CROWDSOURCED PAIRWISE-COMPARISON FOR SOURCE SEPARATION EVALUATION

Mark Cartwright[1], *Bryan Pardo*[2], *Gautham J. Mysore*[3]

[1] New York University, USA
[2] Northwestern University, USA
[3] Adobe Research, USA

ABSTRACT

Automated objective methods of audio source separation evaluation are fast, cheap, and require little effort by the investigator. However, their output often correlates poorly with human quality assessments and typically require ground-truth (perfectly separated) signals to evaluate algorithm performance. Subjective multi-stimulus human ratings (e.g. MUSHRA) of audio quality are the gold standard for many tasks, but they are slow and require a great deal of effort to recruit participants and run listening tests. Recent work has shown that a crowdsourced multi-stimulus listening test can have results comparable to lab-based multi-stimulus tests. While these results are encouraging, MUSHRA multi-stimulus tests are limited to evaluating 12 or fewer stimuli, and they require ground-truth stimuli for reference. In this work, we evaluate a web-based pairwise-comparison listening approach that promises to speed and facilitate conducting listening tests, while also addressing some of the shortcomings of multi-stimulus tests. Using audio source separation quality as our evaluation task, we compare our web-based pairwise-comparison listening test to both web-based and lab-based multi-stimulus tests. We find that pairwise-comparison listening tests perform comparably to multi-stimulus tests, but without many of their shortcomings.

Index Terms— audio quality evaluation, crowdsourcing, source separation

1. INTRODUCTION

In recent years there has been increased dissatisfaction with the existing automated objective metrics that are typically used to evaluate audio source separation algorithms [1, 2, 3, 4]. It is difficult to make informed decisions to improve source separation algorithms when researchers have shown that metrics such as BSS-Eval [5] and PEASS [6] poorly correlate with human perception of quality [1, 3]. Also, ground-truth signals are required by the most popular objective metrics. These may not be available in many cases. An alternative to automated objective metrics is to run subjective listening tests. However, lab-based subjective listening tests such as MUSHRA [7] typically require at least 20 participants, and recruiting these participants and administering tests can easily consume a week of a researcher's time. Researchers have sought to reduce the time and effort to run listening tests by moving them from the lab to the web [8, 9, 10, 11], but experimental testing is needed to determine how this change in environment affects quality judgments.

Recently, we compared MUSHRA performed in a controlled lab environment (*lab-MS*) to a MUSHRA-like[1] multi-stimulus test performed in an uncontrolled web environment on a population drawn from Mechanical Turk (*web-MS*). In that study, we collected crowdsourced data from over 500 participants in less than 9 hours—a feat that would be difficult to accomplish with a lab-based study. We also showed that web-MS perceptual evaluation scores are comparable to those estimated in the controlled lab environment. However, while crowdsourced MUSHRA-like tests can be a proxy for lab-based MUSHRA tests in some situations, these tests recommend to limit the number of stimuli to only 12 audio signals [7]. This is a limitation when evaluating many source separation algorithms in campaigns such as SISEC [12].

Pairwise-comparison tests have a long history in psychometric testing and are well studied [13]. There is also a history of pairwise-comparison listening tests in audio (e.g. ABX or ITU-R BS.1116-2 [14]). Pairwise-comparison tests are discriminative by design and require participants to attend to fewer stimuli simultaneously than multi-stimulus tests—a characteristic that could be beneficial if recruiting novice participants from crowdsourcing platforms. In pairwise-comparison tests, we can also measure participant reliability by simply observing transitivity in their comparisons [15, 10], whereas MUSHRA/MUSHRA-like tests require repeated trials. Consistency measures are useful for incentivizing workers with consistency-based rewards. Therefore, pairwise-comparison tests may be more suitable for large-scale, crowdsourced quality evaluations.

This work was supported, in part, by USA National Science Foundation Award 1420971

[1] Some of the MUSHRA recommendations are not feasible on the web (e.g. playback system specifications and the requirement for expert users) [1]. However, the multi-stimulus test we implemented still shares many characteristics of MUSHRA such as the inclusion of hidden reference and anchor stimuli. Therefore, we refer to our multi-stimulus tests as "MUSHRA-like".

There are variants of pairwise-comparison tests that can estimate quality scores with or without ground-truth stimuli for reference [16]. This allows evaluation to be performed on any kind of real-world data. Also, there is no limit on the number of stimuli that can be evaluated. When an additional stimulus is compared, individual tasks are not more taxing on the participant—we need more pairwise comparisons per trial, but these comparisons can be distributed among multiple participants. Techniques have also recently been developed to further reduce the number of comparisons [17, 18].

Given the potential advantages of pairwise evaluation of audio over the web, in this work we seek to establish how results from pairwise-comparison listening tests conducted in a crowdsourced setting compare to gold-standard results collected from a multi-stimulus listening test in a lab setting.

2. METHODS

2.1. Baseline Data Set

As in our previous study [1], we compare our quality scores to the results of a lab-based listening test that followed the MUSHRA recommendation and was conducted by the developers of PEASS [6] to train the PEASS objective scoring models. We refer to this lab-based multi-stimulus test as *lab-MS*, and our previous web-based, MUSHRA-like multi-stimulus test as *web-MS*. In our current study of pairwise-comparison listening tests, we use the same test audio as these previous tests, and we treat the lab-MS results as the gold-standard baseline. This test material consists of 10 mixtures (5 speech, 5 music), each 5 seconds long and containing 2–7 sources. Two speech mixtures are mono, and the remaining mixtures are stereo. For each mixture, there are 8 test stimuli: the ground-truth target source (the reference), 3 anchors, and 4 outputs of a variety of source separation algorithms. The lab-MS data was collected in a lab setting from 20 normal-hearing participants who were experts in general audio applications. Each participant performed a MUSHRA trial for each of the 10 sets of test stimuli for 4 different quality scales: *overall quality*, *preservation of the target source*, *suppression of other sources*, and *absence of additional artificial noises*.

2.2. Listening Test Procedure

In a pairwise-comparison listening test, a participant is asked to choose which of two test stimuli [2] is higher on a given quality scale, e.g. "In which recording are the drums louder? Recording A or B?". A participant may answer the same question for several different pairs of stimuli, and each pair will be evaluated by several different participants. Using these pairwise preference decisions, a stimulus preference order and (possibly) absolute stimulus scores may be estimated.

To perform pairwise-comparison listening tests on the web, we recruited participants from Amazon's Mechanical Turk in the same manner as the web multi-stimulus experiment (web-MS)[1]. We also assigned tasks in the same manner as web-MS, assigning participants to one quality scale, and allowing each participant to perform up to 10 randomly-ordered trials. However, for each task, instead of performing a single multi-stimulus trial, participants performed a pairwise-comparison trial, comparing all pairs of stimuli associated with a mixture—$\binom{8}{2}$, or 28, pairwise comparisons in random order (i.e. 1 trial=28 pairwise comparisons). We limited each participant to one quality scale to reduce task confusion and to eliminate any quality scale ordering effects.

Participants completed the pairwise-comparison test using our Crowdsourced Audio Quality Evaluation (CAQE) software[3]. Using CAQE's web-interface, when a participant first selects a stimulus (A, B, mixture, or reference), looped playback of the selected stimulus begins. When a participant selects a subsequent stimulus, the playback loop synchronously switches to the selected stimulus, maintaining the current playback location. A participant can only proceed to the next comparison after they have listened to the stimuli for five seconds and have selected either the A or B stimulus. As in the web-MS training, participants were required to listen to examples of reference and anchor stimuli to familiarize themselves with the quality scales. All instructions were kept as similar to the web-MS experiment as possible.

As in the web-MS listening test [1], we collected *a minimum of* 20 pairwise-comparison trials for each condition (mixture / quality scale pair). However, we limited the analysis in this paper to the data from the first 20 participants to fairly compare to the lab-MS and web-MS data. We paid participants \$0.80 for completing the first trial, which included a hearing evaluation [1], and \$0.50 for subsequent trials. In addition, participants could receive up to a \$0.25 bonus based on the consistency of their comparisons. Only participants with at least 1000 approved Mechanical Turk assignments and a 97% approval rate were recruited. It took 35.5 hours to collect all of the pairwise data. This is longer than the web-MS data collection time (8.2 hours), and we suspect this is due to an imbalance between task length and task reward[19]. We could potentially reduce the completion time by either increasing the task reward to appropriately match the length of the task or also by simply allowing participants to complete trials of more than one quality scale.

We also introduced a fifth quality scale in this study. In both the web-MS and lab-MS studies, participants seem to confuse the *preservation of the target source* and *absence of additional artificial noises* scales—it's as if the participants could not distinguish between additive artifacts (*absence of additional artificial noises*) and subtractive artifacts (*preservation of the target source*). Such confusions can lead to inconsistencies in preference ratings which affect score estima-

[2] In some studies one or more reference stimuli are also provided.

[3] `https://github.com/interactiveaudiolab/CAQE`

tion. Therefore, in this study we introduced a *lack of distortions to the target source*, a scale which is inclusive of both additive and subtractive artifacts.

2.3. Quality Score Estimation

To estimate the latent quality scores from the pairwise-comparison data, we use a Thurstone model. A Thurstone model is a probabilistic latent variable model that maps discrete preference orderings (e.g. pairwise comparisons) of N items $(a_1, a_2, \ldots, a_N)$ to latent scores $(\mu_1, \mu_2, \ldots, \mu_N)$ on an interval scale [20]. The model assumes that the items can be assigned values on this scale with some measurement error, and therefore the model treats the scale values as random variables $(S_1, S_2, \ldots, S_N)$. The distance between two items on the unobserved scale and the measurement error affect the pairwise preference probabilities of the items—the larger the perceived difference between two items on the scale, the greater probability that the higher item on the scale will be preferred by a listener. In our case, the discrete preference orderings of items are the pairwise comparisons of audio stimuli as described earlier, and our interval scale is an audio quality measure. The basic Thurstone model is as follows:

$$S_n \sim \text{Normal}(\mu_n, \sigma_n^2), \text{ for } n \in 1:N \tag{1}$$

$$\Pr(a_i \succ a_j) = \Pr(S_i > S_j), \text{ for } i, j \in 1:N \; ; \; i \neq j \tag{2}$$

$$= \Pr(S_i - S_j > 0) \tag{3}$$

For simplicity and identifiability, it is common to assume that all variances σ_n^2 are equal (i.e., $\sigma_n = \sigma$). This variation is known as Thurstone Model V [20]. With this assumption, [21] showed Eq. 1 can be rewritten as:

$$\Pr(a_i \succ a_j) = \Phi\left(\frac{\mu_i - \mu_j}{\sigma\sqrt{2}}\right) \tag{4}$$

where Φ is the cumulative distribution function of the normal distribution.

Using Eq. 4, we can relate a preference probability of two audio stimuli to their latent quality scores (μ_i and μ_j) and variance (σ). However, we want to jointly estimate the scores of all audio stimuli. We can write the likelihood for a set of T pairwise comparisons (e.g., 20 participants' pairwise comparisons of the 8 stimuli associated with a mixture for a particular audio quality scale—$20 \times 28 = 560$ comparisons) as:

$$\mathcal{L}(\theta | a_{ii[t]} \succ a_{jj[t]} \forall t \in 1:T) = \prod_{t=1}^{T} \Phi\left(\frac{\mu_{ii[t]} - \mu_{jj[t]}}{\sigma\sqrt{2}}\right) \tag{5}$$

$$\mu_n \sim \text{Uniform} \in [0, 100], \text{ for } n \neq h, \; n \notin L$$
$$\mu_h \sim \text{Truncated-Normal}(100, 5) \in [0, 100]$$
$$\mu_n \sim \text{Truncated-Normal}(15, 15) \in [0, 100], \text{ for } n \in L$$
$$\sigma \sim \text{Uniform} \in [0, 100]$$

where $\theta = (\mu, \sigma)$, h is the hidden reference stimulus index, L is the set of hidden anchor stimuli indices, and $ii[t], jj[t] \in 1:N$ are the indices for the stimulus pair in comparison $t \in 1:T$. Note that all of the preference information for a comparison is in the indices, $ii[t]$ and $jj[t]$, since the preferred stimulus is always assigned to $ii[t]$.

Since the MUSHRA scale range is $[0, 100]$, we limit the range of all score variables to $[0, 100]$ and set the priors for scores of the stimuli of systems under test (all stimuli but the hidden reference and anchors) to Uniform$(0, 100)$. The priors on the hidden reference and anchor scores are set according to our expectations as well—very high for the hidden reference and relatively low for the hidden anchors. While these specific priors and limits are not necessary for estimating meaningful scores, they are necessary if we want scores scaled for direct comparison to the results of MUSHRA tests.

We fit a different Thurstone model for each quality scale and mixture (i.e., 40 fitted models in total) using the NUTS algorithm for Markov chain Monte Carlo (MCMC) sampling [22]. When fitting models, we drew two chains of 10,000 samples from the posterior distribution, dropping the first 5,000 and thinning by a factor of 2. Gelman and Rubin's potential scale reduction $\hat{R}$ is an MCMC sampling convergence diagnostic based on the within-chain and between-chain variance of two or more sampling chains [23]. It is generally accepted that chains are adequately mixed and sampling has converged when $\hat{R}$ is near 1.0 with an acceptance threshold of 1.1 [23]. The variables for all models and conditions met the $\hat{R} < 1.1$ acceptance criterion.

3. RESULTS

To establish if our web-based pairwise-comparison listening test can act as a proxy for a lab-based, gold-standard test, we calculated the Pearson correlation between the lab-MS scores and the scores estimated from the Thurstone model. The results are shown in Figure 1. For comparison, we included the scores estimated by the web-MS test and [1] and the scores calculated from the most popular automated measures of audio source separation quality: the BSS-Eval measurements (i.e. SDR, ISR, SIR, and SAR). The Thurstone model's correlations with lab-MS were comparable to those of the web-MS to lab-MS for all qualities except *preservation of the target source*, for which the correlations were lower—the null hypothesis that the correlations were equal for *preservation of the target source* was rejected by a William's t-test with Bonferroni correction ($p = 0.013$), but not for the other quality scales ($p > 0.05$).

Next, we investigated the discriminative power of the lab-MS, web-MS, and Thurstone score estimations. To investigate this, we calculated the widths of the 95% confidence intervals (CIs) for the scores of the systems under tests and aggregated over the original four quality scales. Tighter confidence intervals are preferable because of their greater statis-

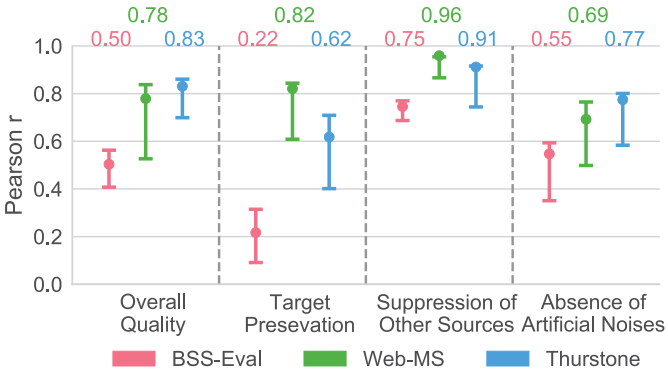

Fig. 1. Pearson correlation with the lab-MS scores and the scores estimated from the pairwise-comparison tests and models (web-MS and BSS-Eval included for comparison). Scores were limited to the systems under test (i.e. excluding the reference and anchors) and estimated using a sample size of 20 participants per mixture. Scores for all mixtures were concatenated before calculating the correlation for each quality scale ($N = 40$). Bars represent 95% CIs calculated from 1000 bootstrap iterations, randomly sampling with replacement from the lab-MS ratings and sampling from posterior distribution of the pairwise model scores.

tical power in discriminating between stimulus scores. The mean and std. deviation of the CI-widths of the lab-MS, web-MS, and Thurstone scores were respectively $mean = 22.0$ ($SD = 10.3$), $mean = 21.7$ ($SD = 7.3$), and $mean = 26.7$ ($SD = 3.6$). A one-way ANOVA rejected the null hypothesis that these means are equal ($F(2, 477) = 21.88$, $p < 0.001$). A post hoc Tukey HSD test ($\alpha = 0.05$) showed that the Thurstone scores have a statistically different CI-width mean than both lab-MS and web-MS, but the difference between lab-MS and web-MS is not significant. Therefore, while the Thurstone scores are comparably as similar to lab-MS as web-MS are to lab-MS, the distributions of the Thurstone scores are less discriminative than web-MS when using the same number of participants—more participants are likely required to get tighter CIs.

Lastly, to evaluate our proposed *lack of distortions to the target source* scale, we calculated both individual and global transitivity measures of the preference ratings. Preference ratings that obey transitivity indicate that the scale is understood by participants and unidimensional—both desirable properties. For individual transitivity, we computed the *transitivity satisfaction rate* (*TSR*) [10]—the fraction of stimulus triples in which an individual participant satisfies transitivity in their pairwise comparisons (e.g., if a participant chooses stimulus A over B, and B over C, then we expect them to choose A over C as well if their preferences satisfy transitivity). For global transitivity, we computed *weak*, *moderate*, and *strong stochastic transitivity* (*WST, MST, SST*) [10] from the empirical pairwise preference probabilities for a population. Let $\hat{P}_{ij}$ be the empirical probability that audio stimuli a_i was chosen over a_j. When $\hat{P}_{ij} \geq 0.5$ and $\hat{P}_{jk} \geq 0.5$, then *WST* is satisfied if $\hat{P}_{ik} \geq 0.5$, *MST* is satisfied if $\hat{P}_{ik} \geq min(\hat{P}_{ij}, \hat{P}_{jk})$, and *SST* is satisfied if $\hat{P}_{ik} \geq max(\hat{P}_{ij}, \hat{P}_{jk})$. In Table 1, we see that *lack of distortions to target* had the highest mean satisfaction rates for all transitivity measures. This highlights the importance of scale clarity to participants and also suggests that *lack of distortions to the target source* should replace *preservation of the target source* and *absence of additional artificial noises*.

Table 1. Mean Pairwise Transitivity Statistics (N=10)

Quality Scale	TSR	WST	MST	SST
Overall Quality	0.91	0.97	0.93	0.61
Target Preservation	0.90	0.97	0.95	0.71
Supp. of Other Sources	0.92	0.99	0.94	0.60
Absence of Artif. Noises	0.91	0.99	0.98	0.71
Lack of Dist. to Target	**0.93**	**1.00**	**0.99**	**0.73**

4. CONCLUSION

In this work, we evaluated a crowdsourced pairwise comparison listening test for source separation evaluation. Our previous work [1] established that we could crowdsource a MUSHRA-like multi-stimulus listening test on the web and obtain scores comparable to a lab-based MUSHRA listening test. However, such multi-stimulus tests are limited to 12 stimuli or less and require ground-truth reference stimuli. Pairwise-comparison tests do not have these limitations. Therefore, we built on our previous work and crowdsourced a pairwise-comparison listening test. We estimated scores from the pairwise-comparison test using a Thurstone model, and then we compared these scores to the scores obtained in both web-based and lab-based multi-stimulus listening tests. The results for the pairwise-comparison listening tests establish that crowdsourced pairwise-comparison tests can also produce results similar to lab-based MUSHRA and can therefore be used when MUSHRA/MUSHRA-like multi-stimulus tests are not appropriate (e.g., when there isn't a reference, or there are more than 12 stimuli, etc.). However, if testing 12 or fewer stimuli, MUSHRA/MUSHRA-like multi-stimulus tests are preferred since they produce more discriminative scores given equal numbers of test participants. While we evaluated this listening test on the task of source separation evaluation, we believe it could generalize to the evaluation of other audio tasks which have differences in stimuli of a similar magnitude. Lastly, we showed the importance of clarity of quality scale definition when using novice crowdsourced participants. We hope that the results of this work will encourage researchers to not rely on poor automated measures of audio quality and to instead evaluate their algorithms using the consumers of their algorithms: the listeners.

5. REFERENCES

[1] Mark Cartwright, Bryan Pardo, Gautham Mysore, and Matthew Hoffman, "Fast and easy crowdsourced perceptual audio evaluation," in *Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, 2016, pp. 619–623.

[2] Andrew JR Simpson, Gerard Roma, Emad M Grais, Russell D Mason, Christopher Hummersone, and Mark D Plumbley, "Psychophysical evaluation of audio source separation methods," in *Proc. of Int'l. Conference on Latent Variable Analysis and Signal Separation*, 2017, pp. 211–221.

[3] E. Cano, D. FitzGerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: human perception vs quantitative metrics," in *Proc. of European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1758–1762.

[4] H. Wierstorf, D. Ward, R. Mason, E. M. Grais, C. Hummersone, and M. D. Plumbley, "Perceptual evaluation of source separation for remixing music," in *Proc. of Audio Engineering Society Convention 143*, 2017.

[5] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[6] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[7] ITU, "Recommendation ITU-R BS.1534-2: Method for the subjective assessment of intermediate quality level of audio systems," 2014.

[8] F. Ribeiro, D. Florencio, Zhang Cha, and M. Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," in *Proc. of Int'l. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[9] Michael Schoeffler, Fabian-Robert Stöter, Bernd Edler, and Jrgen Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R Recommendation BS. 1534 (MUSHRA)," in *Proc. of Web Audio Conference*, 2015.

[10] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proc. of ACM Int'l Conf. on Multimedia*, 2009, pp. 491–500.

[11] N. Jillings, B. De Man, D. Moffat, and J. D. Reiss, "Web audio evaluation tool: A browser-based listening test environment," in *Proc. of Sound and Music Computing Conference*, 2015.

[12] Antoine Liutkus, Fabian-Robert Stöter, Zafar Rafii, Daichi Kitamura, Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave, "The 2016 signal separation evaluation campaign," in *Proc. of Int'l. Conf. on Latent Variable Analysis and Signal Separation*, 2017, pp. 323–332.

[13] Herbert Aron David, *The method of paired comparisons*, vol. 12, DTIC Document, 1963.

[14] ITU, "Recommendation ITU-R BS.1116-2: Methods for the subjective assessment of small impairments in audio systems," 2014.

[15] T. Hofeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[16] Søren Bech and Nick Zacharov, *Perceptual Audio Evaluation—Theory, Method and Application*, John Wiley & Sons, 2007.

[17] X. Qianqian, H. Qingming, J. Tingting, Y. Bowei, L. Weisi, and Y. Yuan, "Hodgerank on random graphs for subjective video quality assessment," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 844–857, 2012.

[18] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz, "Pairwise ranking aggregation in a crowdsourced setting," in *Proc. of ACM Int'l. Conf. on Web Search and Data Mining*, 2013, pp. 193–202.

[19] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, 2010.

[20] L. L. Thurstone, "The method of paired comparisons for social values," *The Journal of Abnormal and Social Psychology*, vol. 21, no. 4, pp. 384–400, 1927.

[21] K. Tsukida and M. R. Gupta, "How to analyze paired comparison data," Report, U. of Washington, 2011.

[22] Matthew D. Hoffman and Andrew Gelman, "The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1593–1623, 2014.

[23] Andrew Gelman, *Bayesian data analysis*, Chapman & Hall/CRC texts in statistical science. CRC Press, Boca Raton, 3rd edition, 2014.