The Astrophysical Journal, 890:153 (12pp), 2020 February 20 https://doi.org/10.3847/1538-4357/ab672b

Bayesian Revisit of the Relationship between the Total Field Strength and the Volume Density of Interstellar Clouds

Hangjin Jiang[1,2], Hua-bai Li[3], and Xiaodan Fan[2]
[1] Center for Data Science, Zhejiang University, Hangzhou, People's Republic of China
[2] Department of Statistics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, People's Republic of China; xfan@cuhk.edu.hk
[3] Department of Physics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, People's Republic of China; hbli@phy.cuhk.edu.hk
Received 2018 April 11; revised 2019 December 13; accepted 2019 December 29; published 2020 February 24

Abstract

The Zeeman effect has been the only method to directly probe the magnetic field strength in molecular clouds. The Bayesian analysis of Zeeman measurements carried out by Crutcher et al. is the only reference for cloud magnetic field strength. Here we extended their model and Bayesian analysis of the relation between field strength (B) and volume density (n) in the following three directions based on the recent observational and theoretical development. First, we take R, the observational uncertainty of n, as a parameter to be estimated from data. Second, the restriction of α, the index of the B–n relationship, is relieved from [0, 0.75] to [0, 1]. Third, we allow f, the minimum-to-maximum B ratio, to vary with n. Our results show that taking R as a parameter provides a better fitting to the B–n relationship and much more reliable estimates on R, f, and the changing point of α. Arguably our most important finding is that α cannot be reliably estimated by any of the models studied here, either from us or Crutcher et al., if $R > 2$, which is indeed the case from our estimate. This is the so-called errors-in-variables bias, a well known problem for statisticians.

Unified Astronomy Thesaurus concepts: Star formation (1569); Magnetic field (994); Molecular clouds (1072); Bayesian statistics (1900)

1. Introduction

It is increasingly clear that gravity, turbulence, and magnetic fields (B-fields) are indispensable in understanding the observed phenomena of star formation (McKee & Ostriker 2007; Crutcher 2012; Li et al. 2014). However, the details are still far from clear. For example, the interpretation of the Zeeman measurements is highly controversial (Crutcher et al. 2009, 2010; Mouschovias & Tassis 2010).

The scatter plot of line-of-sight B-field (B_z) against the number density (n) in Figure 1 includes most, if not all, of the Zeeman measurements ever made. Crutcher et al. (2010) concluded that a dynamically relevant B-field during core formation is inconsistent with Figure 1 because the upper limit of the B_z–n logarithmic plot has a slope α of $2/3$ (however, see the discussion in Section 4.2). Many criticisms focus on the data itself. For example, the OH measurements are from dark clouds while the CN data is mostly from massive cluster forming clumps in giant molecular clouds (GMCs). Since it is unlikely for dense cores of nearby dark clouds to evolve into massive cluster-forming clumps, using the slope to infer an isotropic collapse is questionable. Others are concerned by either the B-field morphologies (Mouschovias & Tassis 2010) or clump shapes (Tritsis et al. 2015). There has been no attempt to examine the Bayesian analysis that leads to the $2/3$-slope.

By observing Figure 1, here we introduce important parameters of the B–n model from Crutcher et al. (2010), and why we believe that the model can be improved. The first thing to notice in Figure 1 is the large vertical error bars, which show how difficult Zeeman measurements are. How about the horizontal error bars? In the analysis of Crutcher et al. (2010), the uncertainty, R, of n is fixed as a factor of two, while they stated that "the actual degree of uncertainty is not precisely known." However, the reliability of the statistical results can be very sensitive to R due to the errors-in-variables problem (see Section 4.1). Thus, instead of setting $R = 2$ as in Crutcher et al. (2010), we take it as an unknown parameter to be estimated from the Zeeman data in the Bayesian way, especially when there are good reasons to expect $R > 2$. Estimates of n in Figure 1 largely depend on the critical densities of the tracers, e.g., CN and CS (1–0), whose effective densities can be off from the critical densities by more than an order of magnitude (Shirley 2015); see more reasons given in Tritsis et al. (2015).

Another characteristic of Figure 1 is an increasing upper envelope for n above some threshold, n_0. The slope, α, of the upper envelope is limited to [0, 0.75] in Crutcher et al. (2010). This excludes the possibility of super-Alfvenic shocks, which can result in an α as high as 1. Our models accept all the physically possible α, which ranges within [0, 1].

Finally, three apparent factors contribute to the vertical scattering in Figure 1: projections, measurement uncertainties, and the intrinsic distribution of B. For a given n, Crutcher et al. (2010) assumed B uniformly distributed between a maximum and a minimum, which is a fraction, f, of the maximum. For simplicity, they set f as a constant over all n. We try to free f a little based on the following reason. The threshold n_0 may be related to the magnetic critical density (e.g., Li et al. 2013, 2014). Below this critical density, gas can only accrete along the field and move horizontally toward the right in Figure 1. Accretion happens in all directions above the critical density, which can compresses field lines and result in a positive slope in Figure 1. The lower the B, the lower the critical density and thus the accretion track in Figure 1 can turn upwards at lower density. The above fact will reduce f for densities beyond n_0, which is indeed also observed in simulations (see, for example, Mocz et al. 2017).

The remainder of the paper is organized as follows. In Section 2, we detail the model in Crutcher et al. (2010) and

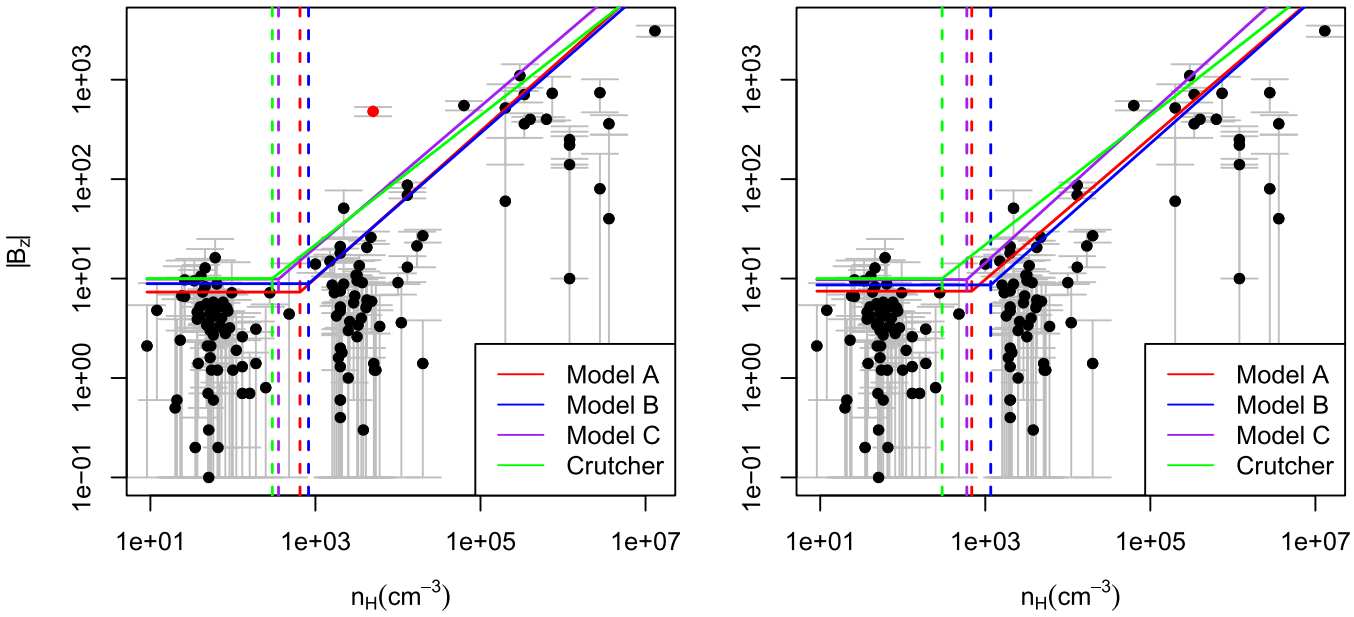

Figure 1. The Zeeman measurements of interstellar clouds from Crutcher et al. (2010) and all model fitting. The vertical axis represents the magnitude of the line-of-sight component $|B_z|$ (i.e., ignoring the direction of the line-of-sight component). The horizontal axis represents the number density $n_{\rm H}$. Corresponding to our symbols, the coordinate of a bullet point is $(\hat{n}_i, |B_i|)$ and the location of the error bar represents the 1σ uncertainty (i.e., σ_i). The red point in the left panel is the GMC Sgr-B2-North. The colored solid lines are the fitted maximum magnetic field strength (M_i) from different models. Among them, the green solid line (labeled as "Crutcher") in both panels is the result of Crutcher et al. (2010). All model fitting in the left panel used the data set with the red point (Dataset1), while the model fitting of Model A, B, and C in the right panel did not use the red point (Dataset2). The vertical dashed line marks the threshold n_0 of the corresponding fitting.

generalize it from aspects discussed above. Numerical results from simulation studies and Zeeman measurements based on different models are presented in Section 3 and discussed in Section 4. Finally, Section 5 concludes the paper. Technical details and extra results are presented in the Appendix.

2. Model and Inference

2.1. Extended Models

We adopt the same Zeeman data set as in Crutcher et al. (2010). The 137 observations of the line-of-sight component (denoted by B_z in Crutcher et al. 2010), the corresponding 1σ uncertainty, and number density are denoted by B_i, σ_i, and $\hat{n}_i$, $i = 1, \cdots, 137$, respectively. For the ease of presentation, we define the following frequently used symbols: $P(x)$ denotes the probability density function (PDF) of the continuous random variable x (or $P(x|Q)$ if emphasizing the parameter Q); $G(x|\lambda, \tau) \equiv \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{(x-\lambda)^2}{2\tau^2}}$ denotes the Gaussian PDF; $U(x|a, b) \equiv \frac{I(a \leqslant x \leqslant b)}{b-a}$ denotes the uniform distribution PDF; $I(A)$ denotes the indicator function that equals 1 if the statement A is true and 0 otherwise.

Following Crutcher et al. (2010), we assume the unobserved true number density (n_i) and the upper limit of cloud field (M_i) of a cloud satisfy the equation $M_i = B_0 I(n_i < n_0) + B_0 \left(\frac{n_i}{n_0}\right)^{\alpha} I(n_i \geqslant n_0)$. That is, the maximum magnetic field strength is some constant B_0 for clouds with density lower than some threshold n_0 and increases with density beyond n_0 as $M_i \propto n_i^{\alpha}$.

For the low density part ($n_i < n_0$), the observed line-of-sight field B_i is modeled as a Gaussian random variable centered at the true line-of-sight field A_i and with known variance σ_i^2; A_i equals the total magnetic field C_i times the cosine of the unknown angle between C_i and the observed line of sight to the cloud, thus $P(A_i|C_i) = U(A_i| -C_i, C_i)$; the total magnetic field C_i is assumed to be uniformly distributed in the interval $(f_1 M_i, M_i)$, where $0 < f_1 \leqslant 1$ and $M_i = B_0$. Thus, the observed line-of-sight field B_i is given by the convolution $P(B_i|n_0, B_0, \alpha, n_i, f_1) = \int G(B_i|A_i, \sigma_i) U(A_i|-C_i, C_i) U(C_i|f_1 B_0, B_0) dA_i dC_i$. For the high density part ($n_i \geqslant n_0$), the total magnetic field C_i is modeled as uniformly distributed between $f_2 M_i$ and M_i, i.e., $U(C_i|f_2 M_i, M_i)$, where $M_i = B_0 \left(\frac{n_i}{n_0}\right)^{\alpha}$. The observed line-of-sight field B_i in this case is given by the convolution $P(B_i|n_0, B_0, \alpha, n_i, f_2) = \int G(B_i|A_i, \sigma_i) U(A_i|-C_i, C_i) U(C_i|f_2 M_i, M_i) dA_i dC_i$. Finally, the observed number density $\hat{n}_i$ is modeled as $P(\hat{n}_i|R, n_i) = \frac{1}{2\hat{n}_i \ln R} I\left(\frac{1}{R} \leqslant \frac{\hat{n}_i}{n_i} \leqslant R\right)$, where R is the uncertainty of the observed number density $\hat{n}_i$ and n_i is the corresponding unknown real density. Thus, the likelihood function of the observed data point $(B_i, \hat{n}_i)$ is given by $P(B_i, \hat{n}_i|\theta, n_i) = P(B_i|\theta, n_i) P(\hat{n}_i|\theta, n_i)$ with $\theta = (f_1, f_2, \alpha, B_0, n_0, R)$.

In summary, we extended the model in Crutcher et al. (2010) from the following three aspects: (1) The distribution of the total magnetic field C_i is assumed to follow different uniform distributions for the lower density part and the higher density part. That is, the field minimum-to-maximum ratio f for the lower and higher density parts is assumed to be different. Specifically, we assume $C_i \sim U(C_i|f_1 M_i, M_i)$ for the lower density part, and $C_i \sim U(C_i|f_2 M_i, M_i)$ for the higher density part. (2) The uncertainty of observed number density R is taken as a parameter to be estimated from the Zeeman data under a prior distribution rather than fixing at 2 as in Crutcher et al. (2010). (3) The constraint on α is relaxed from [0, 0.75] to [0, 1]. We refer to this extended model as Model A. For comparison, we also study its reduced versions. Fixing $f_1 = f_2 \equiv f$ in Model A, we arrive at Model B. If we further fix $R = 2$, we reach at exactly the model in Crutcher et al. (2010),

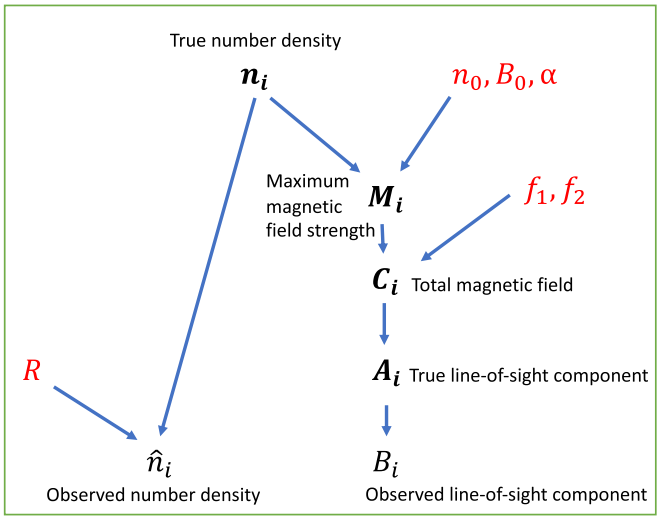

Figure 2. Dependency graph of Models A, B, and C. Among all the variables, the red ones are model parameters to be estimated, the bold ones are latent variables that are unobservable but of no direct interest, and the others are observed. Note that M_i is a deterministic function of other variables, thus not a new variable.

which we call Model C. We summarized these models in the following with the dependency graph given in Figure 2.

Model A

Model for B_i: $P(B_i|A_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} e^{-\frac{(B_i-A_i)^2}{2\sigma_i^2}}$, $P(A_i|C_i) = U(A_i| - C_i, C_i)$,

$$P(C_i|f_1, f_2, \alpha, B_0, n_0, n_i) = \begin{cases} U(C_i|f_1 M_i, M_i) & \text{if } n_i \leqslant n_0 \\ U(C_i|f_2 M_i, M_i) & \text{if } n_i > n_0 \end{cases},$$

where $M_i = B_0 I(n_i < n_0) + B_0 \left(\frac{n_i}{n_0}\right)^{\alpha} I(n_i \geqslant n_0)$.

Model for $\hat{n}_i$: $P(\hat{n}_i|R, n_i) = \frac{1}{2\hat{n}_i \ln R} I\left(\frac{1}{R} \leqslant \frac{\hat{n}_i}{n_i} \leqslant R\right)$, $R \geqslant 1$.

Parameters: $\theta = (f_1, f_2, \alpha, B_0, n_0, R)$.

Model B

Take $f_1 = f_2 \equiv f$.

Parameters: $\theta = (f, \alpha, B_0, n_0, R)$.

Model C (the model in Crutcher et al. 2010)

Take $f_1 = f_2 \equiv f$ and $R = 2$.

Parameters: $\theta = (f, \alpha, B_0, n_0)$.

2.2. Bayesian Inference

In this section, we present necessary elements of Bayesian inference for the above models to make them accessible for readers unfamiliar with Bayesian inference. Prior distribution, likelihood function, and posterior distribution are three key elements in Bayesian inference. The prior distribution models the prior knowledge about the parameter, the likelihood function summarizes the knowledge from the data, while the posterior distribution summarizes the knowledge about the parameter from the prior distribution and the data.

Likelihood function. The likelihood function summarizes the evidence from the data. Let D be the observed data set $\{B_i, \hat{n}_i\}_{i=1}^m$ with $m = 137$. For Models A, B, and C, the likelihood function is $P(D|\theta) = \prod_{i=1}^m \{I_{i1} + I_{i2}\}$, with I_{i1} and I_{i2} given in the A.1.

Prior distribution. In Bayesian analysis, the prior $P(\theta)$ should be set such that the resultant posterior distribution is proper. That is, the integral $\int P(D|\theta)P(\theta)d\theta$ should be finite for any D. No valid inference or summary can be made based on an improper posterior distribution. For Models A, B, and C, the prior distribution of the number density n_i is set as $P(n_i) \propto 1/n_i$, following that given in Crutcher et al. (2010). The prior distributions of f_1 and f_2 in Model A are set as a uniform distribution on [0, 1]. The prior distribution of α, is set as $P(\alpha) \propto I(0 < \alpha < 1)/\alpha$, more noninformative than that given in Crutcher et al. (2010; $P(\alpha) \propto I(0 < \alpha < 0.75)/\alpha$). Instead of following Crutcher et al. (2010) to set the prior distribution of n_0 as $P(n_0) \propto 1/n_0$, which has an infinite integral, we set it as $P(n_0) \propto 1/n_0^2$, which has a finite integral. Our prior penalizes more heavily the bigger n_0 values and improves the Markov chain Monte Carlo (MCMC) algorithm. The prior of B_0 is the same as that in Crutcher et al. (2010), i.e., $P(B_0) \propto$ constant. The prior distribution of the uncertainty R is set as $P(R) \propto I(R \geqslant 1)/R^2$ since a larger uncertainty is usually less probable. Note that the prior expectation of R and n_0 is infinity, which means the prior is pretty noninformative and their posterior estimates will be determined by data.

Posterior distribution. The posterior distribution of parameters is given by $P(\theta|D) \propto P(D|\theta)P(\theta)$, where $P(\theta)$ is the joint prior distribution.

Sampling from posterior distribution. In the Bayesian framework, all of the statistical inference are based on the posterior distribution. However, as shown in Appendix A.1, the posterior distribution in our problem is too complex to summarize analytically, thus an MCMC algorithm (Robert & Casella 2004) is designed to sample from this complex joint posterior distribution. The converged samples are used for statistical inference. Generally speaking, MCMC algorithms achieve the posterior sampling of a target density function $g(\theta)$ by evolving a Markov chain over the parameter space of θ iteratively. In the $(t+1)$-th iteration, we propose a candidate parameter y given the previous sample θ_t from a proposal distribution $q(y|\theta_t)$, then set $\theta_{t+1} = y$ with probability $r = \min\left(1, \frac{g(y)q(\theta_t|y)}{g(\theta_t)q(y|\theta_t)}\right)$ and $\theta_{t+1} = \theta_t$ with the remaining probability. In our case, $g(\theta)$ is the posterior distribution of parameters $P(\theta|D)$ as calculated in Equation (1) in Appendix A.1. An MCMC algorithm, more specifically, a Metropolis-within-Gibbs algorithm (Robert & Casella 2004), is used to sample from the posterior distribution, where we iteratively update each parameter by sampling from its univariate conditional posterior distribution with other parameters fixed at their latest values. When a conditional posterior distribution is difficult to sample directly, we use a Metropolis–Hastings algorithm to sample from it. The algorithms for Models B and C are very similar to Algorithm 1, thus we do not describe them here.

Convergence diagnosis of MCMC. To make sure that the Markov chain from an algorithm is converged, we run multiple Markov chains of the same algorithm independently starting from random initial values and compute the potential scale reduction factor (PSRF; Brooks & Gelman 1998) to diagnose the convergence. Usually PSRF $\leqslant 1.1$ indicates that the Markov chains are converged.

In our analysis, we run the Metropolis-within-Gibbs sampling algorithm for 20,000 iterations in each of three

independent chains. The PSRF of the second half iterations shows that the Markov chains have converged. We thin the second half of each chain by taking every tenth observation as an effort to reduce autocorrelation, and then merge all selected observations for posterior inference. The maximum a posterior (MAP), posterior mean and median can be reported as the point estimates of parameters in a model. If the posterior distribution of the parameter is unimodal and symmetric, MAP is the same as posterior mean and posterior median. In this paper, following Crutcher et al. (2010), we report the posterior median as the point estimate of parameters due to its robustness.

2.3. Model Comparison

Compared with Model C, Model B has an extra parameter, which means a better fitness of Model B to data may be due to its higher complexity. Thus, we should compare Models A, B, and C by taking into consideration the complexity of these models. To this aim, the Bayesian information criteria (BIC; Schwarz 1978) is used to compare these models, which is defined as $\mathrm{BIC}(M) = \ln(n)p_M - 2L(\hat{\theta}_M)$, where n is the sample size, p_M is the number of parameters in Model M, $\hat{\theta}_M$ is the estimate of parameter θ (here we use posterior median) in Model M, and $L(\hat{\theta}_M)$ is the log-likelihood at $\hat{\theta}_M$ of Model M. A smaller BIC value indicates a more preferable model.

3. Results

3.1. Simulation Study

Before applying the Bayesian procedure developed in Section 2.2 to Zeeman measurements, synthesized data sets with known underlying parameter values are used to evaluate the effectiveness of our algorithms, and understand the behavior of Bayesian estimator under different cases. In the simulation study, we know the true values of parameters in the model, and the Bayesian procedure can be evaluated through comparing the Bayesian estimates with the true values. However, we do not have the ground truth for the real data and do not know to what extent we can believe the results from our method if we apply the Bayesian procedure directly on the real data. Thus, the simulation study is necessary and important to evaluate the statistical method.

If a Bayesian algorithm is effective, the true parameter value underlying the data set shall be covered well by the posterior distribution inferred from the data, preferably in the high density area. More specifically, the empirical coverage rate of 95% highest posterior density interval (HPDI) of each parameter in the model shall be around 95% if the algorithm is effective in estimating the parameter (Cook et al. 2006), where the 95% HPDI is defined as the shortest interval in which the posterior samples located with probability 95%. Thus, the coverage rate of 95% HPDI of each parameter, estimated from 200 independent replicates for each combination of a model and an uncertainty level R, is used to evaluate the performance of Bayesian estimator.

To mimic the Zeeman data set, the true number density values (n_i) are set as the observed number density of the Zeeman data set and the uncertainty of observed line-of-sight field (σ_i) are set as corresponding values in the Zeeman data set. The uncertainty of observed number density R is set at $2r$ with $r = 1, 2, \cdots, 19$ to represent different levels of uncertainty for observations of the number density. Other parameter values are sampled from their prior distributions. Observed line-of-sight field and number density are then generated according to Models A, B, or C.

Our algorithm yields accurate estimates if number density has little observation uncertainty. The results shown in Figure 3 indicate that our algorithm produced satisfied coverage rates when data is generated from Model A/B/C with $R = 2$, i.e., with small observation uncertainty on number density. This suggests that our Bayesian algorithm is correct and can effectively recover parameter values if the uncertainty of number density is small.

Model B is preferred in terms of coverage rate when the uncertainty of number density is unknown. As shown in Figure 3, the coverage rate of f, α, and n_0 based on Model C decreases faster when the uncertainty of number density R increases but one estimates them by fixing $R = 2$. By comparison, Model B enjoys a much higher coverage rate for these parameters, which suggests that one should not fix R at some value when little is known about it.

Estimates on α and n_0 are unreliable when the uncertainty of number density is high. Although our Bayesian algorithm can recover true parameter values well when the uncertainty on number density observation is small, i.e., $R \leqslant 2$, the same Bayesian algorithm performed less accurately for α and n_0 when the uncertainty is cannot be neglected. As shown in Figure 3, for all three models, when the true uncertainty of observed number density (R) is 2, the coverage rates of 95% HPDI of all parameters are around 95%. However, the coverage rate of 95% HPDI of α decrease below 60% when R gets larger, and that of n_0 drops to around 80%, which is undesirable. These facts suggest that the Bayesian algorithm, although correct, can no longer effectively recover the true values for α and n_0 when the number density has a significant amount of observation uncertainty.

3.2. Zeeman Measurement

In this section, we apply our Bayesian procedure to Zeeman measurement. Since the red point in Figure 1, which is for the GMC Sgr-B2-North, might be an influential point as pointed out by Crutcher et al. (2010), we work on both the full data set (labeled Dataset1) and the data set without the red point (labeled Dataset2). We fit both Dataset1 and Dataset2 to Models A, B, and C using our MCMC algorithm, each with three independent runs starting from different initial parameter values. The three Markov chains are converged with PSRF < 1.1. The posterior median (with 95% HPDI) of each parameter in each model is summarized in Table 1. Furthermore, we show in Figure 1 the fitted lines based on Models A, B, and C. Crutcher et al. (2010) reported the posterior median as $(f, \alpha, B_0, n_0) = (0.03, 0.65, 10{,}300)$. To compare with it, we first set the prior of α as that in Crutcher et al. (2010), i.e., $P(\alpha) \propto I(0 < \alpha < 0.75)/\alpha$, and obtain posterior distributions of parameters similar to those presented in Figure 4 in Crutcher et al. (2010). Next, we set the prior of α as $P(\alpha) \propto I(0 < \alpha < 1)/\alpha$, and the posterior distributions of parameters are shown in Figure 4. Comparing results from $P(\alpha) \propto I(0 < \alpha < 0.75)/\alpha$ and $P(\alpha) \propto I(0 < \alpha < 1)/\alpha$ (see Table 1), we see that the constraint $\alpha < 0.75$ results in a smaller estimate on α and a shorter 95% HPDI due to the fact that it is more informative than $\alpha \in [0,1]$. As discussed before, the true value of α possibly locates in [0.75, 1], but the restriction $\alpha < 0.75$ will definitely lead to an estimate of α less

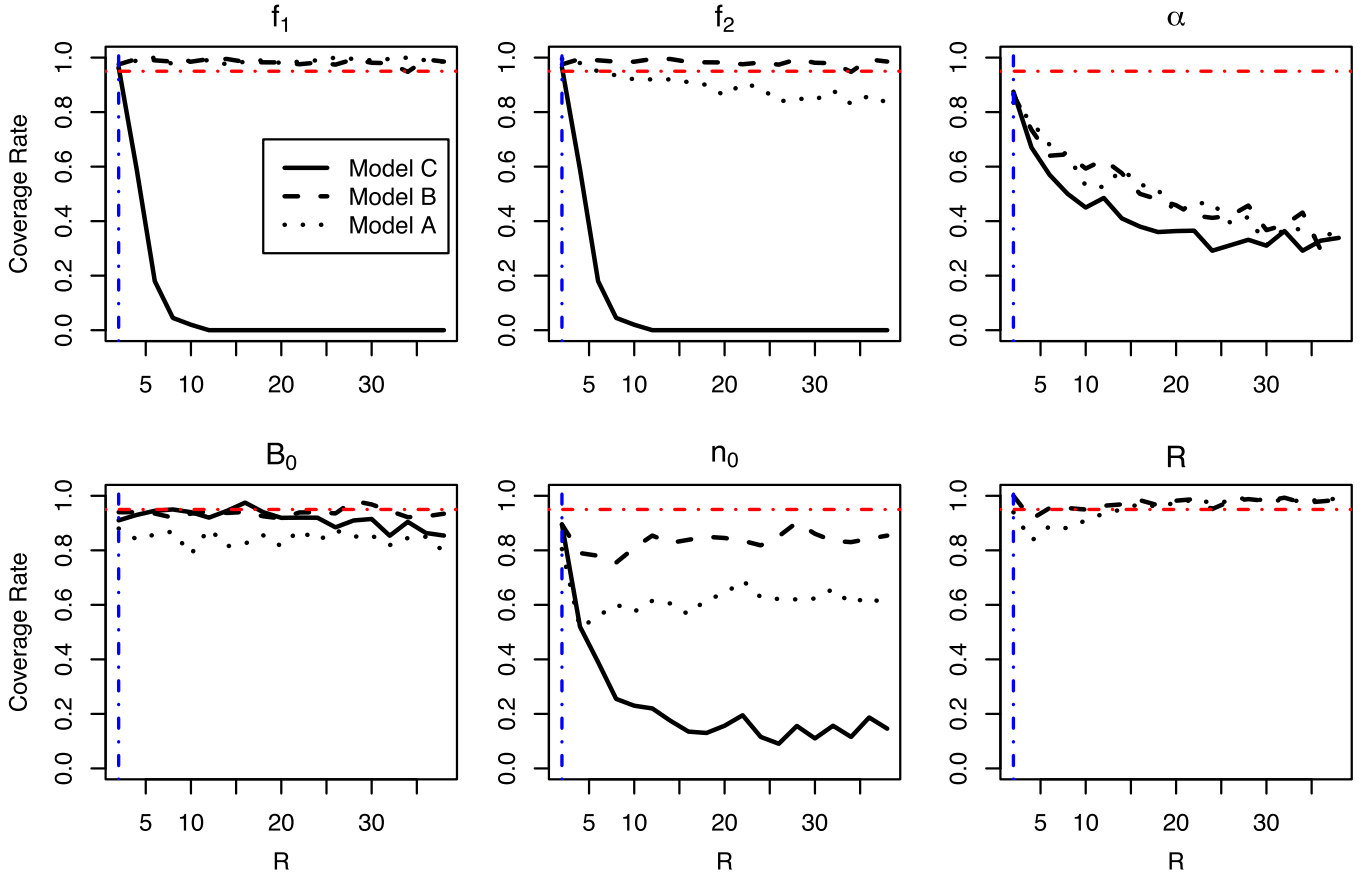

Figure 3. Coverage rate of the 95% HPDI of each parameter under different uncertainty of number density. The red dashed–dotted line corresponds to the target coverage rate 95%, and the blue dashed–dotted line marks $R = 2$. Note that $f_1 = f_2 = f$ in Model C and Model B.

than 0.75 regardless of the data, thus using $P(\alpha) \propto I(0 < \alpha < 0.75)/\alpha$ as the prior may underestimate α. However, there is no harm to allow α to take any value from [0, 1] in the weak prior, even if it actually locates in [0, 0.75].

The GMC Sgr-B2-North has a different impact on results from Model C and Model A/B. Both Model A and Model B give a significantly larger estimate of R from Dataset1 than that from Dataset2. This is consistent with the claim that the GMC Sgr-B2-North behaves much like an outlier with a larger uncertainty on number density than others (Tritsis et al. 2015). However, the GMC Sgr-B2-North shows no significant impact on the results from Model C, which is consistent with the claim made in Crutcher et al. (2010). This is caused by fixing R at an overly optimistic value and ignoring the larger uncertainty on the number density of GMC Sgr-B2-North. In other words, the estimated value of R is very sensitive to one or a small number of data points. Thus, it does not necessarily reflect the actual mean uncertainty. Furthermore, we evaluated the effect of each data point in Dataset2 on estimating R and concluded that each of them has little impact on R since removing each of them gives an estimate of R still around 9.3. This further confirms that GMC Sgr-B2-North is a special point with larger uncertainty on number density.

Model B is the best model among the three models. We compare the three models based on BIC using their posterior median estimates shown in Table 1. On Dataset2, we have BIC(C) − BIC(B) = 4.993 and BIC(B) − BIC(A) = −5.77, which indicates strong evidence that Model B fits the data better than Model A and strong evidence that Model B is more preferable than Model C (consistent with our conclusion based on the coverage rate from the simulation study). Thus, Model B is the best model. On Dataset1, we have BIC(C) − BIC(B) = 47.703 and BIC(B) − BIC(A) = −4.889, thus Model B is also the best. Therefore, our final estimate should be the estimate from Model B on Dataset2, which is (f, α, B_0, n_0, R) = (0.26, 0.72, 8.3, 1125, 9.3; see Figure 5). Comparing with the results presented in Figure 4 of Crutcher et al. (2010), our estimates on f, n_0, and R from Model B are significantly different.

Estimates on α and n_0 are unreliable. According to results reported in Table 1, the estimated uncertainty of number density (posterior median of R) from Dataset2 is 9.3 for Model B and 7.7 for Model A, respectively. These estimates are consistent with the literature survey of R by Tritsis et al. (2015), who compared volume densities adopted in Crutcher et al. (2009) with those appeared in the literature and found differences by factors between 2 and 60 with a mean at 15, if the potential outlier with $R \approx 400$ (the red point in Figure 1) is excluded. Figure 3 shows that the coverage rate of 95% HPDI for R is around 95% for R ranging from 2 to 38, thus our Bayesian algorithms for Models A and B can recover R accurately. On the other hand, the coverage rate of 95% HPDI for α and n_0 when $R = 9.3$ is only about 60% and 80%, respectively. Thus, the estimates on α and n_0 are unreliable, especially for the estimate of α.

In summary, we obtained a better fitting of Zeeman measurements by extending the model in Crutcher et al. (2010). It seems that our estimate of α, 0.72 with the 95% HPDI given by [0.58, 0.86], is significantly larger than 0.5.

Table 1
Posterior Median (and 95% HPDI in Brackets) of Parameters in Models A, B, and C

		f_1	f_2	α	B_0	n_0	R
Model A	DS1[a]	0.59 [0.10, 1.00]	0.05 [0.00, 0.16]	0.71 [0.55, 0.90]	7.1 [4.2, 10.1]	506 [46, 1440]	40.3 [14.1, 88.8]
	DS2[a]	0.53 [0.08, 0.99]	0.11 [0.00, 0.35]	0.72 [0.58, 0.83]	7.7 [5.7, 10.7]	732 [73, 1620]	7.7 [2.7, 16.1]
Model B	DS1[a]	0.08 [0.00, 0.31]	0.08 [0.00, 0.31]	0.73 [0.57, 0.94]	9.1 [6.1, 12.3]	852 [106, 2476]	44.1 [13.3, 117.7]
	DS2[a]	0.26 [0.00, 0.83]	0.26 [0.00, 0.83]	0.72 [0.58, 0.86]	8.3 [5.6, 11.4]	1125 [366, 2616]	9.3 [2.4, 19.0]
Model C^2	DS1[a]	0.02 [0.00, 0.09]	0.02 [0.00, 0.09]	0.71 [0.59, 0.84]	9.7 [7.3, 12.7]	346 [99, 765]	2
	DS2[a]	0.03 [0.00, 0.13]	0.03 [0.00, 0.13]	0.76 [0.63, 0.87]	9.6 [7.2, 12.3]	605 [202, 1091]	2
Model C^1	DS1[a]	0.02 [0.00, 0.08]	0.02 [0.00, 0.08]	0.68 [0.59, 0.75]	9.7 [7.2, 12.7]	286 [89, 573]	2
	DS2[a]	0.03 [0.00, 0.11]	0.03 [0.00, 0.11]	0.71 [0.63, 0.75]	9.8 [7.3, 12.9]	462 [184, 805]	2

Notes. For Models B and C, we have $f_1 = f_2 = f$.
The prior of α is $P(\alpha) \propto I(0 < \alpha < 0.75)/\alpha$ in Model C^1 and $P(\alpha) \propto I(0 < \alpha < 1)/\alpha$ in Model C^2.
[a] DS1 and DS2 are abbreviations for Dataset1 and Dataset2, respectively.

However, we should keep in mind that, due to the errors-in-variables problem (see Section 4.1), such 95% HPDIs only cover the true value of α with a probability around 60%. More efforts are needed to improve this result (see the discussion in Section 4.1).

On the other hand, since our simulation results in Section 3.1 show that our estimates on (f_1, f_2, B_0, R) are reliable and our estimate on n_0 is not too bad, one may ask whether the estimated α is underestimated or overestimated if the estimated values of (f_1, f_2, B_0, n_0, R) from Model B under Dataset2 are indeed the reality. We conducted another simulation study to check this. Datasets mimicking the Zeeman data set (Dataset2; see Section 3.1 for the mimicking procedure) are synthesized from Model B with (f_1, f_2, B_0, n_0, R) = (0.26, 0.26, 8.3, 1125, 9.3) and $\alpha = 0.05 + 0.05t$, $t = 1, ..., 18$. For each α, 200 data sets are synthesized. For each data set, our Bayesian method is used to fit Model B to the data. Figure 6 compares the posterior median estimates of α with corresponding true α values. It shows that, when (f_1, f_2, B_0, n_0, R) = (0.26, 0.26, 8.3, 1125, 9.3), the Bayesian method tends to underestimate α and a true $\alpha \leqslant 0.6$ can rarely lead to an estimate $\hat{\alpha} = 0.72$. That is, if the truth underlying Zeeman measurements is Model B with (f_1, f_2, B_0, n_0, R) = (0.26, 0.26, 8.3, 1125, 9.3), the real α is most likely larger than 0.6.

4. Discussion

4.1. Errors-in-variables Model

We studied through simulation the accuracy of the Bayesian approach for estimating parameters in Models A, B, and C, and found that the results for α and n_0 are unreliable, especially for α, when the uncertainty of observed number density is large, i.e., $R > 2$ (see Figure 3). We discuss in the following how to further improve the estimates. The discussion goes in three directions: (1) statistical inference on errors-in-variables models; (2) uncertainty of observations; and (3) sample size of the observations.

Statistically, it is difficult to estimate accurately parameters in Models A, B, and C when the uncertainty of number density is large. When the uncertainty of number density is large, Models A, B, and C are essentially errors-in-variables models. The incapability to recover true parameters as suggested by the low coverage rates of 95% HPDIs of α and n_0 (see Figure 3) is not strange to statisticians when dealing with such errors-in-variables models. The estimators of parameters in errors-in-variables models tend to be biased no matter whether Bayesian or frequentist approach is used and no matter how much data are collected. That is, these models lead to inconsistent estimates and may be intrinsically non-identifiable. A theoretical analysis of a simple linear errors-in-variables model is given to illustrate such biasedness problem in the A.2. For the Zeeman measurement, we show the bias of (f_1, f_2, α, B_0, n_0, R) in Models A, B, and C in Figure 7 obtained through 200 independent simulations under different R values when the true value of (f_1, f_2, α, B_0, n_0) is (0.03, 0.03, 0.65, 10, 300; i.e., the estimates reported by Crutcher et al. 2010). From the results, we see that the bias of parameters in these models is very small and the coverage rate of 95% HPDI is satisfied (see Figure 3), when $R = 2$. When R gets larger, the bias of α and n_0 tends to be larger, which leads to a lower coverage rate of 95% HPDIs. Interestingly, combining with results in Figures 3 and 7, we find that the uncertainty of number density has little impact on the accuracy of f and B_0 in Model B. Since the bias may depend on the true value of parameters, the trend of the bias shown in Figure 7 may be only true around the particular parameter setting, i.e., (f_1, f_2, α, B_0, n_0) = (0.03, 0.03, 0.65, 10, 300). It should not be considered as a general trend for all parameter settings, which is too complicated to obtain for models with so many parameters. The bias of parameter estimates in other errors-in-variables models can be found in many fields, such as survival analysis (Kong & Gu 1999), economics (Hsiao 1989; Li 2002), and epidemiological study (Frost & Thompson 2002). Some researchers proposed to correct the bias for some simple models based on a corrected version of the log-likelihood function, see Kong & Gu (1999) for an example. However, it is still an open problem to correct complex nonlinear errors-in-variables models.

Reducing the uncertainty of observations is helpful to improve the accuracy. One way to improve the reliability of results from these models is to reduce the uncertainty of observed number density, which, however, is very challenging. Another way is to reduce the uncertainty of observed B_z (Li & Pan 2016; Li et al. 2019; the Five-hundred-meter Aperture Spherical radio Telescope gives us hope). However, this method can not exclude the estimation bias but may only reduce the bias, which can be concluded from the simple linear errors-in-variables model (see Equation (2) in the A.2).

More data will be helpful to improve the results. One motivation of using Bayesian analysis is the possibility of including the "none detections" (data with signal to noise

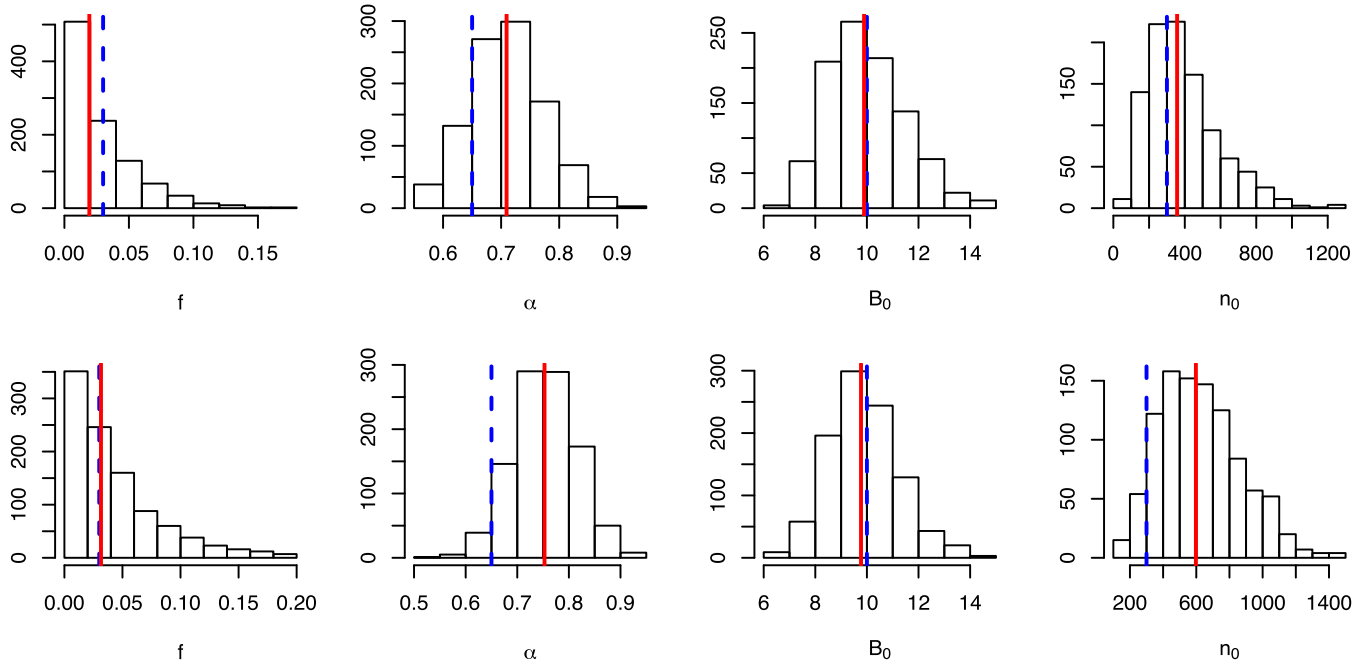

Figure 4. The histogram of converged posterior samples of parameters in Model C taking prior of α as $I(0 < \alpha < 1)/\alpha$ based on Dataset1 (the first row) and Dataset2 (the second row). The blue dashed line in each subplot denotes the estimated value of the corresponding parameter from Crutcher et al. (2010), and the red solid line corresponds to the posterior median from our algorithm.

ratio less than three), which are almost twice as much as the detections of the Zeeman measurements (Figure 1), into the analysis (Crutcher et al. 2010). However, we have already seen that larger uncertainties from observations can result in larger uncertainties in the estimate. Extra data will be helpful only if its uncertainty is under a certain limit, which may also be related to the amount of such extra data. To better understand this point, we show theoretically in the A.3 the effect of extra noisy data in estimating parameters for a simple linear model. The exact criterion for an acceptable noise level for Models A, B, or C is more difficult to acquire and is out of the focus of this work, but should be something kept in mind for future analyses.

4.2. Comparing with Numerical Simulations

The role of statistics in data analysis is to objectively infer in face of the uncertainty in data and give a full uncertainty quantification of the results such that we know to what extent the results we obtained should be believed. The significant uncertainty (the low HPDI coverage rate) we showed here can explain the discrepancy between the results from Bayesian analysis and from magnetohydrodynamic simulations of molecular clouds. All the simulations we can find from the literature (e.g., Li et al. 2015; Mocz et al. 2017 and Zhang et al. 2019), no matter super- or sub-Alfvenic, predict $n_0 > 10^4$, yet the n_0 estimated by Crutcher et al. (2010) is 300. As shown in Figure 3, our Model B improves the coverage rate of HPDI of n_0 from Model C, which is equivalent to the model in Crutcher et al. (2010), by more than 200% for $R > 8$ and bring the estimated n_0 one order of magnitude closer toward the value predicted by the physical simulations.

Another thing worthy of emphasis is f, for which Models A and B estimate significantly higher values compared to Model C with much higher coverage rates of HPDIs. Simulations, however, result in even higher f values, especially f_2 (see for example Figure 4 of Mocz et al. 2017). This can be explained by the bias of Zeeman measurements due to $B_{\rm los}$ reversals within a telescope beam. Field reversals are considered minor in Crutcher et al. (2010), so is the bias of f. However, recent interferometer polarimetry shows that field morphologies can be quite complicated in cloud cores (Hull et al. 2014; Zhang et al. 2014) even when the mean core fields are aligned with the cloud fields (Li et al. 2009). To explain this, Zhang et al. (2019) performed numerical simulations to show that cloud cores are always super-Alfvenic, which explains Zhang et al. (2014) and Hull et al. (2014), even when the cloud as a whole is sub-Alfvenic, which explains Li et al. (2009). The fact that cores are super-Alfvenic may explain the difference between the observed and simulated f, especially f_2 as none of the Bayesian models assumes the possibility of B_z reversal within a telescope beam.

Also note that all the simulations mentioned above achieve $\alpha \approx 2/3$ at high densities, even for a sub-Alfvenic cloud with a magnetic criticality of merely two (Zhang et al. 2019). This means that $\alpha \approx 2/3$ is not necessarily a signature of "weak field," which is only true when the total mass is fixed. The cores formed in these simulations kept on accreting from the envelopes, which allows the core mass to grow with the increasing magnetic critical density due to the change of the field morphology after gravitational compression.

5. Conclusion and Future Work

In this paper, we revisited the Zeeman data set for revealing the relationship between the total field strength and the volume density of interstellar clouds with uncertain quantification. We extended the model presented in Crutcher et al. (2010) from three aspects, and showed that the extended model (Model B) is much better for fitting the Zeeman data set, when the uncertainty of number density is unknown to us. Our estimate (posterior median) of (f, α, B_0, n_0, R), (0.26,

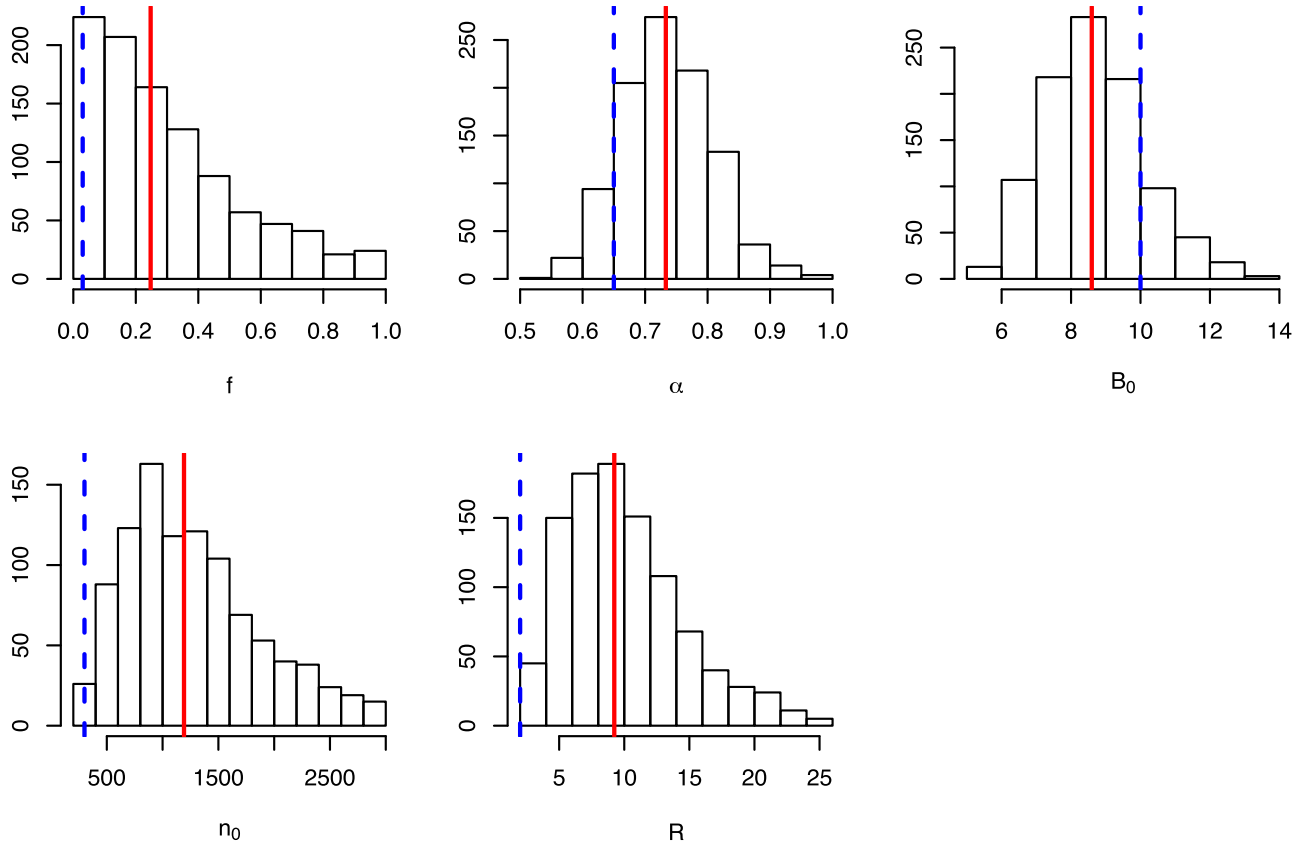

Figure 5. The histogram of posterior samples of parameters in Model B based on Dataset2. The blue dashed line in each sub-plot denotes the estimated value of the corresponding parameter from Crutcher et al. (2010), and the red solid line corresponds to the posterior median from our algorithm.

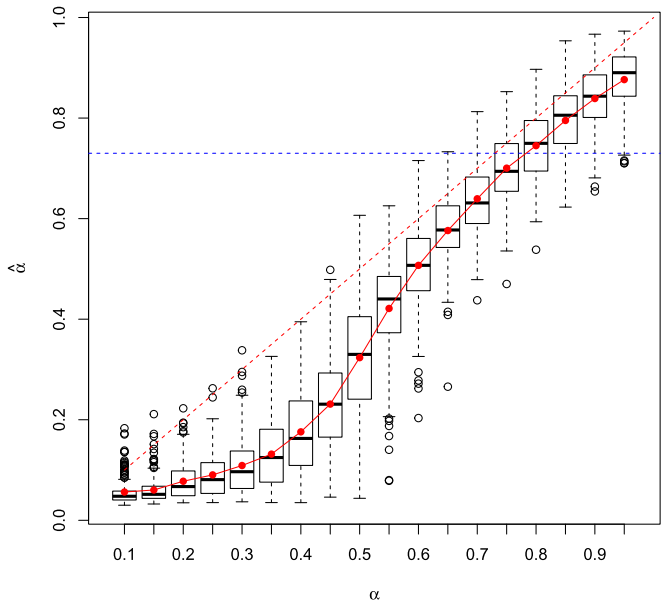

Figure 6. Posterior median estimates vs. the true underlying value of α. For each true α value on the horizontal axis, 200 data sets are synthesized and used to fit Model B. The 200 posterior median estimates $\hat{\alpha}$ for the same underlying true α are presented in a boxplot. The red dotted line is the diagonal line, and the red solid line marks the mean value of the posterior estimates for each true α. The blue horizontal dotted line corresponds to $\hat{\alpha} = 0.72$ from Model B.

0.72, 8.3, 1125, 9.3), is significantly different from that given in Crutcher et al. (2010), (0.03, 0.65, 10, 300, 2), as shown in Figure 5.

Comparing with Model C, the model in Crutcher et al. (2010), our Model B, as shown in Figure 3, provides much more reliable estimates on f, α, and n_0 by taking R as a parameter to be estimated from the data, instead of fixed at two as in Model C. The improvement of Model B on estimating α is not enough, since the coverage rate of the 95% HPDI is only around 60%. The difficulty of estimating α raises from the errors-in-variables model. Note that this problem of errors-in-variables models does not originate from the Bayesian approach. Rather, it is an essential difficulty for inferring from such models that the frequentist approaches also have. In summary, we should be more careful on drawing a conclusion from data sets with errors in both variables.

In Models A, B, and C, the total magnetic field C_i, following Crutcher et al. (2010), is modeled as uniformly distributed in the interval (fM_i , M_i), where M_i is the maximum magnetic field. As pointed out in Section 4.2, the turbulence Alfvén Mach number grows with densities (Zhang et al. 2019), so should be the chance for "B_z reversal" within a telescope beam. Ignoring this effect will bias f toward lower values. Moreover, the model for C_i can be more informative than the currently used uniform distribution if more knowledge on the relationship between the total magnetic filed and maximum magnetic field strength is available. In addition, the model for the observed number density can also be tuned if certain observation process demands.

We would like to thank Professor Richard M. Crutcher and Professor Benjamin Wandelt for helpful discussion on their results, and thank Professor Chi Tim Ng for checking the computing part. This research is partially supported by grants from the Research

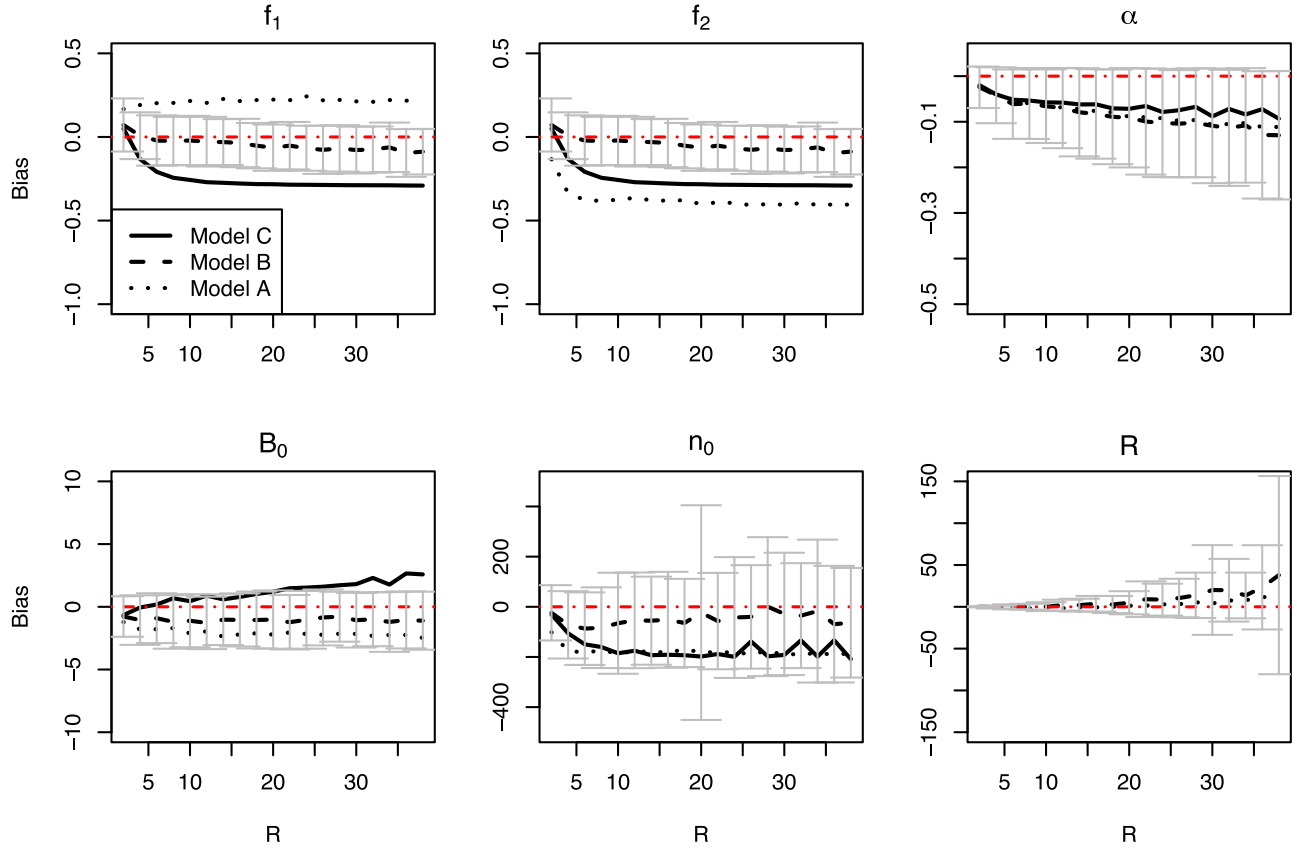

Figure 7. Bias analysis of parameters in Models A, B, and C under different uncertainty of number density when the true value of (f_1, f_2, α, B_0, n_0) is (0.03, 0.03, 0.65, 10, 300). Bias is defined as $E\hat{\theta} - \theta_0$, where $\hat{\theta}$ is the estimated value, θ_0 is the true value, and the expectation is taken over the data. Note that in Model C and Model B, $f_1 = f_2 = f$. For clear presentation, we only show the error bar (mean $\pm$ sd) of the estimated bias of parameters in Model B, since the error bar for Models A and C is similar to that for Model B.

Grants Council of the Hong Kong Special Administrative Region, China (Theme-based Research Scheme T12-402/13N; General Research Fund CUHK14173817, 14203915, 14600915 and 14304616), and The Fundamental Research Funds for the Central Universities (1A3000∗172210192).

Appendix

A.1. Posterior Distribution

The essence of Bayesian inference is to estimate parameters by combining the knowledge built in prior distributions and the evidence from the data incorporated in the likelihood function. In this section, we present mathematical derivation of the posterior distribution $P(H, \theta|D)$ and $P(\theta|D)$ for Model A, since the formula for Models B and C is a special case of that for Model A.

$$\begin{aligned} P(H, \theta|D) &= P(A_1, \cdots, A_m, n_1, \cdots, n_m, C_1, \cdots, \\ &\quad C_m, \theta|B_1, \cdots, B_m, \hat{n}_1, \cdots, \hat{n}_m) \\ &\propto \Pi_{i=1}^{n} P(B_i|A_i)P(\hat{n}_i|n_i)P(A_i|\theta, n_i, C_i) \\ &\quad \times P(C_i|\theta, n_i)P(n_i) \times P(\theta), \end{aligned}$$

where

1. $P(A_i|\theta, n_i, C_i) = \frac{I(|A_i| \leqslant C_i)}{2C_i}$,
2. $P(C_i|\theta, n_i) = h(C_i|f_1, B_0)I(n_i \leqslant n_0) + h(C_i|f_2, M_i)$ $I(n_i > n_0)$,
3. $h(x|y, z) \equiv \delta_z(x)I(y = 1) + \frac{I(yz \leqslant x \leqslant z)}{(1-y)z}I(0 \leqslant y < 1)$.

Since H is unobservable, we integrate it out from the posterior distribution. Hence, the posterior distribution of θ becomes

$$\begin{aligned} P(\theta|D) &= \int P(H, \theta|D)dH \\ &\propto \Pi_{i=1}^{m}\left\{\int\int P(B_i|A_i)P(A_i|\theta, n_i)dA_iP(\hat{n}_i|n_i)P(n_i)dn_i\right\}P(\theta) \\ &= \Pi_{i=1}^{m}\Bigg\{\underbrace{\int\int P(B_i|A_i)dA_iP(\hat{n}_i|n_i)P(n_i)F_1dn_i}_{I_{i1}} \\ &\quad + \underbrace{\int\int P(B_i|A_i)dA_iP(\hat{n}_i|n_i)P(n_i)F_2dn_i}_{I_{i2}}\Bigg\}P(\theta) \\ &= \Pi_{i=1}^{m}\{I_{i1} + I_{i2}\}P(\theta) \end{aligned} \tag{1}$$

where

$$\begin{aligned} P(A_i|\theta, n_i) &= \underbrace{\int P(A_i|\theta, n_i, C_i)P(C_i|\theta, n_i)I(n_i \leqslant n_0)dC_i}_{F_1} \\ &\quad + \underbrace{\int P(A_i|\theta, n_i, C_i)P(C_i|\theta, n_i)I(n_i > n_0)dC_i}_{F_2}. \end{aligned}$$

Let $f_{1i} = \max\{f_1, |A_i|/B_0\}$, we have

$$F_1 = \begin{cases} \dfrac{I(|A_i| \leqslant B_0)}{2B_0} I(n_i \leqslant n_0) & \text{if } f_1 = 1 \\ \dfrac{I(n_i \leqslant n_0) I(|A_i| \leqslant B_0)}{2(1-f_1)B_0} \ln(1/f_{1i}) & \text{if } 0 \leqslant f_1 < 1 \end{cases}.$$

Let $R_n = \min\{R^2, Rn_0/\hat{n}_i\}$, we have

$$I_{i1} = \begin{cases} \dfrac{\Phi\left(\frac{B_0-B_i}{\sigma_i}\right) - \Phi\left(\frac{-B_0-B_i}{\sigma_i}\right)}{4B_0\hat{n}_i \ln R} \ln(R_n) I(n_0 > \hat{n}_i/R) & \text{if } f_1 = 1 \\ \dfrac{\Phi\left(\frac{B_0-B_i}{\sigma_i}\right) - \Phi\left(\frac{-B_0-B_i}{\sigma_i}\right)}{4(1-f_1)B_0\hat{n}_i \ln R} \ln(R_n) I(n_0 > \hat{n}_i/R) E \ln(1/f_{1i}) & \text{if } 0 \leqslant f_1 < 1 \end{cases}$$

where the expectation is taken over the density of $A_i I(|A_i| \leqslant B_0)$ with $A_i \sim N(B_i, \sigma_i^2)$.

Furthermore, if $R = 1$, we have

$$I_{i1} = \begin{cases} \dfrac{\Phi\left(\frac{B_0-B_i}{\sigma_i}\right) - \Phi\left(\frac{-B_0-B_i}{\sigma_i}\right)}{2B_0} I(\hat{n}_i \leqslant n_0) & \text{if } f_1 = 1 \\ \dfrac{\Phi\left(\frac{B_0-B_i}{\sigma_i}\right) - \Phi\left(\frac{-B_0-B_i}{\sigma_i}\right)}{2(1-f_1)B_0} I(\hat{n}_i \leqslant n_0) E \ln(1/f_{1i}) & \text{if } 0 \leqslant f_1 < 1 \end{cases}.$$

Let $f_{2i} = \max\{f_2, |A_i|/M_i\}$, we have

$$F_2 = \begin{cases} \dfrac{I(|A_i| \leqslant M_i)}{2M_i} I(n_i > n_0) & \text{if } f_2 = 1 \\ \dfrac{I(n_i > n_0) I(|A_i| \leqslant M_i)}{2(1-f_2)M_i} \ln(1/f_{2i}) & \text{if } 0 \leqslant f_2 < 1 \end{cases}.$$

Let $L = \max\left\{\frac{\hat{n}_i}{R}, n_0\right\}$, $L_1 = \max\{L, n_0(|A_i|/B_0)^{1/\alpha}\}$, $U_1 = R\hat{n}_i$, $L_2 = \max\left\{L, n_0\left(\frac{|A_i|}{f_2 B_0}\right)^{1/\alpha}\right\}$, $U_2 = \min\left\{U_1, n_0\left(\frac{|A_i|}{f_2 B_0}\right)^{1/\alpha}\right\}$, we have

$$I_{i2} = \begin{cases} p_0 u_0 I(n_0 < R\hat{n}_i) E_0[L_1^{-\alpha} - U_1^{-\alpha}] & \text{if } f_2 = 1 \\ \dfrac{u_0 I(n_0 < R\hat{n}_i)}{(1-f_2)} \left\{ p_1 E_1[u_1(L_2^{-\alpha} - U_1^{-\alpha})] + p_2 E_2\left[u_2(L_1^{-\alpha} - U_2^{-\alpha}) + \alpha\left(\frac{\ln L_1}{L_1^{\alpha}} - \frac{\ln U_2}{U_2^{\alpha}}\right)\right]\right\} & \text{if } 0 < f_2 < 1 \\ p_0 u_0 I(n_0 < R\hat{n}_i) E_0\left[u_2(L_1^{-\alpha} - U_1^{-\alpha}) + \alpha\left(\frac{\ln L_1}{L_1^{\alpha}} - \frac{\ln U_1}{U_1^{\alpha}}\right)\right] & \text{if } f_2 = 0 \end{cases}$$

where $u_0 = \frac{n_0^{\alpha}}{4\alpha B_0 \hat{n}_i \ln R}$, $u_1 = \ln(1/f_2)$, $u_2 = \ln\left(\frac{B_0}{n_0^{\alpha} |A_i|}\right) + 1$, $p_0 = \Phi\left(\frac{\hat{M}_i - B_i}{\sigma_i}\right) - \Phi\left(\frac{-\hat{M}_i - B_i}{\sigma_i}\right)$, $p_1 = \Phi\left(\frac{f_2\hat{M}_i - B_i}{\sigma_i}\right) - \Phi\left(\frac{-f_2\hat{M}_i - B_i}{\sigma_i}\right)$, and $p_2 = P(|A_i| \leqslant \hat{M}_i) - P(|A_i| \leqslant f_2\hat{M}_i^L)$. The expectation E_0 is taken over the density of $A_i I(|A_i| \leqslant \hat{M}_i)$ with $A_i \sim N(B_i, \sigma_i^2)$ and $\hat{M}_i = B_0\left(\frac{U_1}{n_0}\right)^{\alpha}$. The expectation E_1 is taken over the density of $A_i I(|A_i| \leqslant f_2\hat{M}_i)$. The expectation E_2 is taken over density of $A_i I(f_2\hat{M}_i^L < |A_i| < \hat{M}_i)$ with $\hat{M}_i^L = B_0\left(\frac{L}{n_0}\right)^{\alpha}$.

Furthermore, if $R = 1$, we have

$$I_{i2} = \begin{cases} \dfrac{n_0^{\alpha} I(\hat{n}_i > n_0)}{2B_0\hat{n}_i^{\alpha}} \left[\Phi\left(\frac{M_i - B_i}{\sigma_i}\right) - \Phi\left(\frac{-M_i - B_i}{\sigma_i}\right)\right] & \text{if } f_2 = 1 \\ \dfrac{n_0^{\alpha} I(\hat{n}_i > n_0)}{2(1-f_2)B_0\hat{n}_i^{\alpha}} \left[\Phi\left(\frac{M_i - B_i}{\sigma_i}\right) - \Phi\left(\frac{-M_i - B_i}{\sigma_i}\right)\right] E_0 \ln(1/f_{2i}) & \text{if } 0 \leqslant f_2 < 1 \end{cases}.$$

A.2. Analysis of the Linear Errors-in-variables Model

In this section, we explain theoretically the difficulty in estimating parameters for errors-in-variables models through a similar but much simpler, thus analytically approachable, errors-in-variables model. The notations in this section are independent from those in the other sections.

Assume that we have observations (x_i, y_i), $i = 1, \cdots, n$, from the following model:

$$x_i = z_i + \eta_i,\ P(\eta_i) = G(\eta_i|0, \sigma_0);$$
$$y_i = z_i\beta + \epsilon_i,\ P(\epsilon_i) = G(\epsilon_i|0, \sigma_1);$$

where z_i is unobservable, σ_0 and σ_1 are known constants. In this model, x_i is the measurement of z_i with a Gaussian error η_i, and y_i follows a linear regression model with regard to the unobserved z_i. The goal here is to estimate the unknown parameter β based on the observations (x_i, y_i), $i = 1, \cdots, n$. We shall demonstrate that, as long as there is error in the measurement x_i, i.e., $\sigma_0 > 0$, the parameter β cannot be estimated consistently. For this purpose, we need an assumption that the observations (x_i, y_i) and unobserved z_i $(i = 1, \cdots, n)$ are bounded. There are basically two

approaches to obtain the estimate: the frequentist approach and Bayesian approach.

1. Frequentist approach—We first consider the frequentist approach to perform a simple linear regression of y_i on x_i. The resulting estimator is

$$\hat{\beta} \equiv \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n}(z_i^2\beta + z_i\beta\eta_i + z_i\epsilon_i + \epsilon_i\eta_i)}{\sum_{i=1}^{n}(z_i^2 + 2z_i\eta_i + \eta_i^2)}.$$

By the law of large numbers, $\frac{1}{n}\sum_{i=1}^{n}\eta_i^2 \to \sigma_0^2$, $\frac{1}{n}\sum_{i=1}^{n} z_i\eta_i \to 0$, $\frac{1}{n}\sum_{i=1}^{n} z_i\epsilon_i \to 0$, and $\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\eta_i \to 0$ as $n \to +\infty$, Thus, the estimator $\hat{\beta} \to \frac{\beta}{1+\sigma_0^2/\sigma_z^2}$, where $\sigma_z^2 = \lim_{n\to+\infty}\sum_{i=1}^{n} z_i^2/n$, and we assume the limitation $\sigma_z^2 < +\infty$. Thus, the linear regression estimate is biased by a multiplicative factor of $1/(1 + \sigma_0^2/\sigma_z^2)$ when $\sigma_0 > 0$.

2. Bayesian approach–The second estimation method is the Bayesian approach. We now illustrate that the Bayesian analysis also produces a biased estimate asymptotically. First, we assume a typical prior distribution for z_i and β as follows.

1. $P(z_i) = G(z_i|\mu_i, \sigma)$, where μ_i and σ are given constants, $i = 1, \cdots, n$;
2. $P(\beta) = G(\beta|\mu_\beta, \sigma_\beta)$, where μ_β and σ_β are given constants.

Then, the posterior density of β can be derived as:

$$\begin{aligned}
&P(\beta|y_1, y_2, \cdots, y_n, x_1, x_2, \cdots, x_n)\\
&\propto P(y_1, y_2, \cdots, y_n, x_1, x_2, \cdots, x_n|\beta)P(\beta)\\
&= P(\beta)\int P(y_1, y_2, \cdots, y_n, x_1, x_2, \cdots, x_n, z_1, z_2, \cdots, z_n|\beta)\\
&\quad \times dz_1 dz_2 \cdots dz_n\\
&= P(\beta)\prod_{i=1}^{n}\int \frac{1}{\beta}\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(z_i-v_i)^2}{2\sigma_2^2}}\\
&\quad \times \frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{(z_i-x_i)^2}{2\sigma_0^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(z_i-\mu_i)^2}{2\sigma^2}}dz_i,
\end{aligned}$$

where $\sigma_2 = \sigma_1/\beta$, $v_i = y_i/\beta$.

To simplify the above equation, we introduce the following lemma, whose proof shall be straightforward.

Lemma 1. *Let* $G(x|\lambda, \tau) \equiv \frac{1}{\sqrt{2\pi}\tau}e^{-\frac{(x-\lambda)^2}{2\tau^2}}$, *then we have:*

$$G(x|\lambda_1, \tau_1)G(x|\lambda_2, \tau_2) = S_{12}G(x|\lambda_{12}, \tau_{12}),$$

where $\lambda_{12} = \frac{\lambda_1\tau_2^2 + \lambda_2\tau_1^2}{\tau_1^2+\tau_2^2}$, $\tau_{12} = \frac{\tau_1\tau_2}{\sqrt{\tau_1^2+\tau_2^2}}$, *and* $S_{12} = \frac{1}{\sqrt{2\pi(\tau_1^2+\tau_2^2)}}e^{-\frac{(\lambda_1-\lambda_2)^2}{2(\tau_1^2+\tau_2^2)}}$.

By Lemma 1, we have:

$$\begin{aligned}
&P(\beta|y_1, y_2, \cdots, y_n, x_1, x_2, \cdots, x_n)\\
&\propto P(y_1, y_2, \cdots, y_n, x_1, x_2, \cdots, x_n|\beta)P(\beta)\\
&\propto P(\beta)\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi(\sigma_1^2 + \beta^2\sigma_{12}^2)}}e^{-\frac{(y_i-\beta\mu_{12}^i)^2}{2(\sigma_1^2+\beta^2\sigma_{12}^2)}},
\end{aligned}$$

where $\mu_{12}^i = \frac{x_i\sigma^2 + \mu_i\sigma_0^2}{\sigma^2+\sigma_0^2}$ and $\sigma_{12} = \frac{\sigma_0\sigma}{\sqrt{\sigma_0^2+\sigma^2}}$.

The final line of the above equation shows that the Bayesian method is essentially estimating β by regressing y_i on s_i based on the linear regression model $y_i = \beta s_i + w_i$, $i = 1, 2, \cdots, n$, where $P(w_i) = G(w_i|0, \sqrt{\sigma_1^2 + \beta^2\sigma_{12}^2})$, $s_i = tx_i + (1-t)\mu_i$, $t = \frac{\sigma^2}{\sigma^2+\sigma_0^2}$, and $\sigma_{12} = \frac{\sigma_0\sigma}{\sqrt{\sigma_0^2+\sigma^2}}$. Asymptotically, as $n \to +\infty$, the posterior distribution of β will converge to the same point as the following estimator:

$$\begin{aligned}
\hat{\beta}_{\rm MLE} &\equiv \arg\max_{\beta} F(\beta) \equiv \arg\max_{\beta}\prod_{i=1}^{n}\\
&\quad \times \frac{1}{\sqrt{2\pi(\sigma_1^2 + \beta^2\sigma_{12}^2)}}e^{-\frac{(y_i-\beta\mu_{12}^i)^2}{2(\sigma_1^2+\beta^2\sigma_{12}^2)}}.
\end{aligned}$$

Thus, we just need to show that $\hat{\beta}_{\rm MLE}$ will not converge to the true β value.

Denote the derivative of $\log(F(\beta))$ as $T(\beta)$. $\hat{\beta}_{\rm MLE}$ shall satisfy $T(\hat{\beta}_{\rm MLE}) = 0$. We can derive that $T(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta s_i)s_i + \frac{\beta\sigma_{12}^2}{\sigma_1^2+\beta^2\sigma_{12}^2}\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta s_i)^2 + \beta\sigma_{12}^2$.

Assume that the true value of β is $\beta_0(\neq 0)$, thus $y_i = \beta_0 z_i + \epsilon_i$ according to the model assumption. We have,

$$\begin{aligned}
T(\beta) &= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 z_i + \epsilon_i - \beta s_i)s_i + \frac{\beta\sigma_{12}^2}{\sigma_1^2 + \beta^2\sigma_{12}^2}\\
&\quad \times \frac{1}{n}\sum_{i=1}^{n}(\beta_0 z_i + \epsilon_i - \beta s_i)^2 + \beta\sigma_{12}^2.
\end{aligned}$$

As we can see, the equation $T(\beta) = 0$ is too complicated to have an analytic solution. Thus, we consider instead $T(\beta_0)$. If $T(\beta_0)$ is zero, we have $\beta_{\rm MLE} = \beta_0$. By the law of large numbers,

$$\begin{aligned}
T(\beta_0) &= \frac{1}{n}\sum_{i=1}^{n}(z_i - s_i)s_i + \frac{1}{n}\sum_{i=1}^{n}(z_i - s_i)^2\frac{\beta_0^2\sigma_{12}^2}{\sigma_1^2 + \beta_0^2\sigma_{12}^2}\\
&\quad + \frac{\sigma_1^2\sigma_{12}^2}{\sigma_1^2 + \beta_0^2\sigma_{12}^2} + \sigma_{12}^2 + o_p(1),
\end{aligned} \tag{2}$$

where $o_p(1)$ denotes a term that will converge in probability to zero as $n \to +\infty$.

First, we note that $T(\beta_0) = \frac{\sigma_1^2\sigma_{12}^2}{\sigma_1^2+\beta_0^2\sigma_{12}^2} + \sigma_{12}^2 + o_p(1) > 0$, if we have the best guess of μ_i such that $s_i = z_i$. In other words, $\beta_{\rm MLE}$ is a biased estimate of β_0 under this best guess. In addition, if there is no measurement error ($\sigma_0 \Rightarrow \sigma_{12} = 0$ and $s_i = z_i$), $T(\beta_0) = o_p(1)$, that is, $\hat{\beta}_{\rm MLE}$ is unbiased. Second, one may specify $\mu_i = x_i$, which gives $s_i = x_i$. In this case, $T(\beta_0) = o_p(1)$ only if $\beta_0 = \sqrt{\frac{\sigma_1^2}{\sigma_{12}^2}[\frac{\sigma_0^2}{\sigma_{12}^2} - 1 - \frac{1}{\sigma_{12}^2}]}$ and $\sigma_0^2 > 1 + \sigma_{12}^2$. Thus, $\beta_{\rm MLE}$ is a biased estimate of β_0 for this case, since $P(\beta_0 = \sqrt{\frac{\sigma_1^2}{\sigma_{12}^2}[\frac{\sigma_0^2}{\sigma_{12}^2} - 1 - \frac{1}{\sigma_{12}^2}]}) = 0$. Third, note that $T(\beta_0)$ is a function of $\mu = (\mu_1, \cdots, \mu_n) \in R^n$, we rewrite $T(\beta_0)$ as $T(\mu|\beta_0)$. Let $A_{\beta_0} = \{\mu \in R^n: T(\mu|\beta_0) = o_p(1)\}$ be the set of μ on which $\beta_{\rm MLE} \approx \beta_0$, i.e., $\beta_{\rm MLE}$ is an unbiased estimate of β_0. The number of points belonging to set A_{β_0} depends on unknown $z = (z_1, \cdots, z_n)$ and unknown β_0, and it can be finite or infinite. However, we specify μ at random, and for any point $a \in A_{\beta_0}$, we have $P(\mu = a) = 0$ when μ is viewed as a random vector. In other words, we cannot guess exactly what value of μ such that $T(\mu|\beta_0) = o_p(1)$. In summary, $\beta_{\rm MLE}$ is usually a biased estimate of β_0. We cannot recover the true parameter value even if we have infinite number of observations.

A.3. The Effect of Extra Noisy Data

In this section, we analyze the effect of extra noisy samples on estimating the parameter in a linear model, and show the condition under which these extra noisy samples are helpful. The following analysis is based on the frequentist approach for the ease of presentation, but a similar conclusion can be drawn from the Bayesian approach.

Assume that the original data set $\{(x_i, y_i): i = 1, 2, \cdots, n\}$ are from $y_i = x_i\beta + \epsilon_i$ for $i = 1,2, \cdots, n$, and the extra data set $\{(x_i, y_i): i = n + 1, n + 2, \cdots, n + m\}$ are from $y_i = x_i\beta + \eta_i$ for $i = n + 1, \cdots, n + m$, where $P(\epsilon_i) = G(\epsilon_i|0, \sigma_1)$, $P(\eta_i) = G(\eta_i|0, \sigma_2)$.

Similar to the simple linear regression, β based on the whole data set $\{(x_i, y_i): i = 1, 2, \cdots, n + m\}$ can be estimated by

$$\hat{\beta}_{n+m} = \frac{\sum_{i=1}^{n+m} y_i x_i}{\sum_{i=1}^{n+m} x_i^2} = \beta_0 + Z_{n+m},$$

$$P(Z_{n+m}) = G\left(Z_{n+m}|0, \sqrt{\frac{S_1\sigma_1^2 + S_2\sigma_2^2}{(S_1 + S_2)^2}}\right),$$

where β_0 is the true value of β, $S_1 = \sum_{i=1}^{n} x_i^2$, and $S_2 = \sum_{i=n+1}^{n+m} x_i^2$.

Setting $m = 0$, we get the estimate of β based on the original data set,

$$\hat{\beta}_n = \beta_0 + Z_n, \quad P(Z_n) = G\left(Z_n|0, \sqrt{\frac{\sigma_1^2}{S_1}}\right).$$

Both estimator are unbiased regardless of adding the extra noisy data or not. The extra data is helpful for estimating β only if the estimate based on the whole data set has a smaller uncertainty than that based on the original data set. That is, the following relationship holds:

$$\mathrm{Var}(\hat{\beta}_{n+m}) = \frac{S_1\sigma_1^2 + S_2\sigma_2^2}{(S_1 + S_2)^2} < \mathrm{Var}(\hat{\beta}_n) = \frac{\sigma_1^2}{S_1}.$$

Thus, if and only if the number of extra observations m and the corresponding uncertainty σ_2^2 satisfy

$$\frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2}{S_1} + 2, \tag{3}$$

the extra observations are helpful for estimating β.

Note that $S_1/n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \bar{x}^2$, and similarly, $S_2/m = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2 + \bar{x}^2$. Since $\{x_1, \cdots, x_n\}$ and $\{x_{n+1}, \cdots, x_{n+m}\}$ are usually from the same distribution, $S_1/n \simeq S_2/m$ when n and m are large enough. Thus the inequality (2) implies that if the uncertainty of the extra data satisfies $\sigma_2^2 < (m/n + 2)\sigma_1^2$, the extra data is helpful. Thus, if m is much smaller than n such that $\frac{S_2}{S_1} \simeq 0$, the extra data will be helpful if $\sigma_2^2 < 2\sigma_1^2$. On the other hand, if m is quite large as compared with n, the additional data will always be helpful even if they are noisy.

ORCID iDs

Hua-bai Li https://orcid.org/0000-0003-2641-9240
Xiaodan Fan https://orcid.org/0000-0002-2744-9030

References

Brooks, S. P., & Gelman, A. 1998, J. Comput. Graph. Stat., 7, 434
Cook, S. R., Gelman, A., & Rubin, D. B. 2006, J. Comput. Graph. Stat., 15, 675
Crutcher, R. M. 2012, ARA&A, 50, 29
Crutcher, R. M., Hakobian, N., & Troland, T. 2009, ApJ, 692, 844
Crutcher, R. M., Wandelt, B., Heiles, C., Falgarone, E., & Troland, T. H. 2010, ApJ, 725, 466
Frost, C., & Thompson, S. G. 2002, J. R Stat. Soc. A, 163, 173
Hsiao, C. 1989, J. Econ., 41, 159
Hull, C. L., Plambeck, R. L., Kwon, W., et al. 2014, ApJS, 213, 13
Kong, F. H., & Gu, M. 1999, Stat. Sinica, 9, 953
Li, D., Dickey, J. M., & Liu, S. 2019, RAA, 19, 1904
Li, D., & Pan, Z. 2016, RaSc, 51, 1060
Li, H.-b., Dowell, C. D., Goodman, A., Hildebrand, R., & Novak, G. 2009, ApJ, 704, 891
Li, H.-b., Fang, M., Henning, T., & Kainulainen, J. 2013, MNRAS, 436, 3707
Li, H.-b., Goodman, A., Sridharan, T. K., et al. 2014, in Protostars and Planets VI, ed. H. Beuther et al. (Tucson: Univ. Arizona Press), 101
Li, P. S., McKee, C. F., & Klein, R. I. 2015, MNRAS, 452, 2500
Li, T. 2002, J. Econ., 110, 1
McKee, C., & Ostriker, E. 2007, ARA&A, 45, 565
Mocz, P., Burkhart, B., Hernquist, L., Mckee, C., & Springel, V. 2017, ApJ, 838, 40
Mouschovias, T., & Tassis, K. 2010, MNRAS, 409, 801
Robert, C., & Casella, G. 2004, Monte Carlo Statistical Methods (New York: Springer)
Schwarz, G. E. 1978, AnSta, 6, 461
Shirley, Y. L. 2015, PASP, 127, 299
Tritsis, A., Panopoulou, G. V., Mouschovias, T. Ch., Tassis, K., & Pavlidou, V. 2015, MNRAS, 451, 4384
Zhang, Q., Qiu, K., Girart, J. M., et al. 2014, ApJ, 792, 116
Zhang, Y., Guo, Z., Wang, H. H., & Li, H.-b. 2019, ApJ, 871, 98