be large, and the posterior contours, determined by $\boldsymbol{\Omega}$, are elongated in one direction and narrow in another. This provides a guideline that $\{\tau_i^2\}$ should not be set too large relative to $\{\sigma_i^2\}$ in practice, because large $\{\tau_i^2\}$ can lead to near model non-identifiability and consequently more costly computation. A computationally cheaper way of dealing with possible model non-identifiability is to set one of the $\{B_i\}$ equal to a fixed value, which is equivalent to setting the corresponding $\tau_i = 0$. We experiment with this computationally cheap strategy in our empirical evaluations, and find that it does not alter the results in substantive ways, but the resulting estimators for the Effective Areas are relative to some (arbitrarily) chosen values instead of in absolute terms/magnitudes.

D Derivation of Conditional Covariance Matrix

In this section, we give detailed derivations of $\boldsymbol{\Omega}^{-1}(\boldsymbol{\sigma^2})$ when all instruments measure all sources. In this case, W_i defined in (2.11) becomes $W_i = \frac{M\sigma_i^{-2}}{M\sigma_i^{-2}+\tau_i^{-2}}$, $1 \leq i \leq N$. Define $\tilde{\sigma}^2 = \left(N^{-1} \sum_{i=1}^{N} \sigma_i^{-2} \right)^{-1}$.

Let $\boldsymbol{A}$ be the $(N + M) \times (N + M)$ diagonal matrix with diagonal elements equal to those of $\boldsymbol{\Omega}(\sigma^2)$. Let $\boldsymbol{U}$ be an $(N + M) \times 2$ matrix such that $U_{i,1} = \sigma_i^{-2}$, $U_{i,2} = 0$ for $i = 1, \ldots, N$, and $U_{j+N,1} = 0$, $U_{j+N,2} = 1$ for $j = 1, \ldots, M$. Let $\boldsymbol{C}$ be a 2×2 matrix such that $C_{i,j} = I_{i \neq j}$ $(i, j = 1, 2)$. Then $\boldsymbol{\Omega}(\sigma^2) = \boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{U}^\top$. By the Woodbury matrix identity, we have

$$\boldsymbol{\Omega}^{-1}(\boldsymbol{\sigma^2}) = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U} \left(\boldsymbol{C} + \boldsymbol{U}^\top \boldsymbol{A}^{-1} \boldsymbol{U} \right)^{-1} \boldsymbol{U}^\top \boldsymbol{A}^{-1}, \tag{D.1}$$

where $\boldsymbol{A}^{-1}$ is a diagonal matrix with diagonal elements

$$\left(\{ W_i \sigma_i^2 / M \}_{1 \leq i \leq N}, \{ \tilde{\sigma}^2 / N \}_{1 \leq j \leq M} \right).$$

Therefore, we can derive the inverse of 2×2 matrix $\boldsymbol{C} + \boldsymbol{U}^\top \boldsymbol{A}^{-1} \boldsymbol{U}$ as

$$
\left(\boldsymbol{C} + \boldsymbol{U}^\top \boldsymbol{A}^{-1} \boldsymbol{U} \right)^{-1} = \begin{pmatrix} \frac{\sum_{i=1}^N W_i \sigma_i^{-2}}{M} & 1 \\ 1 & \frac{M}{N} \tilde{\sigma}^2 \end{pmatrix}^{-1} = -\frac{\sum_{i=1}^N \sigma_i^{-2}}{\sum_{i=1}^N W_i \tau_i^{-2}} \begin{pmatrix} \frac{M^2 \tilde{\sigma}^2}{N} & -M \\ -M & \sum_{i=1}^N W_i \sigma_i^{-2} \end{pmatrix}.
$$

Further, let $\boldsymbol{W}$ be the $N \times 1$ column vector with ith element W_i, then we have

$$
\boldsymbol{A}^{-1} \boldsymbol{U} = \begin{pmatrix} \boldsymbol{W}/M & 0_{N \times 1} \\ 0_{M \times 1} & \tilde{\sigma}^2/N \; 1_{M \times 1} \end{pmatrix}.
$$

Consequently, $\boldsymbol{A}^{-1} \boldsymbol{U} \left(\boldsymbol{C} + \boldsymbol{U}^\top \boldsymbol{A}^{-1} \boldsymbol{U} \right)^{-1} \boldsymbol{U}^\top \boldsymbol{A}^{-1}$ is equal to

$$
- \left(\sum_{i=1}^N \tau_i^{-2} W_i \right)^{-1} \begin{pmatrix} \boldsymbol{W}\boldsymbol{W}^\top & -\boldsymbol{W} 1_{1 \times M} \\ -1_{M \times 1} \boldsymbol{W}^\top & [N^{-1}\tilde{\sigma}^2] \sum_{i=1}^N W_i \sigma_i^{-2} \; 1_{M \times M} \end{pmatrix}.
$$

Finally, we arrive at the closed-form expression for $\boldsymbol{\Omega}^{-1}(\boldsymbol{\sigma}^2)$:

$$
\left(\boldsymbol{\Omega}^{-1}(\boldsymbol{\sigma}^2) \right)_{i,i} = \frac{1}{M \sigma_i^{-2} + \tau_i^{-2}} \left\{ 1 + \frac{M \sigma_i^{-2} W_i}{\sum_{u=1}^N \tau_u^{-2} W_u} \right\},
$$

$$
\left(\boldsymbol{\Omega}^{-1}(\boldsymbol{\sigma}^2) \right)_{i,j+N} = \left(\boldsymbol{\Omega}^{-1}(\boldsymbol{\sigma}^2) \right)_{j+N,i} = -W_i \left(\sum_{u=1}^N \tau_u^{-2} W_u \right)^{-1},
$$

$$
\left(\boldsymbol{\Omega}^{-1}(\boldsymbol{\sigma}^2) \right)_{j+N,j+N} = \left(\sum_{u=1}^N \sigma_u^{-2} \right)^{-1} \left\{ 1 + \frac{\sum_{i=1}^N W_i \, \sigma_i^{-2}}{\sum_{u=1}^N W_u \, \tau_u^{-2}} \right\}.
$$

E Likelihood Method

E.1 MLEs and Their Asymptotic Variances

Note that the variance-covariance matrix of the MLEs $\{\hat{\boldsymbol{B}}, \hat{\boldsymbol{G}}\}$ is in fact $\boldsymbol{\Omega}^{-1}(\boldsymbol{\sigma}^2)$ as defined in (2.8). Therefore, we have the following proposition.

Proposition 1. *If all detectors measure all objects, i.e., $J_i = \{1, \ldots, M\}$, $I_j = \{1, \ldots, N\}$ and $\{\sigma_i^2, \tau_i^2\}$ are known constants, then the variances of $\{\hat{B}_i\}$, $\{\hat{G}_j\}$ are given by*

$$
\mathrm{Var}(\hat{G}_j) = \left[\sum_{i=1}^N \sigma_i^{-2} \right]^{-1} \mathcal{S}_G, \quad \mathrm{Var}(\hat{B}_i) = \left[M \sigma_i^{-2} + \tau_i^{-2} \right]^{-1} \mathcal{S}_B^{(i)}, \tag{E.1}
$$

where the inflation factors $\mathcal{S}_G, \{\mathcal{S}_B^{(i)}\}$ are given by

$$\mathcal{S}_G = 1 + \frac{\sum_{i=1}^{N} \sigma_i^{-2} W_i}{\sum_{i=1}^{N} \tau_i^{-2} W_i}, \quad \mathcal{S}_B^{(i)} = 1 + \frac{M \sigma_i^{-2} W_i}{\sum_{u=1}^{N} \tau_u^{-2} W_u}.$$

Moreover, we have $\mathrm{Cov}(\hat{B}_i, \hat{G}_j) = -W_i \left[\sum_{k=1}^{N} \tau_k^{-2} W_k\right]^{-1}$.

Remark 1. *Under the additive model, B_i and G_j are negatively correlated for all i, j. The asymptotic variances of $\hat{B}_i$ and $\hat{G}_j$ can be written as*

$$\mathrm{Var}(\hat{G}_j) = \mathrm{Var}(\tilde{G}_j)\mathcal{S}_G, \quad \mathrm{Var}(\hat{B}_i) = \mathrm{Var}(\tilde{B}_i)\mathcal{S}_B^{(i)},$$

where $\mathrm{Var}(\tilde{G}_j) = \left[\sum_{i=1}^{N} \sigma_i^{-2}\right]^{-1}$ is the inverse precision, i.e., asymptotic covariance, of $\hat{G}_j$ when the B_i are known constants; $\mathrm{Var}(\tilde{B}_i) = [M\sigma_i^{-2} + \tau_i^{-2}]^{-1}$ is the inverse precision, i.e., asymptotic covariance, of $\hat{B}_i$ when the G_j are known constants. The inflation factors $\mathcal{S}_G$ and $\mathcal{S}_B^{(i)}$ adjust for the fact that none of the B_i or the G_j are known.

Proposition 1 directly yields the following asymptotic results as $N, M \to \infty$.

Corollary 1. *If $\{\sigma_i/\tau_i\}$ are uniformly bounded from below and above by finite positive constants, and $\sum_{i=1}^{N} \sigma_i^{-2}/N$ converges to a positive constant as $N \to \infty$, then for all i, j, as $N, M \to \infty$,*

$$\mathrm{Var}(\hat{G}_j) = O(N^{-1}), \ \mathrm{Var}(\hat{B}_i) = O(N^{-1} + M^{-1}), \ \mathrm{Cov}(\hat{B}_i, \hat{G}_j) = -O(N^{-1}).$$

Specifically, when $\tau = \tau_1 = \cdots = \tau_N$ and $\sigma = \sigma_1 = \cdots = \sigma_N$, (E.1) simplifies to

$$\mathrm{Var}(\hat{G}_j) = \frac{\sigma^2}{N}, \ \mathrm{Var}(\hat{B}_i) = \frac{1}{M\sigma^{-2} + \tau^{-2}}\left(1 + \frac{M\sigma^{-2}}{N\tau^{-2}}\right), \ \mathrm{Cov}(\hat{B}_i, \hat{G}_j) = -\frac{\tau^2}{N}.$$

Remark 2. *The results above show that the asymptotic variances for $\{B_i\}$ and $\{G_j\}$ are not 'exchangeable' (i.e., switching $\boldsymbol{B}$ and $\boldsymbol{G}$ and correspondingly N and M), mainly for three reasons: first, for each B_i we assign an informative prior $\mathcal{N}(b_i, \tau_i^2)$ whereas for each G_j we assign a flat prior on the real line; second, for each instrument i, besides B_i, we also need to estimate σ_i^2; last, the measurement uncertainty depends only on the instrument but not on the sources (recall that $\sigma_{ij}^2 = \sigma_i^2$ for all i, j).*

E.2 Goodness-of-fit

We now give a goodness-of-fit test statistic for the random-effect regression model. Under the model (2.4), we have the following normalized residual sum of squares:

$$T(\boldsymbol{B}, \boldsymbol{G}) := \sum_{i=1}^{N} \frac{(b_i - B_i)^2}{\tau_i^2} + \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\left(y'_{ij} - B_i - G_j\right)^2}{\sigma_i^2}. \tag{E.2}$$

We see this sum of squares has two parts. The first part involves $\{b_i\}$ only, measuring how good the prior means are relative to the prior variances $\{\tau_i\}$. The second part depends on $\{y_{ij}\}$ only, and it will allow us to access how good the fitted $\boldsymbol{B}, \boldsymbol{G}$ are relative to the sampling variances $\boldsymbol{\sigma}^2$. Here we put them together as an overall model check, but one can certainly use them individually if one wants to check the prior distribution and likelihood model separately.

Theorem 2. *When the variances* σ_i^2, τ_i^2 *are known and we insert the MLEs of* B_i *and* G_j *into (E.2), we obtain* $T(\hat{\boldsymbol{B}}, \hat{\boldsymbol{G}}) \sim \chi_{NM-M}^2$.

Proof. This conclusion regarding χ^2 distribution follows from standard results on residual sum of squares of linear regression with Gaussian error. To figure out the correct degrees of freedom, we have $(NM + N)$ independent observations in total, but with $N + M$ parameters. Therefore, the degrees of freedom for the residual sum of squares is $NM - M$. $\qquad\square$

With unknown variances we do not have a closed-form distribution of T as defined in formula (E.2). Heuristically, we invoke the standard large-sample arguments and to continuously use the χ^2 approximation, but reduce the degrees of freedom to $MN - M - N$ to count for the number of estimated variance parameter $\{\sigma_i^2\}$. The resulting p-values of the fitted data in Sections 4.2 and 4.3 are not significant.

F More Simulation Results Under Misspecified Models

In Simulations IV and V, we generate data as $c_{ij} = \lambda_{ij} X_{ij}$, where $X_{ij} \sim \text{Poisson}(A_i F_j)$, and independently $\lambda_{ij} \sim \text{Uniform}[0.8, 1.2]$ for Simulation IV and $\lambda_{ij} \sim \text{Uniform}[0.4, 1.6]$ for Simulation V. In Simulations VI and VII, we generate data from $c_{ij} \sim \text{Poisson}(\lambda_{ij} A_i F_j)$, where the λ_{ij} are randomly generated from the uniform distribution on $[0.8, 1.2]$. The other parameters are set to be

the same as in Simulation II except that $\beta = 0.01$ for these simulations. Simulations VI and VII resemble the cases where the true model is Poisson and the estimation of T_{ij} is volatile, whereas Simulations IV and V resemble the cases that happen in practice, where the photon counts are multiplied by an adjustment factor, such as $\hat{T}_{ij}^{-1}$, as with the data pre-processing step for the XCAL data.

Figure 10 gives the results of Simulations IV and V. Figure 11 gives the results of Simulation VI with smaller counts ($B_i = 1$ and $G_j = 3$) and VII with larger counts ($B_i = 5$ and $G_j = 3$) under this scenario. It shows with large Poisson counts, controlling the uncertainty in the multiplicative constant can possibly lead to reasonably good results. Thus, even with compounded model mis-specification, the log-Normal model is able to provide reasonable, though not as precise, results, as compared with the correctly-specified case. However, when the misspecified "known constant" is highly variable, the fit result is not as satisfactory; plugging in a "guesstimated" σ_i in this case can give disastrously optimistic but biased results.
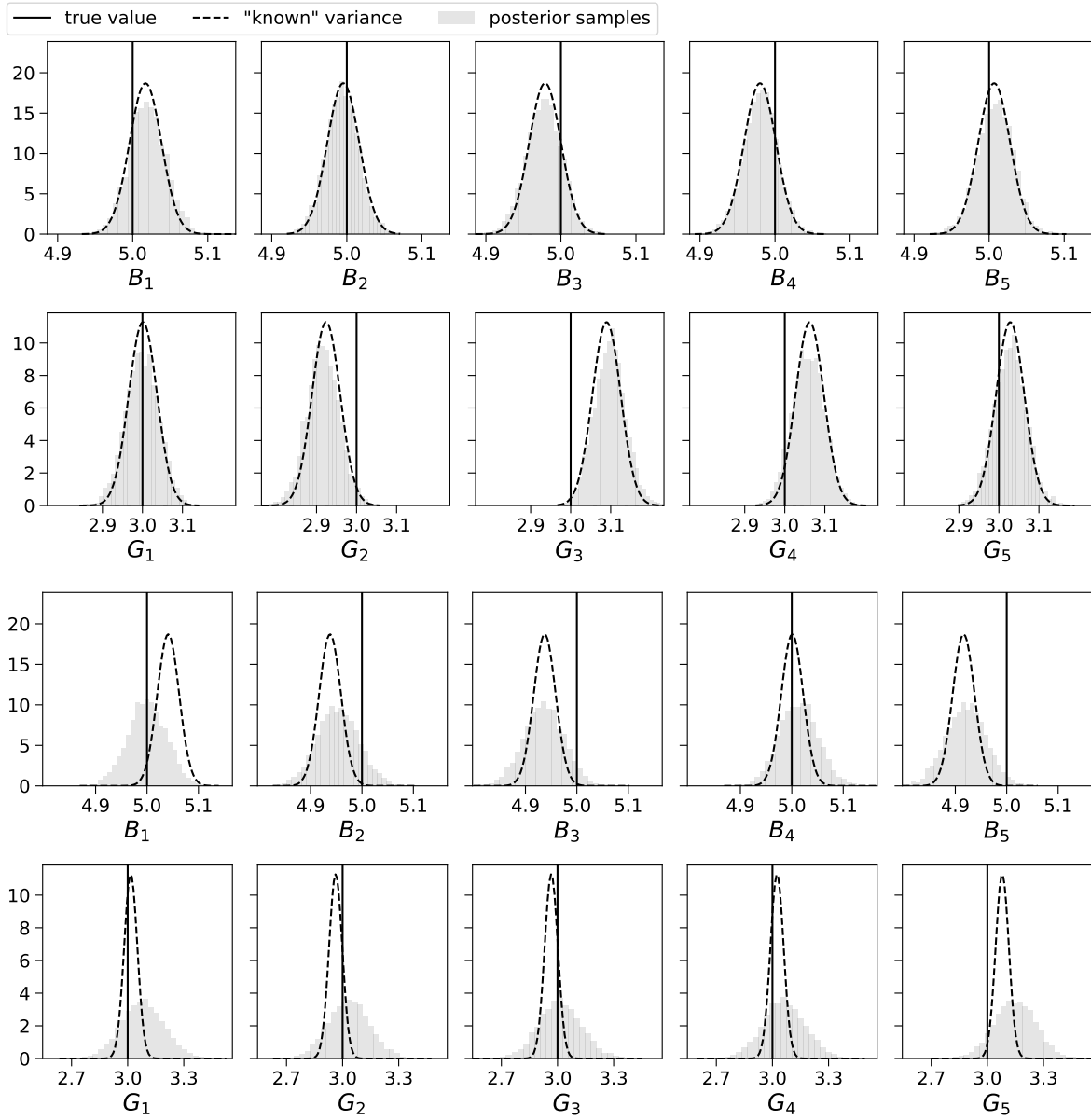
Figure 10: Simulations IV (rows 1 & 2) and V (rows 3 & 4). The gray shades are the posterior distributions of $\{B_i\}_{i=1}^5$ (rows 1 & 3) and $\{G_j\}_{j=1}^5$ (rows 2 & 4) fitted with unknown variances. The solid vertical black lines denote the true values. The black dashed density curves on top of the histograms denote the true posterior densities of $\{B_i\}$ and the $\{G_j\}$ with 'known' variances $\sigma_i^2 = 0.1^2$.
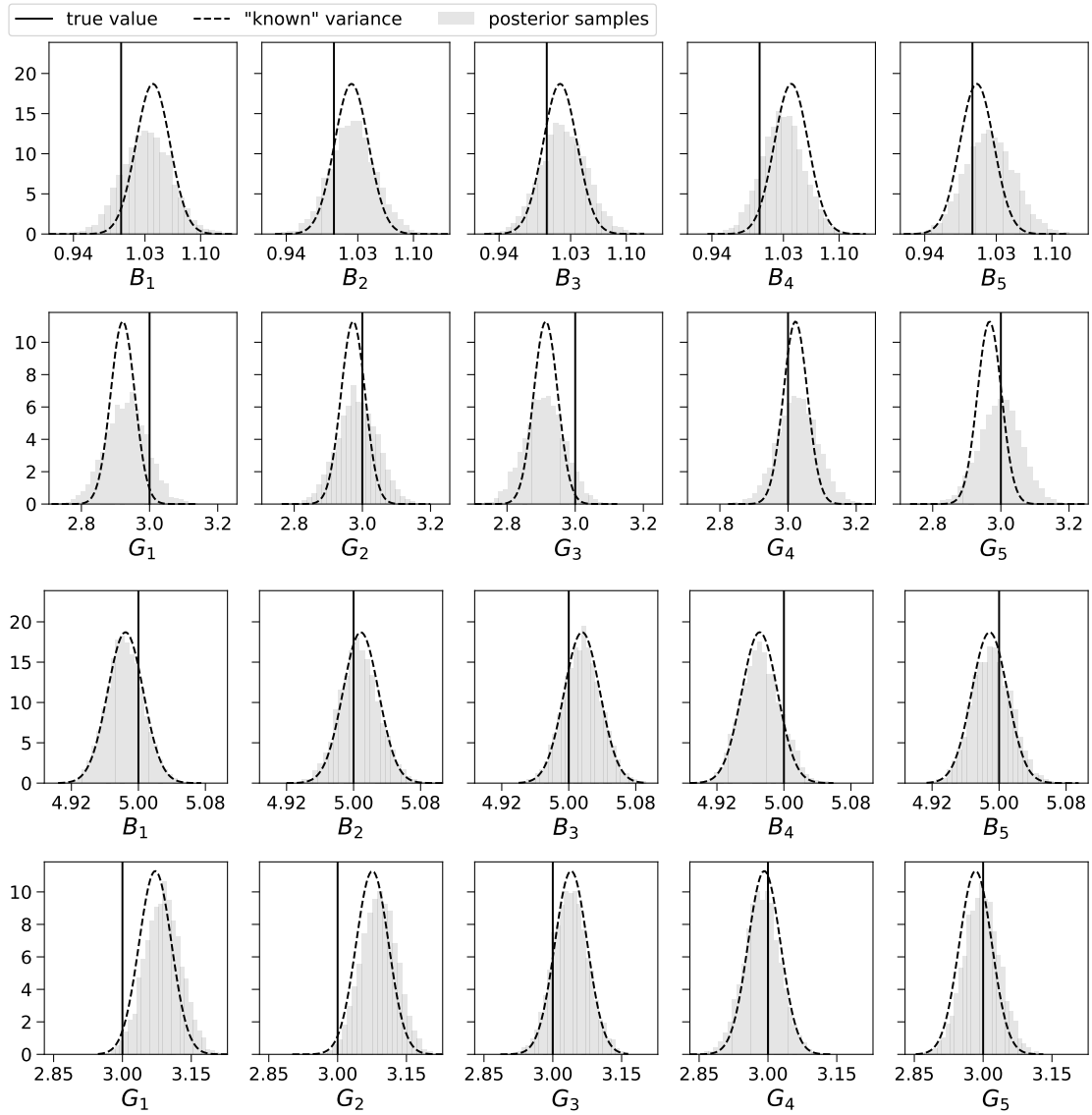
Figure 11: Same as Figure 10 but with Simulations VI (rows 1 & 2) and VII (rows 3 & 4).