3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–8

Optimal Thinning of MCMC Output

Marina Riabiz MARINA.RIABIZ@KCL.AC.UK
King's College London, UK, Alan Turing Institute, UK

Wilson Ye Chen YE.CHEN@SYDNEY.EDU.AU
University of Sydney, Australia

Jon Cockayne JCOCKAYNE@TURING.AC.UK
Alan Turing Institute, UK

Pawel Swietach PAWEL.SWIETACH@DPAG.OX.AC.UK
Oxford University

Steven A. Niederer STEVEN.NIEDERER@KCL.AC.UK
King's College London, UK

Lester Mackey LMACKEY@STANFORD.EDU
Microsoft Research, US

Chris. J. Oates CHRIS.OATES@NEWCASTLE.AC.UK
Newcastle University, UK, Alan Turing Institute, UK

Abstract

The use of heuristics to assess the convergence and compress the output of Markov chain Monte Carlo can be sub-optimal in terms of the empirical approximations that are produced. Typically a number of the initial states are attributed to "burn-in" and removed, whilst the remainder of the chain is "thinned" if compression is also required. In this paper, we consider the problem of retrospectively selecting a subset of states, of fixed cardinality, from the sample path such that the approximation provided by their empirical distribution is close to optimal. We report a method based on greedy minimisation of a kernel Stein discrepancy, that is suitable for problems where heavy compression is required. Theoretical results guarantee consistency of the method and its effectiveness is demonstrated in the challenging context of parameter inference for ordinary differential equations.

1. Introduction

The most popular computational tool for non-conjugate Bayesian inference is Markov chain Monte Carlo (MCMC; Robert and Casella, 2013). The output $(X_i)_{i=1}^n$ from MCMC is usually post-processed to remove burn-in and, if compression is desired, to reduce the total amount of data that are stored (useful, for example, when the MCMC output is used as experimental design in multi-scale models). This results in just a subset of the states being retained, say indexed by $\pi \in \{1, \dots, n\}^m$, corresponding to an empirical approximation $Q_m := \frac{1}{m}\sum_{j=1}^m \delta(X_{\pi(j)})$ of the distribution P being targeted. Here $\delta(x)$ denotes a point mass centred at x. From the perspective of statistical efficiency, this post-processing can be sub-optimal: for a fixed computational budget, imposing tight control on bias through burn-in removal leads to estimators with high variance, due to the limited number of states that remain once the initial portion have been removed.

In this abstract we ask whether an "optimal" index set π can be selected. To this end, we summarize `Stein Thinning`, a method introduced by the authors (Riabiz et al., 2020), that selects an index set π, of specified cardinality m, such that the associated Q_m is close to optimal, and it is a consistent approximation of P. This is achieved by adopting a kernel Stein discrepancy (KSD) as the optimality criterion. The minimisation of KSD is performed using a greedy sequential algorithm and we report results that study the interplay of the greedy algorithm with the randomness inherent to the MCMC output.

This work lies in an active area of research that attempts to cast discrete approximation of a posterior distribution as an optimisation problem. Liu and Lee (2017) considered the use of KSD to optimally weight an arbitrary set $(X_i)_{i=1}^n \subset \mathbb{R}^d$ of states in an manner loosely analogous to importance sampling, and Hodgkinson et al. (2020) studied the effect of applying this to the output of MCMC. Outside of the MCMC framework, various criteria have been proposed to capture how well a discrete measure approximates P, so that global numerical optimisation methods can be used to arrive at a suitable point set: this is the case for the "minimum energy design" (MED) of Joseph et al. (2015, 2019), the "Stein points" (SP) of Chen et al. (2018) and "Support Points" of Mak and Joseph (2018). The reliance on global optimisation renders the theoretical analysis of these methods difficult, and, to obviate this, Chen et al. (2019) considered using Markov chains, in the context of SP, to approximately perform numerical optimisation, allowing a tractable analytic treatment. `Stein Thinning` differs from the contributions cited above, in one or more of the following senses: (1) It is primarily intended for compression, not re-weighting, of the MCMC output. (2) Continuous optimisation is avoided by working directly on the MCMC output. (3) Finite sample size error bounds and consistency of the algorithm can be established. In the remainder we report `Stein Thinning`, together with a theoretical and empirical assessment.

Notation: Throughout we assume that the target distribution P admits a positive and continuously differentiable density p on $\mathbb{R}^d$. Let $Q_n \Rightarrow P$ denote weak convergence of a sequence (Q_n) of measures to P. $\mathcal{H}(k)$ denotes a reproducing kernel Hilbert space (RKHS) of functions on $\mathbb{R}^d$, namely a Hilbert space, equipped with a function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, called a *kernel*, such that $\forall x \in \mathbb{R}^d$ $k(\cdot, x) \in \mathcal{H}(k)$ and $\forall x \in \mathbb{R}^d, h \in \mathcal{H}(k)$, $h(x) = \langle h, k(\cdot, x)\rangle_{\mathcal{H}(k)}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}(k)}$ denotes the inner product in $\mathcal{H}(k)$ and the induced norm is denoted $\| \cdot \|_{\mathcal{H}(k)}$. On the other hand, $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product, with induced Euclidean norm $\|x\| = \langle x, x \rangle^{1/2}$. We denote $\nabla \cdot$ the divergence operator in $\mathbb{R}^d$.

2. Methods

In Section 2.1 we recall KSD and in Section 2.2 we summarize our `Stein Thinning` algorithm for minimisation of KSD.

2.1. Kernel Stein Discrepancy

To select states from the MCMC output we require a notion of optimal approximation for probability distributions. To this end, recall the notion of a *Stein discrepancy*, proposed in Gorham and Mackey (2015). This was based on Stein's method (Stein, 1972), which consists of finding a differential operator $\mathcal{A}_P$, depending on P and acting on d-dimensional vector fields on $\mathbb{R}^d$, and a set $\mathcal{G}$ of sufficiently differentiable d-dimensional vector fields on

$\mathbb{R}^d$ such that $\int_{\mathbb{R}^d} \mathcal{A}_P g \, \mathrm{d}P = 0$ for all $g \in \mathcal{G}$. The seminal paper of Gorham and Mackey (2015) introduced the *Stein discrepancy*

$$D_{\mathcal{A}_P \mathcal{G}}(P, Q) \;=\; \sup_{g \in \mathcal{G}} \left| \int_{\mathbb{R}^d} \mathcal{A}_P g \, \mathrm{d}Q \right|. \tag{1}$$

In this paper we focus on a particular form of (1) due to Liu et al. (2016); Chwialkowski et al. (2016); Gorham and Mackey (2017), called a *kernel* Stein discrepancy (KSD). In this case, $\mathcal{A}_P$ is the *Langevin Stein operator* $\mathcal{A}_P g := p^{-1} \nabla \cdot (pg)$ derived in Gorham and Mackey (2015), where $\mathcal{G} := \{g : \mathbb{R}^d \to \mathbb{R}^d | \sum_{i=1}^d \|g_i\|^2_{\mathcal{H}(k)} \leq 1\}$ is the unit ball in a Cartesian product of RKHS. It follows from construction that the set $\mathcal{A}_P \mathcal{G}$ is the unit ball of another RKHS, denoted $\mathcal{H}(k_P)$, whose elements h satisfy $\int_{\mathbb{R}^d} h \mathrm{d}P = 0$. The explicit form of the kernel k_P was derived in Oates et al. (2017) and depends only on k and the gradient of the log-density of P. KSD is recognised as a maximum mean discrepancy in $\mathcal{H}(k_P)$ (Gretton et al., 2006) and is therefore fully characterised by the kernel k_P; we can thus unambiguously adopt the shorthand $D_{k_P}(Q)$ for $D_{\mathcal{A}_P \mathcal{G}}(P, Q)$.

Under suitable conditions on the kernel k, elements of $\mathcal{H}(k_P)$ have zero mean with respect to P, and under further conditions on k and P, the KSD can be explicitly computed for an empirical measure $Q_n = \frac{1}{n} \sum_{i=1}^n \delta(x_i)$, supported on states $x_i \in \mathbb{R}^d$:

$$D_{k_P}(Q_n) = \sqrt{\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_P(x_i, x_j)}. \tag{2}$$

Theoretical analysis in Gorham and Mackey (2017); Chen et al. (2018); Huggins and Mackey (2018); Chen et al. (2019); Hodgkinson et al. (2020); Gorham et al. (2020) has established sufficient conditions for when $D_{k_P}(Q_n) \to 0$ implies $Q_n \Rightarrow P$ (a property called *convergence control*). Assuming that such conditions for convergence control are met, KSD is a suitable optimality criterion to minimise for the post-processing of MCMC output.

2.2. Greedy Minimisation of KSD

Continuous optimisation algorithms over $\mathbb{R}^d$ were proposed for greedy minimisation of KSD in Chen et al. (2018, 2019), wherein at iteration n a new state x_n is appended to the current sequence $(x_1, \ldots, x_{n-1})$ by searching over a continuous domain in $\mathbb{R}^d$. Our proposed method does not attempt to solve the fundamentally difficult continuous optimisation problem for selection of the next point $x_n \in \mathbb{R}^d$. Instead, we exactly solve the discrete optimisation problem of selecting a suitable element x_n from supplied MCMC output. The method, called `Stein Thinning`, is straight-forward to implement and is succinctly stated in Algorithm 1. (The convention $\sum_{i=1}^0 = 0$ is employed.)

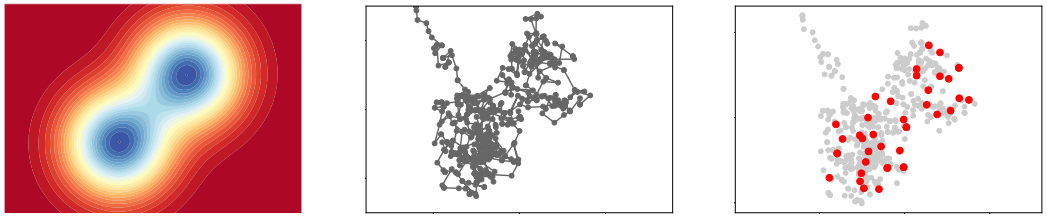
The algorithm is illustrated on a simple bivariate Gaussian mixture in Figure 1. Observe that, when m is fixed, the computational complexity of `Stein Thinning` is equal to $O(n)$, identical to the linear complexity of MCMC, while it is higher when m is increasing with n, being at most $O(nm^2)$. Notice moreover that, in general, the indices in π need not be distinct. That is, Algorithm 1 may prefer to include a duplicate state rather than to include a state which is not useful for representing P.

Data: The output $\{x_i\}_{i=1}^n$ from an MCMC method, a kernel k_P for which convergence control holds, and a desired cardinality $m \in \mathbb{N}$.
Result: The indices π of a sequence $(x_{\pi(j)})_{j=1}^m \subset \{x_i\}_{i=1}^n$ where $\pi \in \{1, \dots, n\}^m$.
for $j = 1, \dots, m$ **do**

$$\pi(j) \in \underset{i=1,\dots,n}{\arg\min} \frac{k_P(x_i, x_i)}{2} + \sum_{j'=1}^{j-1} k_P(x_{\pi(j')}, x_i)$$

end

Algorithm 1: The proposed method; `Stein Thinning`.



(a) (b) (c)

Figure 1: Illustration of `Stein Thinning`: (a) Contours of the distributional target P. (b) Markov chain Monte Carlo (MCMC) output, limited to 500 iterations to mimic a challenging computational context, exhibiting burn-in and autocorrelation that must be identified and mitigated. (c) A subset of $m = 40$ states from the MCMC output selected using `Stein Thinning`, which correctly ignores the burn-in period and stratifies states approximately equally across the two components of the target.

The suitability of KSD to quantify how well Q_m, the empirical measure resulting from `Stein Thinning`, approximates P is determined by the choice of the kernel k that is used in the definition of k_P. Several choices are possible and we follow Chen et al. (2019), who advocated the pre-conditioned inverse multi-quadric kernel $k(x, y) := (1 + \|\Gamma^{-1/2}(x - y)\|^2)^{-1/2}$, where we take a data-driven approach to select Γ, based on the MCMC output. To explore different strategies for the selection of Γ, we focus on the following candidates: (1) Median (`med`) $\Gamma = \ell^2 I$, where $\ell = \text{med} := \text{median}\{\|X_i - X_j\| : 1 \leq i < j \leq n_0\}$ is the median Euclidean distance between states; (2) Scaled median (`sclmed`) $\Gamma = \ell^2 I$, where $\ell = \text{med}/\sqrt{\log(m)}$, proposed in the context of Stein variational gradient descent in Liu and Wang (2016); (3) Sample covariance (`smpcov`), $\Gamma = \hat{\Sigma}$ provided that the sample covariance matrix $\hat{\Sigma}$ of the MCMC output is non-singular. The experiments in Section 4 shed light on which of these settings is the most effective, but we acknowledge that many other settings could also be considered. In what follows, we set $n_0 = \min(n, 10^3)$ for the `med` and `sclmed` settings, to avoid an $O(n^2)$ cost of computing ℓ, and otherwise set $n_0 = n$, so that the whole of the MCMC output is used to select Γ.

3. Theoretical Assessment

This section clarifies the limiting behaviour of `Stein Thinning` as $m, n \to \infty$, reporting convergence in mean-square of the KSD, along with a finite sample size error bound. Let V be a function $V : \mathbb{R}^d \to [1, \infty)$ and, for a function $f : \mathbb{R}^d \to \mathbb{R}$ and a measure μ on $\mathbb{R}^d$, let $\|f\|_V := \sup_{x \in \mathbb{R}^d} \frac{|f(x)|}{V(x)}$, $\|\mu\|_V := \sup_{\|f\|_V \leq 1} \left| \int_{\mathbb{R}^d} f \mathrm{d}\mu \right|$. Then a ψ-irreducible and aperiodic Markov chain $(X_i)_{i \in \mathbb{N}} \subset \mathcal{X}$ with n^{th} step transition kernel P^n is *V-uniformly ergodic* if and only if $\exists R \in [0, \infty), \rho \in (0, 1)$ such that

$$\|\mathrm{P}^n(x, \cdot) - P\|_V \leq RV(x)\rho^n \tag{3}$$

for all initial states $x \in \mathbb{R}^d$ and all $n \in \mathbb{N}$. The notation $\mathbb{E}$ will be used to denote expectation with respect to the law of the Markov chain. The proof of this result can be found in Riabiz et al. (2020).

Theorem 1 *Let P be a probability distribution on $\mathbb{R}^d$. Let $k_P : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a reproducing kernel with $\int_{\mathbb{R}^d} k_P(x, \cdot) \mathrm{d}P(x) = 0$ for all $x \in \mathbb{R}^d$. Consider a P-invariant, time-homogeneous, reversible Markov chain $(X_i)_{i \in \mathbb{N}} \subset \mathbb{R}^d$ generated using a V-uniformly ergodic transition kernel, such that (3) is satisfied with $V(x) \geq \sqrt{k_P(x, x)}$ for all $x \in \mathbb{R}^d$. Suppose that, for some $\gamma > 0$,*

$$b := \sup_{i \in \mathbb{N}} \mathbb{E}\left[e^{\gamma k_P(X_i, X_i)} \right] < \infty, \qquad M := \sup_{i \in \mathbb{N}} \mathbb{E}\left[\sqrt{k_P(X_i, X_i)} V(X_i) \right] < \infty.$$

Let π be an index sequence of length m produced by Algorithm 1 applied to the Markov chain output $(X_i)_{i=1}^n$. Then, with $C = \frac{2R\rho}{1-\rho}$, we have that

$$\mathbb{E}\left[D_{k_P}\left(\frac{1}{m} \sum_{j=1}^{m} \delta(X_{\pi(j)}) \right)^2 \right] \leq \frac{\log(b)}{\gamma n} + \frac{CM}{n} + \left(\frac{1 + \log(m)}{m} \right) \frac{\log(nb)}{\gamma}. \tag{4}$$

Observe that the upper bound in (4) is asymptotically minimised when (up to log factors) m is proportional to n. In practice we are interested in the case $m \ll n$, so we may for example set $m = \lfloor \frac{n}{1000} \rfloor$ if we aim for substantial compression. It is not claimed that the bound in (4) is tight and indeed empirical results in Section 4 endorse the use of `Stein Thinning` in the small m context.

4. Empirical Assessment

In this section we report a subset of the experiments presented in Riabiz et al. (2020), which compare the performance of `Stein Thinning` with traditional approaches to post-processing MCMC output. Our focus is a Bayesian inference problem defined by a system of ordinary differential equations (ODEs), due to Goodwin (1965), that has a $d = 4$ dimensional parameter to be inferred. Random Walk (RW) MCMC was implemented ($n = 10^6$) and trace plots reveal a clear a burn-in period. As in the rest of this abstract, we assume that compression of the MCMC output to $m \ll n$ states is desired. We therefore compare the following methods: (1) the traditional approach, which estimates a burn-in period using the convergence diagnostics of Vats and Knudson (2018) based on L independent

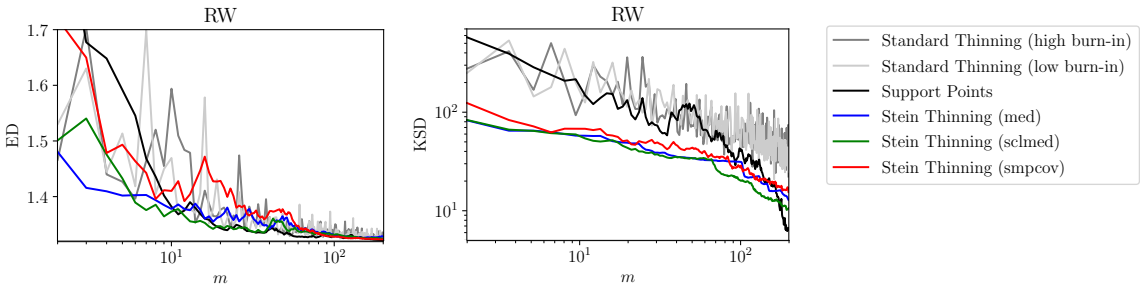


Figure 2: Goodwin oscillator: Energy distance (ED) to the posterior, and Kernel Stein discrepancy (KSD) for empirical distributions obtained through traditional burn-in and thinning (grey lines), `Support Points` (black line) and `Stein Thinning` (colored lines), based on output from four Random Walk (RW) MCMC.

chains[1] and discards it, followed by equidistant thinning of the remainder of the chain to obtain precisely m states; (2) the `Support Points` algorithm proposed in Mak and Joseph (2018), implemented in the `R` package `support`;[2] (3) the `Stein Thinning` algorithm that we have proposed, with each of the kernel choices described in Section 2.2. For a principled assessment, we computed two quantitative measures for how well the resulting empirical distributions approximate the posterior: (a) the energy distance (ED; Székely and Rizzo, 2004; Baringhaus and Franz, 2004), given up to an additive constant by

$$\mathrm{ED} := \frac{2}{m}\sum_{j=1}^{m}\int \|x - x_{\pi(j)}\|_\Sigma \,\mathrm{d}P(x) - \frac{1}{m^2}\sum_{j,j'=1}^{m}\|x_{\pi(j)} - x_{\pi(j')}\|_\Sigma, \tag{5}$$

where we use the norm $\|x\|_\Sigma := \|\Sigma^{-1/2}x\|$ induced by the covariance matrix of P, with both Σ and (5) being estimated from MCMC output, and (b) the KSD based on `med`, the simplest setting for Γ. The results for ED and KSD are shown in Figure 2. `Stein Thinning` based on `med` and `sclmed` performed at least as well as the other methods considered with respect to ED, even if ED is closely related to the quantity that the `Support Points` algorithm attempts to minimise (Mak and Joseph (2018) used the $\|\cdot\|$ norm in place of $\|\cdot\|_\Sigma$). The same holds for KSD, for all but the largest values of m considered. The `smpcov` setting performed well for small m but for large m its performance degraded (this may be because in `smpcov` there are more degrees of freedom in Γ that must be estimated). Note that neither ED nor KSD values tend to 0 as $m \to \infty$ in this experiment, since the number n of MCMC iterations was fixed.

1. The burn-in estimate depends on a number of tuning parameters. Here we report results with both the smallest and the largest estimated burn-in period, to explore the bias-variance trade-off involved with the choice of this parameter.
2. There do not yet exist theoretical results for this algorithm as implemented.

5. Conclusion

In this abstract we summarized `Stein Thinning`, a method that seeks a subset of the MCMC output, of fixed cardinality, such that the associated empirical approximation is close to optimal. Our theoretical analysis is able to handle the effect of the post-processing procedure jointly with the randomness involved in simulating from the Markov chain, such that finite sample size error bounds and the consistency of the overall estimator can be established.

References

Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

Wilson Ye Chen, Lester Mackey, Jackson Gorham, François-Xavier Briol, and Chris Oates. Stein points. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Wilson Ye Chen, Alessandro Barp, François-Xavier Briol, Jackson Gorham, Lester Mackey, Mark Girolami, and Chris Oates. Stein points Markov chain Monte Carlo. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Brian Goodwin. Oscillatory behavior in enzymatic control process. *Advances in Enzyme Regulation*, 3:318–356, 1965.

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Proceedings of the 29th Conference on Neural Information Processing Systems*, 2015.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Jackson Gorham, Anant Raj, and Lester Mackey. Stochastic Stein discrepancies. *arXiv:2007.02857*, 2020.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel method for the two-sample-problem. In *Proceedings of the 20th Conference on Neural Information Processing Systems*, 2006.

Liam Hodgkinson, Robert Salomone, and Fred Roosta. The reproducing Stein kernel approach for post-hoc corrected sampling. *arXiv:2001.09266*, 2020.

Jonathan Huggins and Lester Mackey. Random feature Stein discrepancies. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, 2018.

Roshan Joseph, Tirthankar Dasgupta, Rui Tuo, and Jeff Wu. Sequential exploration of complex surfaces using minimum energy designs. *Technometrics*, 57(1):64–74, 2015.

Roshan Joseph, Dianpeng Wang, Li Gu, Shiji Lyu, and Rui Tuo. Deterministic sampling of expensive posteriors using minimum energy designs. *Technometrics*, 61(3):297–308, 2019.

Qiang Liu and Jason D Lee. Black-box importance sampling. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Proceedings of the 30th Conference on Neural Information Processing Systems*, 2016.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Simon Mak and Roshan Joseph. Support points. *The Annals of Statistics*, 46(6A):2562–2592, 2018.

Chris Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society, Series B*, 79(3):695–718, 2017.

Marina Riabiz, Wilson Chen, Jon Cockayne, Pawel Swietach, Steven A Niederer, Lester Mackey, Chris Oates, et al. Optimal thinning of MCMC output. *arXiv preprint arXiv:2005.03952*, 2020.

Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of 6th Berkeley Symposium on Mathematical Statistics and Probability*, pages 583–602. University of California Press, 1972.

Gábor J Székely and Maria L Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

Dootika Vats and Christina Knudson. Revisiting the Gelman-Rubin diagnostic. *arXiv:1812.09384*, 2018.