Bayesian constraint relaxation

Leo L Duan, Alexander L Young, Akihiko Nishimura, David B Dunson

Prior information often takes the form of parameter constraints. Bayesian methods include such information through prior distributions having constrained support. By using posterior sampling algorithms, one can quantify uncertainty without relying on asymptotic approximations. However, *sharply* constrained priors are (a) unrealistic in many settings; and (b) tend to limit modeling scope to a narrow set of distributions that are tractable computationally. We propose to solve both of these problems via a general class of Bayesian *constraint relaxation* methods. The key idea is to replace the sharp indicator function of the constraint holding with an exponential kernel. This kernel decays with distance from the constrained space at a rate depending on a relaxation hyperparameter. By avoiding the sharp constraint, we enable use of off-the-shelf posterior sampling algorithms, such as Hamiltonian Monte Carlo, facilitating automatic computation in broad models. We study the constrained and relaxed distributions under multiple settings, and theoretically quantify their differences. We illustrate the method through multiple novel modeling examples.

KEY WORDS: Constrained Bayes, Constraint functions, Shrinkage on Manifold, Support Expansion, Ordered Simplex

1 Introduction

It is extremely common to have prior information available on parameter constraints in statistical models. For example, one may have prior knowledge that a vector of parameters lies on the probability simplex or satisfies a particular set of inequality constraints. Other common examples include shape constraints on functions, positive semi-definiteness of matrices and orthogonality. There is a very rich literature on optimization subject to parameter constraints. A common approach relies on Lagrange and Karush-Kuhn-Tucker multipliers (Boyd and Vandenberghe, 2004). However, simply producing a point estimate is often insufficient, as uncertainty quantification (UQ) is a key component of most statistical analyses. Usual large sample asymptotic theory, for example showing asymptotic normality of statistical estimators, tends to break down in constrained inference problems. Instead, limiting distributions may have a complex form that needs to be re-derived for each new type of constraint, and may be intractable.

An appealing alternative is to rely on Bayesian methods for UQ, including the constraint through a prior distribution having restricted support, and then applying Markov chain Monte Carlo (MCMC) to avoid the need for large sample approximations (Gelfand et al., 1992). Although this strategy appears conceptually simple, there are two clear limitations in practice. First, it is often too restrictive to assume that the parameter exactly satisfies the constraint, and one may want to allow slight deviations. Second, it is in general very difficult to develop tractable posterior sampling algorithms except in special cases. For example, one may be forced to focus on particular forms for the prior and likelihood function to gain tractability and/or may need to develop specially tailored algorithms on a case-by-case basis.

To overcome the first problem, one can attempt to tune the parameters in an unconstrained prior to place high probability close to the constrained space. However, this approach is limited in scope and often cannot be used. An alternative strategy is to partially enforce the constraint. For example, in considering monotone function estimation, one may impose monotonicity only at a subset of points (REF). However, it is typically difficult to decide exactly which constraints to remove, and outside of specialized cases, this strategy is not appropriate. Hence, to our knowledge, there is essentially no solution to the general problem of how to choose a prior that is concentrated *close* to a particular constrained space.

There is a richer literature on the second problem - posterior sampling under constraints. A common strategy is to reparameterize to reduce the number of constraints and/or simply the constraints. Common examples include reparameterizations of positive semidefinite covariance matrices, probability vectors on the simplex, and spherical parameters. Unfortunately, convenient reparameterizations often are not available and/or may lead to challenges in prior elicitation. A simple and intuitive prior in the original parameterization may induce a complex prior in the new parameterization, motivating the use of *black-box* convenience priors that may conflict with prior knowledge. There are several alternatives focused on particular models for

particular constraints. These include both specialized distributions on specific manifolds (e.g., von Mises-Fisher and extensions (Khatri and Mardia, 1977; Hoff, 2009)) and projection-based approaches (Lin and Dunson, 2014). There is also a recent literature developing algorithms for certain classes of manifolds; for example Byrne and Girolami (2013) develop a geodesic Monte Carlo algorithm.

The primary contribution of this article is to propose a broad class of Bayesian priors that are formally close to a constrained space, and can effectively solve both of our problems simultaneously. Importantly, the constrained subset of the original parameter space can be very small or even have measure zero, leading to some technical complications. The proposed class is very broad and acts to modify an initial unconstrained prior, having an arbitrary form, to be concentrated near the constrained space to an extent controlled by a hyperparameter. In addition, due to the simple form and lack of any sharp parameter constraints, general off-the-shelf sampling algorithms can be applied directly. Bypassing the need to develop or code complex and specialized algorithms is a major practical advantage.

2 Constraint Relaxation Methodology

Assume that $\theta \in \mathcal{D} \subset \mathcal{R}$ is an unknown continuous parameter, with $\dim(\mathcal{R}) = r < \infty$. The constrained sample space $\mathcal{D}$ is embedded in the r-dimensional Euclidean space $\mathcal{R}$, and can have either zero or positive measure with respect to Lesbesgue measure on $\mathcal{R}$.

The traditional Bayesian approach to including constraints requires a prior density $\pi_{\mathcal{D}}(\theta)$ with support on $\mathcal{D}$. The posterior density of θ given data Y and $\theta \in \mathcal{D}$ is then

$$\pi_{\mathcal{D}}(\theta \mid Y) \quad \propto \quad \pi_{\mathcal{D}}(\theta)\mathcal{L}(\theta; Y). \tag{1}$$

We assume in the sequel that the restricted prior $\pi_{\mathcal{D}}(\theta) \propto \pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)$, with $\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)$ an initial unconstrained prior on $\mathcal{R}$ and $\mathbb{1}_{\mathcal{D}}(\theta)$ an indicator function that the constraint is satisfied.

As noted in Section 1, there are two primary problems motivating this article. The first is that it is often too restrictive to assume that θ is *exactly* within $\mathcal{D}$ *a priori*, and often is more plausible to assume that θ has high probability of falling within a small neighborhood of $\mathcal{D}$. The second is that the difficulty of posterior sampling from (1) has greatly limited the scope of modeling, and there is a critical need for general algorithms that are tractable for a broad variety of choices of prior, likelihood and constraint.

In attempting to address these problems, we propose to replace (1) with the following *COnstraint RElaxed* (CORE) posterior density:

$$\tilde{\pi}_{\lambda}(\theta) \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\left(-\frac{1}{\lambda}\|v_{\mathcal{D}}(\theta)\|\right), \tag{2}$$

where we repress the conditioning on data Y in $\tilde{\pi}_{\lambda}(\theta)$ for concise notation. We assume the the initial prior

$\pi_{\mathcal{R}}(\theta)$ is proper and absolutely continuous with respect to Lesbesgue measure $\mu_{\mathcal{R}}$ on $\mathcal{R}$. We use $\|v_{\mathcal{D}}(\theta)\|$ as a distance from θ to the constrained space. For example, $\|v_{\mathcal{D}}(\theta)\| = \inf_{x\in\mathcal{D}} \|\theta - x\|$ with $\|.\|$ an appropriate metric.

The hyperparameter $\lambda > 0$ controls how concentrated the prior is around $\mathcal{D}$, and as $\lambda \to 0$ the kernel $\exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big)$ converges to $\mathbb{1}_{\mathcal{D}}(\theta)$ in a pointwise manner, excluding $\theta \in \partial\mathcal{D}$ on the boundary of $\mathcal{D}$. For all $\lambda > 0$, $\tilde{\pi}(\theta)$ introduces support outside of $\mathcal{D}$, creating a relaxation of the constraint.

The relaxation is carried out by placing a probability in a neighborhood surruonding $\mathcal{D}$. One could consider a d-neighborhood of $\mathcal{D}$

$$\{\theta \in \mathcal{R} : \|v_{\mathcal{D}}(\theta)\| \leq d)\}.$$

Clearly, the choice of distance $\|v_{\mathcal{D}}(\theta)\|$ is critical as it describes how support expansion occurs around $\mathcal{D}$. We first discuss the distances in the next subsection, then describe the relaxation strategy.

2.1 Distance to Constrained Space

To induce a neighborhood surrounding only $\mathcal{D}$, a minimal condition is that $\|v_{\mathcal{D}}(\theta)\|$ is zero for $\theta \in \mathcal{D}$ and positive for $\theta \notin \mathcal{D}$. There are many choices of distances that satisfy this minimal condition; however, depending on the focus of relaxation, one may be interested in choosing one of following two distances:

(I) The ones directly measuring how far the parameter $\theta \in \mathcal{R}$ is from the original constrained space $\mathcal{D}$. This is useful when one hopes to directly control and/or measure the support expansion based on the parameter θ. For example, when $\mathcal{D}$ is a symmetric compact manifold, one might be interested in isotropically expanding the support of θ, so that the relaxation does not create any directional asymmetry. In this article, we consider a simple distance

$$\|v_{\mathcal{D}}(\theta)\| = \inf_{x\in\mathcal{D}} \|\theta - x\|_k, \tag{3}$$

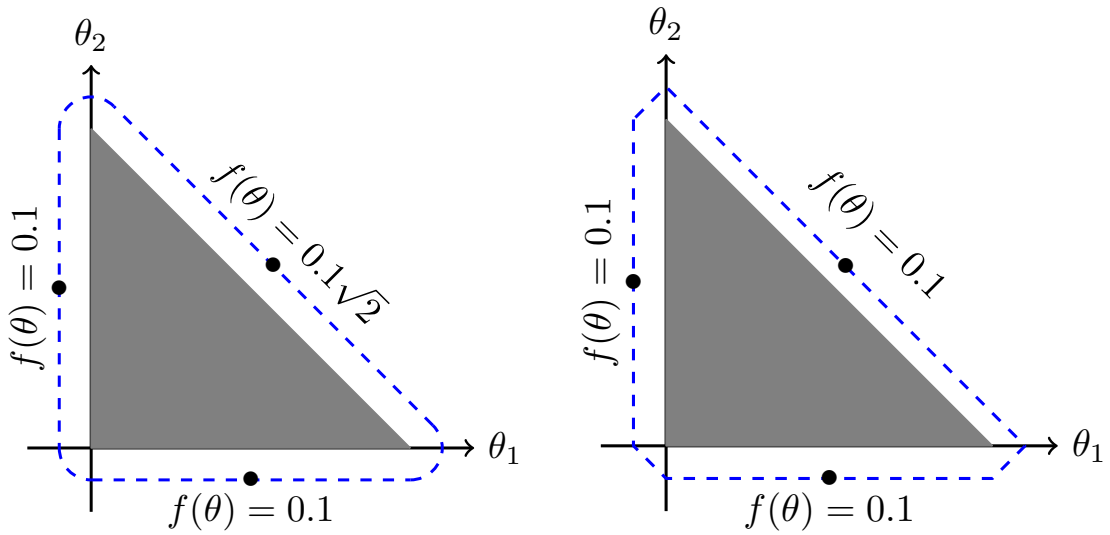
where $\|\cdot\|_k$ denotes a k-norm distance, typically $k = 1$ or 2.

(II) The ones measuring how far a *function* of $\theta \in \mathcal{R}$ is from its constrained value when $\theta \in \mathcal{D}$. This is particularly useful when the interest of relaxation is not on parameter θ, but on the function $f(\theta)$ associated with the constraint. For example, when a model is constrained by assigning an upper bound to a function, one may be interested in how much the posterior of the function deviates from that bound. Similar to (i), we also consider a simple distance

$$\|v_{\mathcal{D}}(\theta)\| = \inf_{x\in\mathcal{D}} \|f(\theta) - f(x)\|_k. \tag{4}$$

Note that when $f(x)$ is a constant for $\theta \in \mathcal{D}$, this distance has a closed form.

The two distances create different geometries of the neighborhood $\{\theta \in \mathcal{R} : \|v_{\mathcal{D}}(\theta)\| \leq d)\}$. To illustrate, we consider a constrained space under 3 inequalities $\mathcal{D} = \{(\theta_1, \theta_2) : \theta_1 > 0, \theta_2 > 0, \theta_1 + \theta_2 < 1\}$. For type-I distance, we use 2-norm $\inf_{x \in \mathcal{D}} \|\theta - x\|_2$. This creates a space expansion of θ equally along the boundary $\partial\mathcal{D}$ (Figure 1(a)). On the other hand, one might be interested in the total violation to the inequalities, as represented in a function $f(\theta) = (-\theta_1)_+ + (-\theta_2)_+ + (\theta_1 + \theta_2 - 1)_+$ where $(x)_+ = x$ if $x > 0$ and 0 otherwise. The associated type-II distance is $\|v_{\mathcal{D}}(\theta)\| = f(\theta)$. Using type-II distance, its expanded neighborhood has same value of $f(\theta)$ along its boundary (Figure 1(b)).



(a) Expanded neighborhood using type-I distance (b) Expanded neighborhood using type-II distance

Figure 1: The boundary of the neighborhood (blue dashed line) along $\|v_{\mathcal{D}}(\theta)\| = 0.1$ formed by two type of distances, around a constrained space formed by three inequalities (gray area). Panel(a) shows type-I distance expands the support of the θ equally further from the original boundary of $\mathcal{D}$; panel (b) shows type-II distance, while does not equally expands the support, maintains a constant amount of total violation to the three inequalities (quantified by a function $f(\theta)$) along $\|v_{\mathcal{D}}(\theta)\| = 0.1$.

The choice of distance largely depends on the prior belief about how the probability should decrease outside $\mathcal{D}$. For instance, for expanding support from a compact constrained space (assuming a uniform prior), simple isotropic type-I distance reflects the belief that θ should be found eqaully likely at a given distance away from $\mathcal{D}$; whereas simple type-II distance reflects the belief that the prior probability would drop at a rate depending on a function, such as total violation to inequalities in the last example. The former can be particularly useful for modeling geometric object, while the latter could be utilized in settings such as hypothesis testing about certain function. Obviously, customized distance might be needed for specific tasks. As our focus is to demonstrate the methodology, in this article we focus on the simple distances.

2.2 Constraint Relaxation

The sharply constrained space is equivalent to fixing the distance $\|v_{\mathcal{D}}(\theta)\| = 0$. Intuitively, one could achieve constraint relaxation by assigning a distribution that allows $\|v_{\mathcal{D}}(\theta)\|$ to be slightly positive. However, this raises two important questions: (i) what is the relationship between the relaxed and constrained densities? (ii) what are the limitations of this methodology? It turns out the answers are quite different depending on whether the constrained space $\mathcal{D}$ has a zero measure with respect to $\mu_{\mathcal{R}}$. Therefore, we discuss them separately in the next two subsections.

2.2.1 Constrained Space with Positive Measure

We start with the case when $\mathcal{D}$ is a subset of $\mathcal{R}$ with positive Lesbesgue measure, $\mu_{\mathcal{R}}(\mathcal{D}) > 0$. The sharply constrained density is simply a space-truncated version of the unconstrained one, for posterior

$$\pi_{\mathcal{D}}(\theta \mid Y) = \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta),$$

which is defined with respect to $\mu_{\mathcal{R}}$.

For constraint relaxation, we replace the indicator with an exponential function of distance

$$\tilde{\pi}_{\lambda}(\theta \mid Y) = \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big)}{\int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big)d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big) \tag{5}$$

which is also absolutely continuous with respect to $\mu_{\mathcal{R}}$.

To understand the role of the exponential function, one could view it as part of a folded Laplace density

$$\frac{1}{\lambda}\exp(-\frac{w}{\lambda})$$

for random variable $w = \|v_{\mathcal{D}}(\theta)\|$ with scale λ. Laplace has a sharp concentration near 0 with small λ and is routinely used in shrinkage literature. Observing the exponential tail probability for folded Laplace $P(w > k\lambda) = \exp(-k)$, one could fix a small λ and induce only a small relaxation from $\|v_{\mathcal{D}}(\theta)\| = 0$ as a priori. Alternatively, one could take a fully Bayesian approach by treating λ as a hyper-parameter and estimating it from its posterior. More generally, one could consider other densities such as folded generalized double Pareto (REFS). In this article, we focus on the folded Laplace case for its simplicity.

For this relaxation, we now address the two motivating questions regarding its relationship to sharply constrained density and its suitability. First, as $\lambda \to 0$, the relaxed density converges to the constrained density pointwise. On the other hand, even with small $\lambda > 0$, the relaxed density always has positive support outside of $\mathcal{D}$, as opposed to 0 support in sharply constrained case. As illustrated in the following example,

this can be more realistic for modeling. Second, the limitation is that relaxation should yield proper density $\int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big) d\mu_{\mathcal{R}}(\theta) < \infty$, which can be guaranteed if $\pi_{\mathcal{R}}(\theta)$ is proper.

Example: Gaussian under relaxed inequality

For now, we illustrate a Gaussian with a relaxed inequality. Suppose θ is the normal mean of the data, with the prior of θ follows a weakly informative Gaussian

$$y_i \overset{iid}{\sim} \mathrm{No}(\theta, 1) \text{ for } i = 1, \ldots, n, \qquad \theta \sim \mathrm{No}(0, 1000).$$

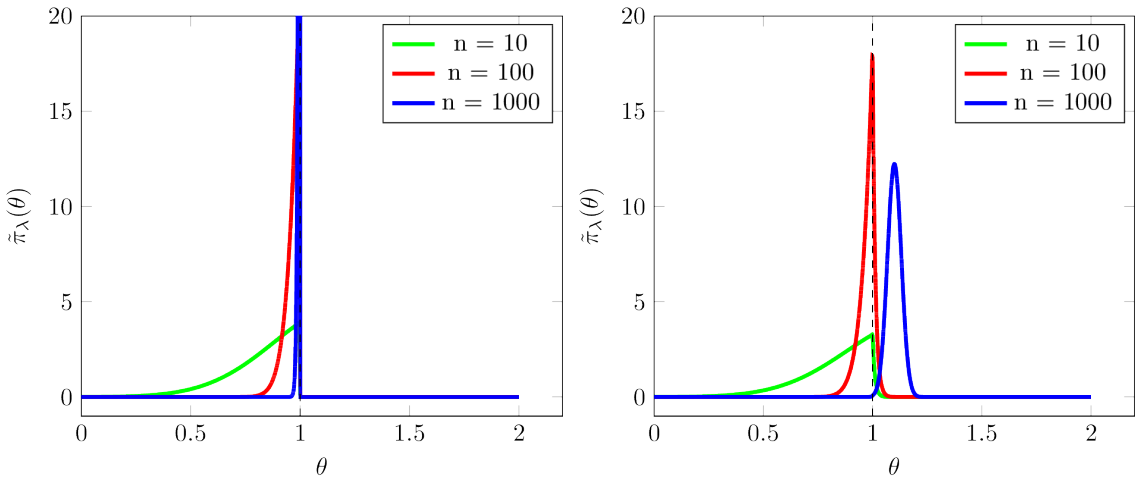
Suppose there is a prior belief that inequality $\theta < 1$. The posterior under a sharply constrained model is

$$\pi_{\mathcal{D}}(\theta \mid Y) \propto \sigma^{-1}\phi(\frac{\theta - \mu}{\sigma})\mathbb{1}_{\theta<1}, \quad \mu = \frac{\bar{y}n}{1/1000 + n}, \quad \sigma^2 = \frac{1}{1/1000 + n},$$

where ϕ denotes the density of standard Gaussian. However, suppose the sample size n keeps growing, but the data sample mean remains $\bar{y} = 1.2 > 1$. This suggests that the inequality is much less plausible, as the posterior become increasingly concentrated on the boundary (Figure 2(a)). Instead, it is more appealing to consider model under a relaxed constraint, with some preference in upholding the constraint as a priori, but allows the posterior to possibly deviate from this presumption:

$$\pi_{\lambda}(\theta \mid Y) \propto \sigma^{-1}\phi(\frac{\theta - \mu}{\sigma}) \exp(-\frac{(\theta - 1)_+}{\lambda}), \quad \mu = \frac{\bar{y}n}{1/1000 + n}, \quad \sigma^2 = \frac{1}{1/1000 + n}$$

where $(\theta - 1)_+$ is both type-I and type-II distance to constrained space. With $\lambda = 10^{-2}$, at $n = 10$ and 100, the relaxed posteriors are similar to the sharply constrained ones; however, a large sample size $n = 1000$ with $\bar{y} = 1.2$ leads to result that show the true normal mean is very likely outside the presumed constrained region.



(a) Sharply constrained posterior

(b) Constraint relaxed posterior

Figure 2: Densities of posteriors for Gaussian mean under sharp and relaxed constraints. With small sample ($n = 10$ and 100), relaxed constraint behaves very similarly to the sharp one; however, with strong evidence of sample mean $\bar{y} = 1.2 > 1$ in large sample ($n = 1000$), the relaxed constraint allow the likelihood to dominate in the posterior, generating more reasonable estimated mean outside the constrained region.

2.2.2 Constrained Space with Zero Measure

Now we consider when $\mathcal{D}$ is a measure zero subset of $\mathcal{R}$ with respect to $\mu_{\mathcal{R}}$. This introduces complication for associating the constrained $\pi_{\mathcal{D}}(\theta \mid Y)$ with an unconstrained density $\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)$. The former is no longer a re-normalized density of the latter. As a motivating example, the constrained space $\{\theta : \theta_1 + \theta_2 = 1\}$ of two independent uniforms $\theta_1, \theta_2 \overset{iid}{\sim} \text{Uniform}(0, 1)$ has exactly measure 0, using $\pi_{\mathcal{R}}(\theta_1, \theta_2) = 1$ can not lead to an appropriate constrained density. However, one could consider a lower-dimensional measure of θ_1 alone for deriving a valid constrained density; on the other hand, one could imagine "thickening" the space to $\{\theta : \theta_1 + \theta_2 \in (1 - \epsilon, 1 + \epsilon)\}$ for a positive 2-dimensional measure, then derived a relaxed density. Building upon this intuition, we will formalize the constrained density via a geometric measure theory concept known as Hausdorff measure, and construct a relaxed density using the constraint relaxation methodology.

For constructing a valid constrained density, we restrict ourselves to the cases where $\mathcal{D}$ can be represented implicitly as the unique solution set of a consistent system of equations $\{v_j(\theta) = 0\}_{j=1}^{s}$, so that $\mathcal{D} = \{\theta \mid v_j(\theta) = 0, j = 1, \ldots, s\}$ is a $(r - s)$-dimensional subspace of $\mathcal{R}$. Obviously, this is equivalent to defining constrained space via distance $\|v_{\mathcal{D}}(\theta)\| = \sum_j \|v_j(\theta)\| = 0$, with $\|v_j(\theta)\|$ corresponding to the distance as described in Section 2.1, provided the space formed satisfy the $(r - s)$-dimensionality requirement.

Recalling that $\mathcal{R}$ have dimensionality r, while $\mathcal{D}$ has zero r-dimensional Lesbesgue measure (that is, zero "volume"), it has a positive $(r-s)$-dimensional "surface area". This is formally known as $(r-s)$-dimensional Hausdorff measure, denoted by $\bar{\mathcal{H}}^{(r-s)}$. Taking the Hausdorff measure of a Borel subset of $\mathcal{D}$ and normalizing it by $\bar{\mathcal{H}}^{(r-s)}(\mathcal{D})$ yields a valid probability in the constrained space. More technical details will be provided

in the theory section. This sharply constrained density is a Hausdorff density

$$\pi_{\mathcal{D}}(\theta \mid Y) = \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)J^{-1}(\nu_{\mathcal{D}}(\theta))\mathbb{1}_{\mathcal{D}}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)J^{-1}(\nu_{\mathcal{D}}(\theta))d\bar{\mathcal{H}}^{(r-s)}(\theta)} \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)J^{-1}(\nu_{\mathcal{D}}(\theta))\mathbb{1}_{\mathcal{D}}(\theta),$$

defined with respect to $\bar{\mathcal{H}}^{(r-s)}$, where $J(\nu_{\mathcal{D}}(\theta)) = \sqrt{(D\nu_{\mathcal{D}})'(D\nu_{\mathcal{D}})}$ is the Jacobian of $\nu_{\mathcal{D}}$, which we assume is positive. Sometimes it is possible to reparameterize $(r-s)-$ dimensional Hausdorff density (with r parameters) and convert to a $(r-s)-$parameter Lesbesgue density, as more commonly seen in statistical literature; but such reparameterization may not exist.

Similar to the positive case, for relaxation we now replace the indicator with an exponential function of $\|v(\theta)\|$, adding support for $\|v(\theta)\| > 0$. The difference is that the range of $v(\theta)$, $\mathcal{X}$ should be controled such that each subspace defined by $\{\theta : v_{\mathcal{D}}(\theta) = x\}$ has the same dimension $(r-s)$ as $\{\theta : v_{\mathcal{D}}(\theta) = 0\}$ almost everywhere in $\mathcal{X}$. For a Borel subset $\mathcal{F} \subset \mathcal{R}$, this would allow us to compute the "surface area" of each slice $\{\theta : v_{\mathcal{D}}(\theta) = x\}$, then compute the "volume" of the expanded support over $x \in \mathcal{X}$. As opposed to the measure in the sharply constrained case, this leads to a more conventional Lesbesgue measure:

$$\int_{\mathcal{F}} \tilde{\pi}_\lambda(\theta)d\mu(\theta) = \frac{\int_{\mathcal{X}} \left[\int_{\{\theta: v(\theta)=x\}\cap\mathcal{F}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)J^{-1}(\nu_{\mathcal{D}}(\theta)))d\bar{\mathcal{H}}^{(r-s)}(\theta) \right] \exp(-\lambda^{-1}\|x\|)dx}{\int_{\mathcal{X}} \left[\int_{\{\theta: v(\theta)=x\}\cap\mathcal{F}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)J^{-1}(\nu_{\mathcal{D}}(\theta)))d\bar{\mathcal{H}}^{(r-s)}(\theta) \right] \exp(-\lambda^{-1}\|x\|)dx}. \tag{6}$$

Further, when $\mathcal{X}$ is the Lipschitz image of a bounded subset in $\mathcal{R}$, i.e. r-rectifiable, the double integral can be further simplified to a single Lesbesgue integral (Federer, 2014).

$$\int_{\mathcal{F}} \tilde{\pi}_\lambda(\theta)d\mu(\theta) = \frac{\int_{\mathcal{F}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\left(-\frac{1}{\lambda}\|\nu_{\mathcal{D}}(\theta)\| \right) \mathbb{1}_{\mathcal{X}}(v(\theta))d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\left(-\frac{1}{\lambda}\|\nu_{\mathcal{D}}(\theta)\| \right) \mathbb{1}_{\mathcal{X}}(v(\theta))d\mu_{\mathcal{R}}(\theta)}. \tag{7}$$

Omitting the integral over $\mathcal{F}$, this reduces to a constraint relaxed density:

$$\tilde{\pi}_\lambda(\theta) \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\left(-\frac{1}{\lambda}\|v_{\mathcal{D}}(\theta)\| \right) \mathbb{1}_{\mathcal{X}}(v(\theta)), \tag{8}$$

which is defined with respect to $\mu_{\mathcal{R}}$. Note the remarkable similarity to the relaxed density in the positive measure case, except for a more restricted suitable range of $v_{\mathcal{D}}(\theta)$.

For this relaxation, we again address the two important questions regarding its relationship to sharply constrained density and its limitations. First, the relaxed density no longer converges pointwise to the sharply constrained one, as they correspond to two different types of measure. Nevertheless, as $\lambda \to 0$, the relaxed

measure concentrates to an arbitrarily small layer around the constrained space. Second, there are several limitations imposed on $\mathcal{D}$, $v(\theta)$ and $\mathcal{X}$, such as the dimensionality and rectifiability. However, this still include a large variety of constrained spaces, such as compact manifolds embedded in Euclidean space via Lipschitz functions (e.g. orthonormal space). More details will be discussed in the theory section.

Example: Support expansion near a curved torus

For now, we consider a curved torus

$$\mathcal{D} = \big\{\theta : (\theta_1, \theta_2, \theta_3) = \big((1 + 0.5\cos\alpha_1)\cos\alpha_2, (1 + 0.5\cos\alpha_1)\sin\alpha_2, 0.5\sin\alpha_1\big), (\alpha_1, \alpha_2) \in [0, 2\pi)^2\big\}.$$

As the intrinsic dimension is 2 hence the 3-dimensional Lesbesgue measure $\mu_{\mathcal{R}}(\mathcal{D}) = 0$, with $\mathcal{R}$ a compact 3-dimensional Euclidean subspace. Diaconis et al. (2013) previously considered a uniform Hausdorff density $\pi_{\mathcal{D}}(\theta \mid Y) \propto 1$ over this compact manifold and utilized the transformed Lesbesgue density based on (α_1, α_2) for sampling within the manifold.

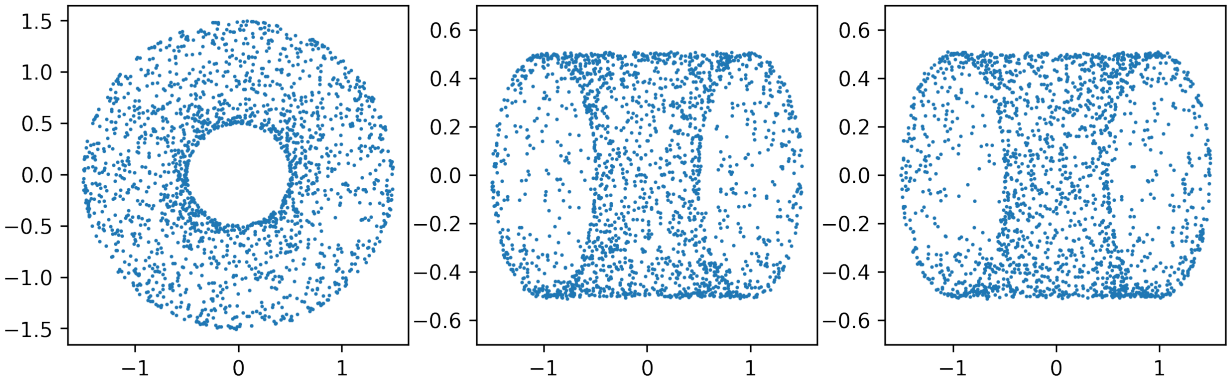
We now consider a different task by considering a relaxed distribution near the manifold. It is intuitive to consider type-I distance to $\mathcal{D}$, with

$$v(\theta) = \|v(\theta)\| = \inf_{(\alpha_1,\alpha_2)\in[0,2\pi)^2} \|(\theta_1 - (1 + 0.5\cos\alpha_1)\cos\alpha_2, \theta_2 - (1 + 0.5\cos\alpha_1)\sin\alpha_2, \theta_3 - 0.5\sin\alpha_1)\|_2.$$

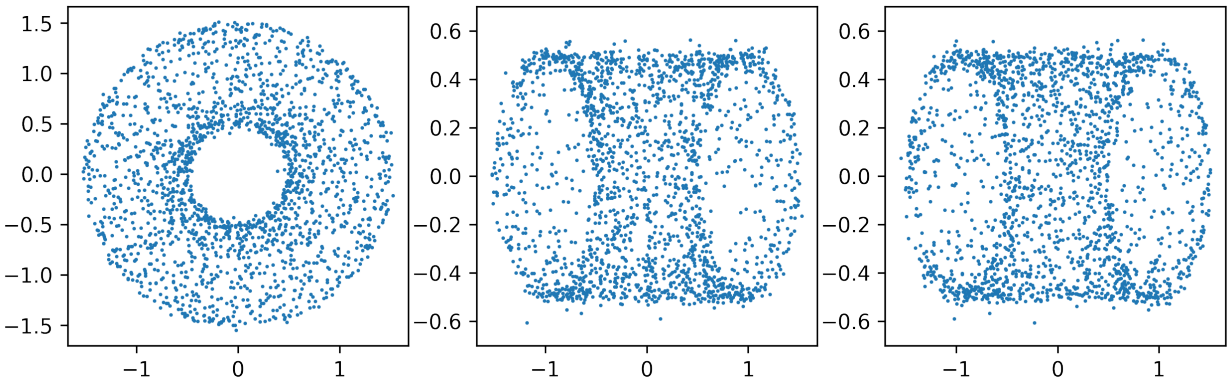
We obtained a relaxed density:

$$\tilde{\pi}_\lambda(\theta) \propto J(v_{\mathcal{D}}(\theta)) \exp(-\lambda^{-1}\|v(\theta)\|)$$

which is with respect to 3-dimensional Lesbesgue measure. Note $v(\theta)$ and $J(v_{\mathcal{D}}(\theta))$ do not have closed-form, fortunately, they can be numerically approximated by automatic optimization and differentiation tools. Figure 3 plots the random samples from constraint relaxed posteriors under two different values of λ, corresponding to different degrees of relaxation. The samples were obtained using random walk Monte Carlo.



(a) Constraint relaxed posterior with $\lambda = 0.01$



(b) Constraint relaxed posterior with $\lambda = 0.1$

Figure 3: Support expansion near a curved torus. Small λ controls the samples to be very close to the manifold (upper panel), while large λ adds support farther away from the constrained space (lower panel). Both densities correspond to more conventional Lesbesgue measure, instead of Hausdorff measure.

3 Theory

In this section, we provide theoretic details, mainly on two aspects: (i) the suitable constraints for relaxation; (ii) the error when using relaxed model as an approximation to constrained model.

3.1 Constrained Space with Positive Measure

For constrained space with positive measure, generally, as long as a tractable distance function exists such that $\mathcal{D} = \{\theta : \|v(\theta)\| = 0\}$, CORE is applicable.

We now focus on quantifying the difference between constrained and relaxed densities. Both of these densities are absolutely continuous with respect to Lesbesgue measure on $\mathcal{R}$. Thus, the expectation of g with respect to constrained density is

$$\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] = \int_{\mathcal{D}} g(\theta)\pi_{\mathcal{D}}(\theta)d\mu_{\mathcal{R}} = \frac{\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}. \tag{9}$$

Similarly, the expected value of g with respect to the relaxed density,

$$\mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)] = \int_{\mathcal{R}} g(\theta)\tilde{\pi}_\lambda(\theta)d\mu_{\mathcal{R}} = \frac{\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{R}} \mathcal{L}(\theta;Y)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}. \tag{10}$$

We can now consider the behavior of $E_{\tilde{\pi}_\lambda}[g]$ as $\lambda \to 0^+$.

Lemma 1. *Suppose $g \in \mathbb{L}^1(\mathcal{R}, \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}d\mu_{\mathcal{R}})$. Then,*

$$\left|\mathbb{E}[g(\theta) \mid \theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| \leq \frac{\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}\mathbb{E}|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)}{\left[\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right]^2}$$

where $E|g(\theta)| \propto \int_{\mathcal{R}} |g(\theta)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}d\mu_{\mathcal{R}}(\theta)$ is the expected value of $|g(\theta)|$ with respect to the unconstrained posterior density and $C_{\mathcal{R}} = \int_{\mathcal{R}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$ is the normalizing constant of this unconstrained posterior density. Furthermore, if $\|v_{\mathcal{D}}(\theta)\|$ is zero for all $\theta \in \mathcal{D}$ and positive for $\theta \in (\mathcal{R}\setminus\mathcal{D})^o$, it follows from the dominated convergence theorem that

$$\left|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| \to 0 \text{ as } \lambda \to 0^+.$$

Thus, one can obtain sufficiently accurate estimates of $\mathbb{E}[g|\theta \in \mathcal{D}]$ by sampling from $\tilde{\pi}_\lambda$ when λ is sufficiently small. From a practical standpoint, it is desirable to understand the rate at which $\mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]$ converges to $\mathbb{E}[g(\theta) \in \mathcal{D}]$. This question is addressed in the following theorem.

Theorem 1. *Suppose $g \in \mathbb{L}^2(\mathcal{R}, \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}d\mu_{\mathcal{R}})$, $v_{\mathcal{D}}(\theta) = \inf_{x\in\mathcal{D}} \|\theta - x\|_2$, $\mathcal{D}$ has a piecewise smooth boundary, and that $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is continuous on a open neighborhood containing $\mathcal{D}$. Then for $0 < \lambda \ll 1$,*

$$\left|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| = O(\sqrt{\lambda}).$$

This theorem follows by applying the Cauchy-Schwartz inequality to the term in the numerator of the bound given in Lemma 1. One can attain a bound depending on the surface area of $\mathcal{D}$ when it is bounded. The proofs of Lemma 1 and Theorem 1 are contained in Appendix A.

These results have some important implications both analytically and numerically. First, in addition to point estimates, $\mathbb{E}[\theta \mid \theta \in \mathcal{D}]$, it is possible to approximate probabilities $P(\theta \in \mathcal{F}|\theta \in \mathcal{D})$ and higher moments, e.g. $\mathbb{E}[\Pi_j \theta_j^{k_j}|\theta \in \mathcal{D}]$, so long as these moments exist for the unconstrained posterior density. Second, these bounds demonstrate that the error in using the relaxed density to approximate $\mathbb{E}[g(\theta)|\theta \in \mathcal{D}]$ is proportional to $\sqrt{\lambda}[\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)]^{-2}$ although this rate may not be optimal. In practice, λ may need to be very small, particularly in the case where $0 < P(\theta \in \mathcal{D}) \ll 1$. Of course, specific details of the scaling of $\left|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right|$ will depend upon $\mathcal{D}$ and $\|v_{\mathcal{D}}(\theta)\|$.

3.2 Constrained Space with Zero Measure

For the constrained space with zero measure, we review a few important concepts of geometric measure theory which are used throughout this section. First, recall the definition of Hausdorff measure.

Definition - Hausdorff Measure. *Let $A \subset \mathbb{R}^r$. Fix $s \leq r$. Then*

$$\mathcal{H}^s(A) = \lim_{\delta \to 0} \inf \left\{ \sum [diam(S_i)]^s : A \subseteq \cup S_i, diam(S_i) \leq \delta, diam(S_i) = \sup_{x,y \in S} \|x - y\| \right\}$$

.

We denote the normalized Hausdorff measure as $\bar{\mathcal{H}}^s(A) = \frac{\Gamma(\frac{1}{2})^s}{2^s \Gamma(\frac{s}{2}+1)} \mathcal{H}^s(A)$. When $s = r$, Lesbesgue and normalized Hausdorff measures coincide $\mu_{\mathbb{R}^m}(A) = \bar{\mathcal{H}}^s(A)$ (Evans and Gariepy, 2015). Additionally, for a subset $\mathcal{D}$, there exists a unique, critical value d such that

$$\bar{\mathcal{H}}^s(\mathcal{D}) = \begin{cases} 0, & s > d \\ \infty, & s < d. \end{cases}$$

The critical value, d, is referred to as the Hausdorff dimension of $\mathcal{D}$. We note that, when $\mathcal{D}$ is a compact, d-dimensional submanifold of $\mathbb{R}^m$, it will have Hausdorff dimension d and $\bar{\mathcal{H}}^d(\mathcal{D})$ is the d-dimensional surface area of A.

We now state the co-area formula which is used to define a regular conditional probability on the measure zero constrained space $\mathcal{D}$ and is pivotal in all of the proofs of the theorems.

Theorem 2. *Co-area formula (Diaconis et al., 2013; Federer, 2014) Suppose $\nu : \mathbb{R}^r \to \mathbb{R}^s$ with $s < r$ is Lipschitz and that $g \in \mathbb{L}^1(\mathbb{R}^r, \mu_{\mathbb{R}^r})$. Assume $J[\nu(\theta)] > 0$, then*

$$\int_{\mathbb{R}^r} g(\theta) J[\nu(\theta)] d\mu_{\mathbb{R}^r}(\theta) = \int_{\mathbb{R}^s} \left(\int_{\nu^{-1}(y)} g(\theta) d\bar{\mathcal{H}}^{r-s}(\theta) \right) d\mu_{\mathbb{R}^s}(y), \tag{11}$$

The behavior of the pre-images $\nu^{-1}(y)$ in the co-area formula are important for the convergence results presented later in this section. As such, we assume that $\mathcal{D}$ can be defined implicitly as the solution set to a system of s equations, $\{\nu_j(\theta) = 0\}_{j=1}^s$, where

(a) $\nu_j : \mathcal{R} \to \mathbb{R}$ is Lipschitz continuous,

(b) $v_j(\theta) = 0$ only for $\theta \in \mathcal{D}$,

(c) for $j = 1, \ldots, s$, the pre-image $v_j^{(-1)}(x)$ is a co-dimension 1 sub-manifold of $\mathcal{R}$ for $\mu_{\mathbb{R}}$-a.e. x in the range of v_j,

(d) $v_j^{(-1)}(0)$ and $v_k^{(-1)}(0)$ intersect transversally for $1 \leq j < k \leq s$.

We refer to the functions $v_1, \ldots, v_s$ as constraint functions. In this case, if we let $\nu : \mathcal{R} \to \mathbb{R}^s$ be the vector-valued function $\nu(\theta) = [\nu_1(\theta), \ldots, \nu_s(\theta)]^T$, then $\mathcal{D} = \ker(v)$ is a co-dimension s submanifold of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-a.e. x the range of v. Recall, the ambient space, $\mathcal{R}$, is $r-$dimensional. Therefore, it follows that $\mathcal{D}$ is a $(r-s)-$dimensional submanifold of $\mathcal{R}$, and it is natural to discuss the $(r-s)$-dimensional surface area of $\mathcal{D}$.

Property (a), guarantees that ν is itself Lipschitz. The remaining properties (b)-(d) are constructed so that $\nu^{(-1)}(x)$ for $x \in \mathbb{R}^s$ is also a submanifold which is close to $\mathcal{D} = \nu^{(-1)}(0)$ when x is near zero. In particular, the assumption of transversality ensures that $\nu^{(-1)}(x)$ with also be $r-s$ dimensional for x sufficiently close to 0, defined by the set $\mathcal{X}$.

The existence and uniqueness of the constraints must be addressed. In the case where $\mathcal{D}$ is specified by a collection of equality constraints – such as the probability simplex or the Stiefel manifold for example– it is not difficult to find a suitable set of constraint functions. Table 1 contains a number of examples of common constrained spaces and appropriate choices of constraint functions.

$\mathcal{R}$	$\mathcal{D}$	$\dim(\mathcal{R})$	$\dim(\mathcal{D})$	Constraint functions
$[0,1]^r$	Probability simplex, Δ^{r-1}	r	$r-1$	$v(\theta) = \sum(\theta) - 1$
$\mathbb{R}^r$	Line, span$\{\vec{u}\}$ $\vec{u} \neq \vec{0}$	r	1	$\nu_j(\vec{\theta}) = \vec{\theta}^T \vec{b}_j$ $\{\vec{b}_1, \ldots, \vec{b}_{r-1}\}$ a basis for span$\{\vec{u}\}^\perp$
$[-1,1]^r$	Unit sphere, $\mathbb{S}^{r-1}$	r	$r-1$	$v(\theta) = (\|\theta\|^2 - 1)$
$[-1,1]^{n\times k}$	Stiefel manifold, $\mathcal{V}(n,k)$	nk	$nk - \frac{1}{2}k(k+1)$	$v_{i,j}(\theta) = (\vec{\theta}_i'\vec{\theta}_j - \delta_{i,j})$ $1 \leq i \leq j \leq k$ and $\delta_{i,j} = \mathbb{1}_{i=j}$

Table 1: Table of constraints for some commonly used constrained spaces.

With regards to uniqueness, we note that the constraints cannot be unique in any case. For example, rescaling in each $v_j(\theta)$ will also satisfy (a)-(d). Naturally, an optimal choice will depend largely on the properties of the constrained distribution that one wishes to estimate making the choice of $\{\nu_j\}_{j=1}^s$ context dependent.

Under the given construction of the constrained space, we can now specify the regular conditional probability of θ, given $\theta \in \mathcal{D}$.

Theorem 3. *(Diaconis et al., 2013) Assume that $J(v(\theta)) > 0$ and that for each $z \in \mathbb{R}^s$ there is a finite non-negative p_z such that,*

$$m^{p_z}(z) = \int_{v^{-1}(z)} \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)}{J(v(\theta))} d\bar{\mathcal{H}}^{p_z}(\theta) \in (0, \infty).$$

Then, for any Borel subset F of $\mathcal{R}$, it follows that

$$P(F \mid v(\theta) = z) = \begin{cases} \frac{1}{m^{p_z}(z)} \int_F \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=z}}{J(v(\theta))} d\bar{\mathcal{H}}^{p_z}(\theta) & m^p(z) \in (0,\infty) \\ \delta(F) & m^p(z) \in \{0,\infty\} \end{cases}$$

is a valid regular conditional probability for $\theta \in \mathcal{D}$. Here, $\delta(F) = 1$ if $0 \in F$ and 0 otherwise.

By construction, $\{\theta : v(\theta) = z\}$ is a $(r-s)$ dimensional submanifold of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-a.e. z in $\mathcal{X}$, the suitable range of v. As such, it follows that one should take $p_z = r - s$. It is possible that $m^p(z) \in \{0,\infty\}$ for some $z \notin \mathcal{X}$; however, they are excluded during our construction. Most importantly, $0 \in \mathcal{X}$, therefore, Theorem 3 allows us to define

$$\pi_{\mathcal{D}}(\theta|\theta \in \mathcal{D}, Y) = \frac{1}{m^{r-s}(0)} \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=0}}{J(v(\theta))} \tag{12}$$

as the constrained posterior density.

As a result, we can define the conditional expectation of $g(\theta)$ given $\theta \in \mathcal{D}$ as

$$\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] = \mathbb{E}[g(\theta)|\nu(\theta) = 0\,] = \int_{\mathcal{R}} g(\theta)\pi_{\mathcal{D}}(\theta) d\bar{\mathcal{H}}^{r-s}(\theta).$$

The expected value of $g(\theta)$ with respect to the relaxed density, denote $\mathbb{E}_{\tilde{\Pi}}[g(\theta)]$, is

$$\mathbb{E}_{\tilde{\Pi}}[g(\theta)] = \frac{1}{m_\lambda} \int_{\mathcal{R}} g(\theta)\pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y) \exp\left(-\frac{1}{\lambda}\|\nu(\theta)\|_1 \right) d\mu_{\mathcal{R}}(\theta)$$

with $m_\lambda = \int_{\mathcal{R}} \pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y) \exp(-\lambda^{-1}\|\nu(\theta)\|_1) d\mu_{\mathcal{R}}(\theta)$. The primary results of the section are the following statements regarding the use of $\mathbb{E}_{\tilde{\Pi}}[g]$ to estimate $\mathbb{E}[g|\theta \in \mathcal{D}]$.

Theorem 4. *Let $m : \mathbb{R}^s \to \mathbb{R}$ and $G : \mathbb{R}^s \to \mathbb{R}$ be defined as follows*

$$m(x) = \int_{\nu^{-1}(x)} \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y)}{J(\nu(\theta))} d\mathcal{R}^{r-s}(\theta)$$

$$G(x) = \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y)}{J(\nu(\theta))} d\mathcal{R}^{r-s}(\theta).$$

Suppose that both m and G are continuous on an open interval containing the origin and that $g \in \mathbb{L}^1(\mathcal{R}, \pi_R\mathcal{L}(\theta;Y)d\mu_{\mathcal{R}})$. Then,

$$\left| \mathbb{E}_{\tilde{\Pi}}[g] - \mathbb{E}[g|\theta \in \mathcal{D}] \right| \to 0 \text{ as } \lambda \to 0^+.$$

Corollary 1. *In addition to the assumptions of Theorem 4, suppose that both m and G are differentiable at 0. Then*

$$\left|\mathbb{E}_{\tilde{\Pi}}[g] - \mathbb{E}[g|\theta \in \mathcal{D}]\right| = O\left(\frac{\lambda}{|\log \lambda|^s}\right)$$

as $\lambda \to 0^+$.

Like the results from Section 3.1, the convergence rates are sub-linear. Unlike the positive measure case, the convergence rates are dimension dependent. We assess the approximation error with different λ in the von Mises–Fisher distribution as described above. The result is provided in the appendix.

4 Posterior Computation

Compared to constrained density in space $\mathcal{D}$, relaxed density is supported in $\mathcal{R}$ and can be directly sampled via off–the–shelf tools such as slice sampling, adaptive Metropolis-Hastings and Hamiltonian Monte Carlo (HMC). In this section, we focus on HMC for its easiness to use and good performance in block updating of parameters.

4.1 Hamiltonian Monte Carlo under Constraint Relaxation

We provide a brief overview of HMC for continuous θ^* under constraint relaxation. Discrete extension is possible via recent work of Nishimura et al. (2017).

In order to sample θ, HMC introduces an auxiliary momentum variable $p \sim \mathrm{No}(0, M)$. The covariance matrix M is referred to as a *mass matrix* and is typically chosen to be the identity or adapted to approximate the inverse covariance of θ. HMC then sample from the joint target density $\pi(\theta, p) = \pi(\theta)\pi(p) \propto \exp(-H(\theta, p))$ where, in the case of the posterior under relaxation,

$$\begin{aligned} &H(\theta, p) = U(\theta) + K(p), \\ &\text{where } U(\theta) = -\log \pi(\theta), \\ &\qquad K(p) = \frac{p' M^{-1} p}{2}. \end{aligned} \tag{13}$$

with $\pi(\theta)$ is the unnormalized density in (5) or (8).

From the current state $(\theta^{(0)}, p^{(0)})$, HMC generates a proposal for Metropolis-Hastings algorithm by simulating Hamiltonian dynamics, which is defined by a differential equation:

$$\frac{\partial \theta^{(t)}}{\partial t} = \frac{\partial H(\theta, p)}{\partial p} = M^{-1}p,$$
$$\frac{\partial p^{(t)}}{\partial t} = -\frac{\partial H(\theta, p)}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta}. \tag{14}$$

The exact solution to (14) is typically intractable but a valid Metropolis proposal can be generated by numerically approximating (14) with a reversible and volume-preserving integrator (Neal, 2011). The standard choice is the *leapfrog* integrator which approximates the evolution $(\theta^{(t)}, p^{(t)}) \to (\theta^{(t+\epsilon)}, p^{(t+\epsilon)})$ through the following update equations:

$$p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial \theta}, \quad \theta \leftarrow \theta + \epsilon M^{-1}p, \quad p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial \theta} \tag{15}$$

Taking L leapfrog steps from the current state $(\theta^{(0)}, p^{(0)})$ generates a proposal $(\theta^*, p^*) \approx (\theta^{(L\epsilon)}, p^{(L\epsilon)})$, which is accepted with the probability

$$1 \wedge \exp\left(-H(\theta^*, p^*) + H(\theta^{(0)}, p^{(0)}))\right)$$

We refer to this algorithm as CORE-HMC.

4.2 Computing Efficiency in CORE-HMC

Since CORE expands the support from $\mathcal{D}$ to $\mathcal{R}$, it is useful to study the effect of space expansion on the computing efficiency of HMC. In this subsection, we provide some quantification.

In understanding the computational efficiency of HMC, it is useful to consider the number of leapfrog steps to be a function of ϵ and set $L = \lfloor \tau/\epsilon \rfloor$ for a fixed integration time $\tau > 0$. In this case, the mixing rate of HMC is completely determined by τ in the limit $\epsilon \to 0$ (Betancourt, 2017). In practice, while a smaller stepsize ϵ leads to a more accurate numerical approximation of Hamiltonian dynamics and hence a higher acceptance rate, it takes a larger number of leapfrog steps and gradient evaluations to achieve good mixing. For an optimal computational efficiency of HMC, therefore, the stepsize ϵ should be chosen only as small as needed to achieve a reasonable acceptance rate (Beskos et al., 2013; Betancourt et al., 2014). A critical factor in determining a reasonable stepsize is the *stability limit* of the leapfrog integrator (Neal, 2011). When ϵ exceeds this limit, the approximation becomes unstable and the acceptance rate drops dramatically. Below the stability limit, the acceptance rate $a(\epsilon)$ of HMC increases to 1 quite rapidly as $\epsilon \to 0$ and in fact satisfies $a(\epsilon) = 1 - \mathcal{O}(\epsilon^4)$ (Beskos et al., 2013).

For simplicity, the following discussions assume the mass matrix M is taken to be the identity, and $\mathcal{D} = \cap_{j=1}^{s}\{\theta : v_j(\theta) = 0\}$. We denote $\mathcal{D}_j = \{\theta : v_j(\theta) = 0\}$ and consider a directional relaxation, which is equivalent to replacing a single λ with several λ_j's in the relaxation part, i.e. $\exp(-\sum_j \|v_j(\theta^*)\|\lambda_j^{-1})$. There are generally two factors limiting the efficiency of HMC: (i) the width of support in constrained space; (ii) the largest eigenvalue of the Hessian matrix. For the former, using $\mathcal{Q}$ to denote a support, the width of support is related to the shortest distance to the boundary $\eta(\theta; \mathcal{Q}) = \inf_{\theta' \notin \mathcal{Q}} \|\theta' - \theta\|$. If $\eta(\theta; \mathcal{Q}) \approx 0$ for all $\theta \in \mathcal{Q}$, the proposal would likely be rejected using a large leap-frog step size. In such case, it is useful to utilize CORE to expand support and increase $\eta(\theta; \mathcal{Q})$ for better computing efficiency. For the eigenvalue, let $\mathbf{H}_U(\theta)$ denote the hessian matrix of $U(\theta) = -\log \pi(\theta)$. The linear stability analysis and empirical evidences suggest that, for stable approximation of Hamiltonian dynamics by the leapfrog integrator in $\mathbb{R}^p$, the condition $\epsilon < 2\xi_1(\theta)^{-1/2}$ must hold on most regions of the parameter space (Hairer et al., 2006), with $\xi_1(\theta)$ the largest eigenvalue of $\mathbf{H}_U(\theta)$. The hessian is

$$\mathbf{H}_U(\theta) = -\mathbf{H}_{\log(\mathcal{L}(\theta;y)\pi_{\mathcal{R}}(\theta))}(\theta) + \sum_j \lambda_j^{-1}\mathbf{H}\|v_j(\theta)\|\mathbb{1}_{\theta \notin \mathcal{D}_j}. \tag{16}$$

Note the second term is zero unless θ is outside of $\mathcal{D}_k$. As λ_j^{-1} in the second term often dominates the eigenvalue in the first term, hence the effective eigenvalue often is proportional to $\min_{j:\theta \notin \mathcal{D}_j} \lambda_j^{1/2}$.

Lastly, one may want to use CORE for approximate estimation under constrained model, while maintaining some computing efficiency. For this purpose, we now provide a practical guide on choosing λ_j. For $\mathcal{D}_j$'s with very small distance to support boundary $\eta(\theta; \mathcal{D}_j) \approx 0$, one should use moderate λ_j to increase support width; for $\mathcal{D}_j$'s without this issue, one should use very small $\lambda_j \approx 0$ to keep $\theta \in \mathcal{D}_j$, so that it has almost no influence on the hessian eigenvalue. The high rejection of HMC near the boundary of $\mathcal{D}_j$ can be avoided with random step size ϵ at each iteration, which is recommended for HMC in general (Livingstone et al., 2016). We evaluate the computing efficiency with different λ in sampling the von Mises–Fisher distribution. The result is provided in the appendix.

5 Simulated Examples

The simple computation of CORE frees up the modeling flexibility. We now illustrate more utility of the method via simulated examples.

Example: Sphere t Distribution

We now derive a new distribution on a $(p-1)$-sphere $\mathcal{D} = \{\theta \in \mathbb{R}^p : \|\theta\|_2 = 1\}$. Recall that von Mises–Fisher distribution (Khatri and Mardia, 1977) is the result of constraining a multivariate Gaussian

$\theta \sim \text{No}(F, I\sigma^2)$ with $F \in \mathcal{D}$

$$\pi_{\mathcal{D}}(\theta) \propto \exp(-\frac{\|F-\theta\|^2}{2\sigma^2})\mathbb{1}_{\theta'\theta=1} \propto \exp(\frac{F'}{\sigma^2}\theta)\mathbb{1}_{\theta'\theta=1}.$$

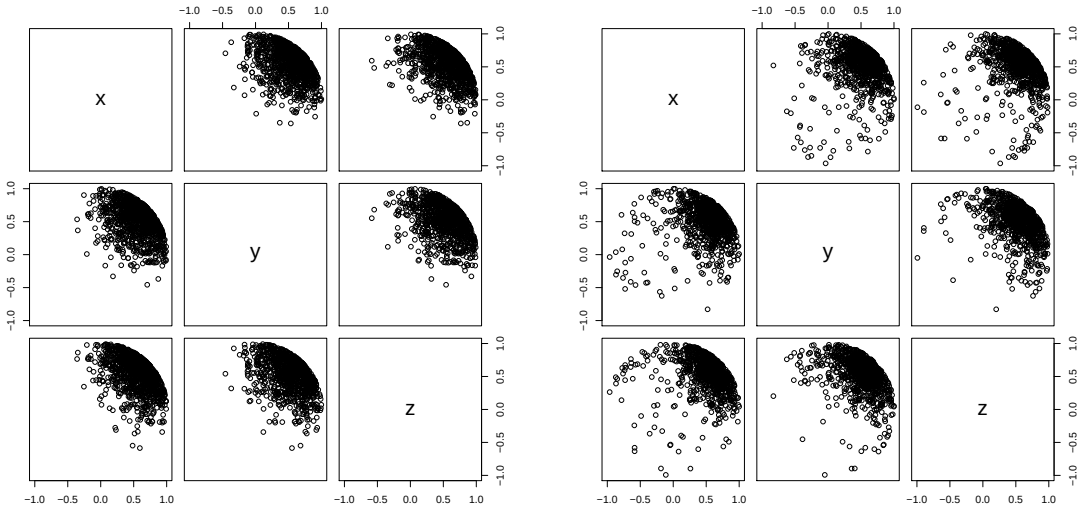
Although the final form appears more like an exponential, the behavior of von Mises–Fisher on sphere can be largely explained by its unconstrained parent Gaussian. In the Gaussian $\pi_{\mathcal{R}}(\theta)$, θ is symmetrically distributed around F, with density decaying exponentially as $\|\theta - F\|^2$ increases with rate $(2\sigma^2)^{-1}$; as the constrained density $\pi_{\mathcal{D}}(\theta)$ is proportional $\pi_{\mathcal{R}}(\theta)$, it concentrates similarly.

This naturally suggests we could use another distribution to induce different behavior on the sphere; then one could use CORE to generate approximate sample. We start from a multivariate t-distribution $\pi_{\mathcal{R}}(\theta)$, $t_m(F, I\sigma^2)$ with m degrees of freedom, mean $F \in \mathcal{D}$ and variance $I\sigma^2$, using (12) to generate a density

$$\begin{aligned}\pi_{\mathcal{D}}(\theta) \propto & (1 + \frac{\|F-\theta\|^2}{m\sigma^2})^{-\frac{(m+p)}{2}}\mathbb{1}_{\theta'\theta=1} \\ \propto & (1 - \frac{F'\theta}{1+m\sigma^2/2})^{-\frac{(m+p)}{2}}\mathbb{1}_{\theta'\theta=1}\end{aligned} \tag{17}$$

As in the t-distribution, the density decays polynomially as $\|F-\theta\|^2$ increases, as opposed to the exponential decay in Gaussian. We refer to this new distribution as sphere t-distribution.

CORE allows us to easily obtain approximate sample via relaxation function $\exp(-\lambda^{-1}\|\theta'\theta - 1\|)$. Figure 4 shows that the sphere t-distribution with $m = 3$ exhibits much less concentration than von Mises–Fisher on the sphere,. This can be useful for robust modeling when there could be 'outlier' on the sphere.



(a) von Mises–Fisher distribution.

(b) Sphere t-distribution with $m = 3$.

Figure 4: Sectional view of random samples from constrained distributions on a unit sphere inside $\mathbb{R}^3$. The distributions are derived through conditioning on $\theta'\theta = 1$ based on unconstrained densities of (a) $\text{No}(F, \text{diag}\{0.1\})$, (b) $t_3(F, \text{diag}\{0.1\})$, where $F = [1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}]'$. The samples are generated via CORE-HMC with $\lambda = 10^{-3}$.

Example: Ordered Dirichlet Distribution

We derive an ordered Dirichlet distribution. We build it upon the canonical Dirichlet distribution $\text{Dir}(\alpha)$ with $\pi_{\mathcal{D}}(\theta) \propto \prod_{j=1}^{J} \theta_j^{\alpha-1} \mathbb{1}_{\sum_{j=1}^{J} \theta_j = 1}$ and further impose order constraint, $1 > \theta_1 \geq \ldots \geq \theta_J > 0$, yielding

$$\pi_{\mathcal{D}}(\theta) \propto \prod_{j=1}^{J} \theta_j^{\alpha-1} \cdot \mathbb{1}_{\sum_{j=1}^{J} \theta_j = 1} \cdot \prod_{j=1}^{J-1} \mathbb{1}_{\theta_j \geq \theta_{j+1}} \,. \tag{18}$$

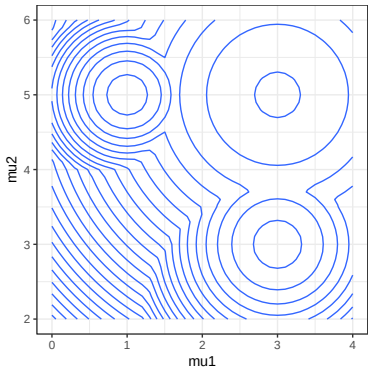
As commonly used in mixture model, canonical Dirichlet prior has its index j exchangeable. Since its permutation does not change the likelihood, label-switching problem often occurs (reviewed in Jasra et al. (2005)). Naturally, order constraint in θ can alleviate this problem, especially in preventing the switch between large θ_j and small $\theta_{j'}$.

To illustrate, we consider a hierarchical normal distribution with a common variance but the mean from a mixture, for data $y_i \in \mathbb{R}^2$ indexed by $i = 1, \ldots, n$:

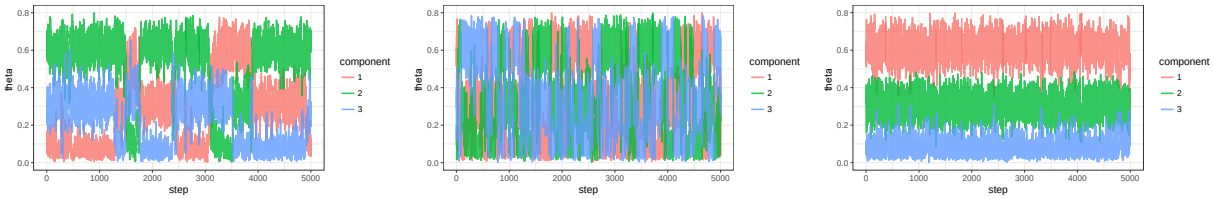
$$y_i \overset{indep}{\sim} \text{No}(\mu_i, \Sigma), \qquad \mu_i \overset{iid}{\sim} G, \qquad G(.) = \sum_{j=1}^{J} \theta_j \delta_{\mu_j}(.),$$

We generate $n = 100$ samples from 3 components with $\{\theta_1, \theta_2, \theta_3\} = \{0.6, 0.3, 0.1\}$, $\{\mu_1, \mu_2, \mu_3\} = \{[1, 5], [3, 3], [3, 5]\}$ and $\Sigma = I_2$. We assign weakly informative priors $\text{No}(0, 10I_2)$ for each μ_j and inverse–

Gamma prior for the diagonal element in $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$ with $\sigma_1^2, \sigma_2^2 \sim \text{IG}(2, 1)$. Figure 5(a) shows the contour of posterior density of μ. The small component sample size leads to large overlap among the posterior.



(a) Posterior density of the component means $\{\mu_j\}_{j=1}^3$.



(b) Gibbs sampling of unordered Dirichlet

(c) HMC sampling of unordered Dirichlet

(d) HMC sampling of ordered Dirichlet

Figure 5: Contour of the posterior density of component means and traceplot of the posterior sample for the component weights w, in a 3-component normal mixture model. Panel (a) shows that there is significant overlap among component means $\{\mu_j\}_{j=1}^3$. Without ordering in θ, its traceplot shows label-switching issue in both Gibbs (b) and HMC (c) sampling of Dirichlet distribution. The ordered Dirichlet distribution has significantly less label-switching issue (d), where we utilize CORE to obtain approximate posterior sample.

The ordering disrupts the traditional Gibbs sampling (Ishwaran and James, 2001), however, one could still obtain approximate posterior using CORE. We use $\exp(-\lambda_1^{-1} \| \sum_{j=1}^J \theta_j - 1) \|) \prod_{j=1}^{J-1} \exp[-\lambda_2^{-1} \max(\theta_{j+1} - \theta_j, 0)]$ to relax the constraints. We use $\lambda_1 = 10^{-3}$ on simplex constraint to allow efficient sampling and $\lambda_2 = 10^{-6}$ to induce almost no relaxation on the ordering. The posterior estimates of θ in CORE are close to the true values and indistinguishable from the other methods, except for a very small relaxation $\sum_{j=1}^J \theta_j - 1$ at posterior mean 0.001 ($(-0.001, 0.003)$ for 95% credible interval).

We compare the traceplots of ordered Dirichlet and unordered Dirichlet. Without the order constraint, significant label-switching occur in both Gibbs and HMC (traceplots in Figure 5(b,c)), whereas ordered Dirichlet has almost no label-switching(Figure 5(d)).

6 Application: Sparse Latent Factor Model in a Population of Brain Networks

We apply CORE in a real data application of analyzing a population of brain networks. The brain connectivity structures are obtained in the data set KKI-42 (Landman et al., 2011), which consists of $n = 21$ healthy subjects without any history of neurological disease. For each subject, we take the first scan out of the scan-rescan data as the input data, and reserve the second scan for model validation later. Each observation is a $V \times V$ symmetric network, recorded as an adjacency matrix A_i for $i = 1, \ldots, n$. The regions are constructed via the Desikan et al. (2006) atlas, for a total of V = 68 nodes (brain regions). For the ith matrix A_i, $A_{i,k,l} \in \{0, 1\}$ is the element on the kth row and lth column of A_i, with $A_{(i,k,l)} = 1$ indicating there is an connection between kth and lth region, $A_{(i,k,l)} = 0$ if there is no connection. The matrix is symmetric with the diagonal records empty $A_{(i,k,k)}$ for all i and k.

One interest in neuroscience is to quantify the variation of brain networks and identify the brain regions contributing to difference. Extending latent factor model to multiple matrices, one appealing approach is to have the networks share a common factor matrix but let the loadings vary across subjects.

$$A_{(i,k,l)} \sim \text{Bern}(\frac{1}{1+\exp(-\psi_{(i,k,l)} - z_{(k,l)})})$$

$$\psi_{(i,k,l)} = \sum_{r=1}^{d} v_{(i,r)} u_{(k,r)} u_{(l,r)}$$

$$z_{(k,l)} \sim \text{No}(0, \sigma_z^2), \quad \sigma_z^2 \sim \text{IG}(2, 1)$$

$$v_{(i,r)} \sim \text{No}(0, \sigma_{v,(r)}^2), \quad \sigma_{v,(r)}^2 \sim \text{IG}(2, 1)$$

for $k > l$, $k = 2, \ldots, V$, $i = 1, \ldots, n$; we choose weakly informative prior inverse Gamma IG$(2, 1)$, as appropriate for the scale parameters $\sigma_\cdot^2$ under the logistic link; $Z = \{z_{(k,l)}\}_{k=1,\ldots,V;l=1,\ldots,V}$ is a symmetric unstructured matrix that serves as the latent mean; $\{v_{(i,r)}\}_{r=1,\ldots,d}$ is the loading for the ith network, with each $v_{(i,r)} > 0$; $U = \{u_{(k,r)}\}_{k=1,\ldots,V;r=1,\ldots,d}$ is the $V \times d$ shared factor matrix.

To help convergence, it is common to let the factor matrix U on a Stiefel manifold $\mathcal{V}(n, d) = \{U : U'U = I_d\}$, so that free rotation or rescaling of U is less likely to occur (Hoff et al., 2016). Obviously, slightly relaxing this constraint via CORE can still retain this convergence property, at this time, much more flexible priors can be assigned for U.

We now consider apply a shrinkage prior near the Stiefel manifold, in order to identify the important nodes. We use the Dirichlet-Laplace prior (Bhattacharya et al., 2015):

$$u_{(k,r)} = \eta_{(k,r)}\kappa_{(k,r)}\sigma_u$$

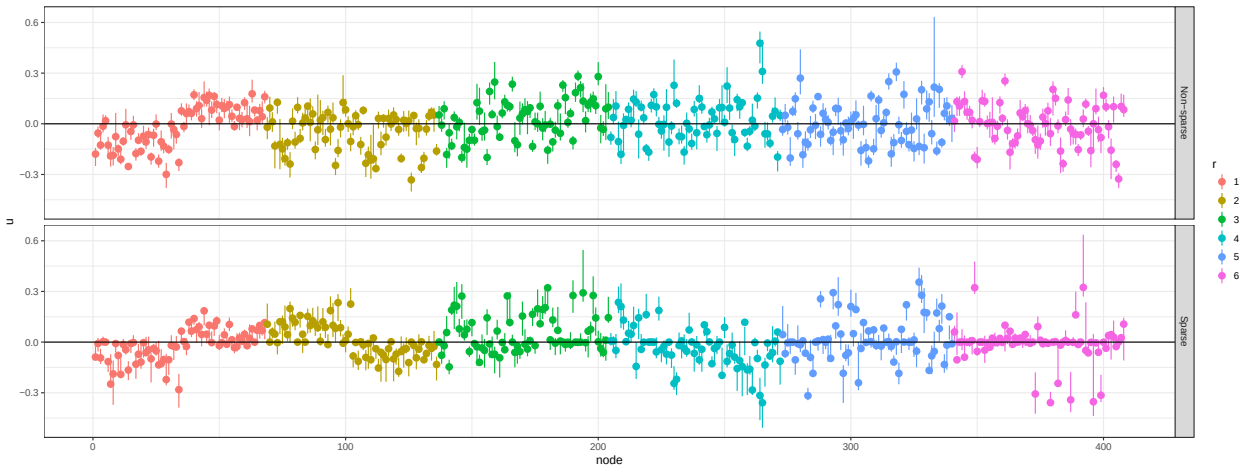
$$\eta_{(k,r)} \sim \text{Lap}(0,1), \quad \{\kappa_{(1,r)} \dots \kappa_{(V,r)}\} \sim \text{Dir}(\alpha), \quad \sigma_u^2 \sim \text{IG}(2,1)$$

for $k = 1, \dots, V$, $\text{Lap}(0,1)$ denotes the Laplace distribution centered at 0 with scale 1. To induce sparsity in each Dirichlet, we use $\alpha = 0.1$ as suggested by Bhattacharya et al. (2015). We replace the constraint by relaxation function $\exp(-\lambda^{-1}\|U'U - I\|)$, with $\lambda = 10^{-3}$.

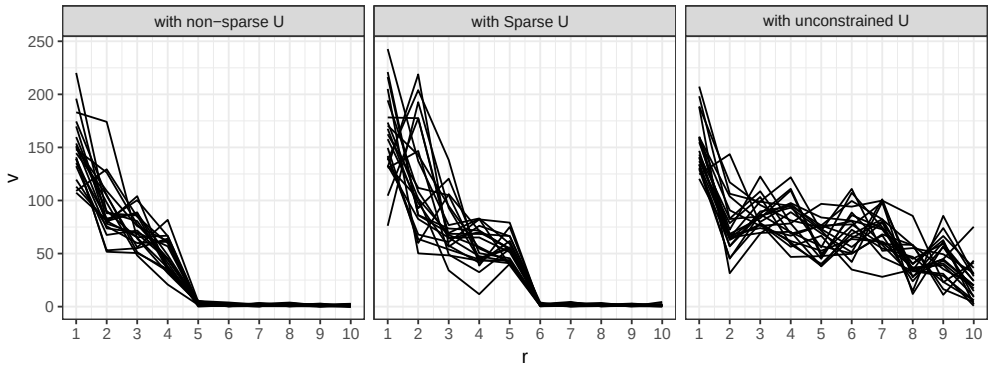
For comparison, we test with the specified model against two baseline models: one without shrinkage prior; one without the shrinkage prior and close-to orthonormality constraint. We use simple $u_{(k,r)} \sim \text{No}(0,1)$ in those two models. We run all models for $10,000$ iterations and discard the first $5,000$ iteration as burn-in. For each iteration, we run 300 leap-frog steps. For efficient computing, we truncated $d = 20$.

Figure 6(a) plots the top 6 factors U_r estimated near the Stiefel manifold, without and with shrinkage prior. Under the shrinkage prior, sparsity starts to show as early as the third factor U_3; the second factor U_2 shows a clear partition of first 34 and latter 34 nodes, which correspond to the two hemispheres of the brain. Accordingly, in the estimated loadings $v_{(i,r)}$ (Figure 6(b)), model under shrinkage prior detects more variability in the subject-specific loadings (represented by each line), especially over the second factor.

For the model with a completely unconstrained U, the factors and loadings fail to converge. And the loadings have a much slower drop to 0, compared to the two models with U near the Stiefel manifold. This indicates that near-orthogonal factors are more efficient representation of the span.



(a) Posterior mean and pointwise 95% credible interval of the factors $U_1, \ldots, U_6$ in the two constrained models.



(b) Posterior mean of the loadings $v_{i,r}$ for 21 subjects using three models. Each line represents the loadings for one subject over $r = 1, \ldots 10$.

Figure 6: Factors and loadings estimates of the network models. Panel (a) shows that shrinkage model shows difference starting from the second factor (model with unconstrained U is omitted due to non-convergence in the factor); Panel (b) compares the varying loadings of the subjects in three models.

We further validate the models by assessing the area under the receiver operating characteristic curve (AUC). We compute the posterior mean of estimated connectivity probability for each individual, then evaluate AUC against the observed binary data (fitted AUC) and the unobserved binary data from the second scan of the same subjects (prediction AUC). Table 2 lists the benchmark results. The two models with near-orthonormality show much better performance, especially in prediction. Although we do not see a clear improved prediction by further using shrinkage prior, the sparse loadings it discover could be more useful for scientific interpretation. In terms of computing efficiency, CORE-HMC generates reasonable effective samples (ESS) per 1000 iterations; while the model with no constraint suffers from extremely small ESS due to non-convergence.

Model	(i).with shrinkage & near-orthonormality	(ii).with near-orthonormality only	(iii).completely unconstrained
Fitted AUC	97.9%	97.1%	96.9%
Prediction AUC	96.2%	96.2%	93.6%
ESS /1000 Iterations	193.72	188.10	8.15

Table 2: Benchmark of 3 models for 21 brain networks. Models with near-orthonormality show much better performance in both AUC and computing efficiency.

7 Discussion

Using constraint relaxation, we circumvent the common difficulties of constrained modeling, such as prior specification and posterior estimation. One interesting further direction perhaps is to tackle the 'doubly intractable' problem. This issue emerges when data (instead of parameters) are on the constrained space, forcing some associated parameters into an intractable constant. It is worth studying how to exploit CORE to approximate this normalization constant. Another task under the CORE framework may involve development of a formal test on whether the parameters reside on the constrained space.

References

Beskos, A., N. Pillai, G. Roberts, J. M. Sanz-Serna, and A. Stuart (2013, 11). Optimal Tuning of the Hybrid Monte Carlo Algorithm. *Bernoulli 19*(5A), 1501–1534.

Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434*.

Betancourt, M., S. Byrne, and M. Girolami (2014). Optimizing the Integrator Step Size for Hamiltonian Monte Carlo. *arXiv:1411.6669*.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2015). Dirichlet–Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association 110*(512), 1479–1490.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge university press.

Byrne, S. and M. Girolami (2013). Geodesic Monte Carlo on Embedded Manifolds. *Scandinavian Journal of Statistics 40*(4), 825–845.

Desikan, R. S., F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, et al. (2006). An Automated Labeling System for Subdividing The Human Cerebral Cortex on MRI Scans into Gyral Based Regions Of Interest. *Neuroimage 31*(3), 968–980.

Diaconis, P., S. Holmes, M. Shahshahani, et al. (2013). Sampling from a Manifold. In *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, pp. 102–125. Institute of Mathematical Statistics.

Evans, L. C. and R. F. Gariepy (2015). *Measure Theory and Fine Properties of Functions*. CRC press.

Federer, H. (2014). *Geometric Teasure Theory*. Springer.

Gelfand, A. E., A. F. Smith, and T.-M. Lee (1992). Bayesian Analysis of Constrained Parameter and Truncated Data Problems using Gibbs Sampling. *Journal of the American Statistical Association 87*(418), 523–532.

Hairer, E., C. Lubich, and G. Wanner (2006). *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer-Verlag.

Hoff, P. D. (2009). Simulation of The Matrix Bingham–von Mises–Fisher Distribution, with Applications to Multivariate and Relational Data. *Journal of Computational and Graphical Statistics 18*(2), 438–456.

Hoff, P. D. et al. (2016). Equivariant and Scale-free Tucker Decomposition Models. *Bayesian Analysis 11*(3), 627–648.

Ishwaran, H. and L. F. James (2001). Gibbs Sampling Methods for Stick-breaking Priors. *Journal of the American Statistical Association 96*(453), 161–173.

Jasra, A., C. C. Holmes, and D. A. Stephens (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, 50–67.

Khatri, C. and K. Mardia (1977). The von Mises-Fisher Matrix Distribution in Orientation Statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 95–106.

Landman, B. A., A. J. Huang, A. Gifford, D. S. Vikram, I. A. L. Lim, J. A. Farrell, J. A. Bogovic, J. Hua, M. Chen, S. Jarso, et al. (2011). Multi-parametric Neuroimaging Reproducibility: a 3-T Resource Study. *Neuroimage 54*(4), 2854–2866.

Lin, L. and D. B. Dunson (2014). Bayesian Monotone Regression Using Gaussian Process Projection. *Biometrika*.

Livingstone, S., M. Betancourt, S. Byrne, and M. Girolami (2016). On the Geometric Ergodicity of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1601.08057*.

Neal, R. M. (2011). MCMC using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo 2*, 113–162.

Nishimura, A., D. Dunson, and J. Lu (2017). Discontinuous Hamiltonian Monte Carlo for Sampling Discrete Parameters. *arXiv preprint arXiv:1705.08510*.

A Proofs for Section 3.1

Proof. Proof of Lemma 1

Recall, that the distance function $\|v_{\mathcal{D}}(\theta)\|$ is chosen so that $\|v_{\mathcal{D}}(\theta)\|$ is zero for all $\theta \in \mathcal{D}$. It follows that for any function g

$$\begin{aligned}&\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)\\ &= \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta) + \int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta).\end{aligned} \tag{19}$$

For brevity, we let $f(\theta) = \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ and use $df(\theta) = \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$ throughout the proof. Then,

$$\begin{aligned}&\left|E[g(\theta)|\theta\in\mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right|\\ &= \left|\frac{\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)} - \frac{\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{R}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)}\right|\\ &= \left|\frac{\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\cdot\int_{\mathcal{D}} g(\theta)df(\theta) - \int_{\mathcal{D}} df(\theta)\cdot\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)}{\int_{\mathcal{D}} df(\theta)[\int_{\mathcal{D}} df(\theta) + \int_{\mathcal{R}\setminus\mathcal{D}}\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)]}\right|\end{aligned}$$

where the second equality follows from combining the fractions and making use of (19). We can bound the denominator from below by $C_{\mathcal{D}}^2 = \big[\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\big]^2 > 0$ so that

$$\begin{aligned}&\left|\mathbb{E}[g(\theta)|\theta\in\mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right|\\ &\leq \frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\cdot\int_{\mathcal{D}} g(\theta)df(\theta) - \int_{\mathcal{D}} df(\theta)\cdot\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|}{C_{\mathcal{D}}^2}\end{aligned}$$

If we add and subtract

$$\int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)\cdot\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)$$

within the numerator, we can apply the triangle inequality. Thus,

$$\begin{aligned}&\left|\mathbb{E}[g(\theta)|\theta\in\mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right|\\ &\leq \frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|\cdot\left|\int_{\mathcal{D}} g(\theta)df(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|}{C_{\mathcal{D}}^2}\\ &\qquad + \frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|\cdot\left|\int_{\mathcal{D}} df(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}}\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|}{C_{\mathcal{D}}^2}\end{aligned}$$

Since $g \in \mathbb{L}^1(\mathcal{R}, \mathcal{L}(\theta; Y)\pi_{\mathcal{R}} d\mu_{\mathcal{R}})$, we can then bound the numerators as follows. First,

$$\begin{aligned}
&\left| \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right| \cdot \left| \int_{\mathcal{D}} g(\theta) df(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta) \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right| \\
&\leq \left| \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right| \cdot \left(\left| \int_{\mathcal{D}} g(\theta) df(\theta) \right| + \left| \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta) \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right| \right) \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \cdot \left(\int_{\mathcal{D}} |g(\theta)| df(\theta) + \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right) \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \cdot \int_{\mathcal{R}} |g(\theta)| df(\theta) = C_{\mathcal{R}} \mathbb{E}|g(x_i)| \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta).
\end{aligned}$$

Here, $C_{\mathcal{R}} = \int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta)$ is the normalizing constant of $\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)$. Secondly,

$$\begin{aligned}
&\left| \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta) \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right| \cdot \left| \int_{\mathcal{D}} df(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right| \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \cdot \left| \int_{\mathcal{D}} df(\theta) \right| + \left| \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right| \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \left(\int_{\mathcal{D}} df(\theta) + \int_{\mathcal{R}\setminus\mathcal{D}} df(\theta) \right) \\
&= C_{\mathcal{R}} \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta).
\end{aligned}$$

Thus, we have the bounds specified by the theorem,

$$\begin{aligned}
&\left| \mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)] \right| \\
&\leq \frac{C_{\mathcal{R}} \mathbb{E}|g(\theta)| \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)}{C_{\mathcal{D}}^2} + \frac{C_{\mathcal{R}} \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)}{C_{\mathcal{D}}^2} \\
&= \frac{C_{\mathcal{R}} \int_{\mathcal{R}\setminus\mathcal{D}} (\mathbb{E}|g(\theta)| + |g(\theta)|) \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2}.
\end{aligned}$$

It remains to be shown that

$$\left| \mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)] \right| \to 0 \text{ as } \lambda \to 0^+.$$

Again, by the assumptions that $g \in \mathbb{L}^1(\mathcal{R}, \mathcal{L}(\theta; Y)\pi_{\mathcal{R}} d\mu_{\mathcal{R}})$ and $\|v_{\mathcal{D}}(\theta)\| > 0$ for $\mu_{\mathcal{R}}$ a.e. $\theta \in \mathcal{R} \setminus \mathcal{D}$, it follows that $(E|g(x_i)|+|g(x_i)|)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)$ is a dominating function of $(E|g(\theta)|+|g(\theta)|)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)$ which converges to zero for $\mu_{\mathcal{R}}$-a.e. $\theta \in \mathcal{R} \setminus \mathcal{D}$ as $\lambda \to 0^+$. Thus, by the dominated convergence theorem, $\left| \mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)] \right| \to 0$ as $\lambda \to 0^+$.

□

Proof. Proof of Theorem 1

We begin with the bound from Lemma 1.

$$\left|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| \leq \frac{C_{\mathcal{R}} \int_{\mathcal{R}\setminus\mathcal{D}} (\mathbb{E}|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2}.$$

For the moment, let us focus on the numerator of the previous expression. By the Cauchy-Schwartz inequality,

$$C_{\mathcal{R}} \int_{\mathcal{R}\setminus\mathcal{D}} (\mathbb{E}|g(\theta)| + |g(\theta)|) \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

$$\leq C_{\mathcal{R}} \left(\int_{\mathcal{R}\setminus\mathcal{D}} (\mathbb{E}|g(\theta)| + |g(\theta)|)^2 df(\theta) \right)^{1/2} \left(\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right)^{1/2}$$

By assumption, $g \in \mathbb{L}^2(\mathcal{R}, \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}\mu_{\mathcal{R}})$. Thus,

$$C_{\mathcal{R}} \int_{\mathcal{R}\setminus\mathcal{D}} (\mathbb{E}|g(\theta)| + |g(\theta)|) \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

$$= \underbrace{\left(3C_{\mathcal{R}}^2 (\mathbb{E}|g|)^2 + C_{\mathcal{R}} \mathbb{E}[|g|^2] \right)^{1/2}}_{C_g} \left(\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) \right)^{1/2} = C_g \left(\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right)^{1/2}$$

We separate the integral

$$\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

over the sets $\{\theta : \|v_{\mathcal{D}}(\theta)\| > -\lambda \log \lambda\}$ and $\{\theta : 0 < \|v_{\mathcal{D}}(\theta)\| < -\lambda \log \lambda\}$.

$$\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

$$= \int_{\{\theta : \|v_{\mathcal{D}}(\theta)\| > -\lambda \log \lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) + \int_{\{\theta : 0 < \|v_{\mathcal{D}}(\theta)\| < -\lambda \log \lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

$$\leq \lambda^2 \int_{\{\theta : \|v_{\mathcal{D}}(\theta)\| > -\lambda \log \lambda\}} df(\theta) + \int_{\{\theta : 0 < \|v_{\mathcal{D}}(\theta)\| < -\lambda \log \lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

$$\leq C_{\mathcal{R}} \lambda^2 + \int_{\{\theta : 0 < \|v_{\mathcal{D}}(\theta)\| < -\lambda \log \lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

To review, to this point we have shown that

$$\left|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| \leq \frac{C_g}{C_{\mathcal{D}}^2} \left(C_{\mathcal{R}} \lambda^2 + \int_{\{\theta : 0 < \|v_{\mathcal{D}}(\theta)\| < -\lambda \log \lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right)^{1/2} \tag{20}$$

From the requirements of Theorem 1, we now let $\|v_{\mathcal{D}}(\theta)\| = \inf_{x \in \mathcal{D}} ||\theta - x||_2$ and assume that $\mathcal{D}$ has a piecewise smooth boundary. In this case, the set $\{\theta : 0 < \|v_{\mathcal{D}}(\theta)\| < -\lambda \log \lambda\}$ forms a 'shell' of thickness $-\lambda \log \lambda$ which encases $\mathcal{D}$.

For the moment, suppose that $\mathcal{D}$ is a bounded subset of $\mathcal{R}$. Furthermore, suppose we take λ sufficiently small so that $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is continuous on $V_\lambda = \{\theta : 0 < \|v_{\mathcal{D}}(\theta)\| < -\lambda \log \lambda\}$. Observe that

$$\int_{V_\lambda} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \leq \sup_{V_\lambda} |\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| \int_{V\lambda} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) d\mu_R(\theta)$$

$$\leq \sup_{V_\lambda} |\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| \int_{V\lambda} d\mu_R(\theta) = \sup_{V_\lambda} |\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| \cdot Vol(V_\lambda)$$

$$\approx \sup_{V_\lambda} |\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| S_{\mathcal{D}} \cdot \lambda |\log \lambda|$$

Here, $S_{\mathcal{D}}$ is the surface area of boundary of $\mathcal{D}$, which is finite by the assumptions that $\mathcal{D}$ is bounded and has a piecewise smooth boundary. Additionally, since V_λ is relatively compact, it follows that $\sup_{V_\lambda} |\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| < \infty$.

Consider the more general case where $\mathcal{D}$ is not a bounded subset of $\mathcal{R}$. Note that, for $\theta \in V_\lambda$, $J(v_{\mathcal{D}}(\theta)) = \sqrt{(Dv_{\mathcal{D}})'(Dv_{\mathcal{D}})} = 2$. By the co-area formula Diaconis et al. (2013); Federer (2014)

$$\int_{V_\lambda} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta) = \int_0^{-\lambda \log \lambda} e^{-\frac{x}{\lambda}} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta) d\bar{\mathcal{H}}^{r-1}(\theta) \right) dx$$

Again, we may take λ sufficiently small so that $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is continuous on V_λ. As such, the function $\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta) d\bar{\mathcal{H}}^{r-1}(\theta)$ is a continuous map from the closed interval, $[0, -\lambda \log \lambda]$, to $\mathbb{R}$. Hence it is bounded. As a result,

$$\int_{\{\theta : 0 < \|v_{\mathcal{D}}(\theta)\| < -\lambda \log \lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta)$$

$$\leq \sup_{x \in [0, -\lambda \log \lambda]} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta) d\bar{\mathcal{H}}^{r-1}(\theta) \right) \cdot \int_0^{-\lambda \log \lambda} e^{-\frac{x}{\lambda}} dx$$

$$= \sup_{x \in [0, -\lambda \log \lambda]} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta) d\bar{\mathcal{H}}^{r-1}(\theta) \right) \cdot (\lambda - \lambda^2) = O(\lambda)$$

This result also applies to the case where $\mathcal{D}$ is bounded. Thus, we may conclude that

$$\left| \mathbb{E}[g(\theta) | \theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)] \right|$$

$$\leq \frac{C_g}{C_{\mathcal{D}}^2} \left(C_{\mathcal{R}} \lambda^2 + \sup_{x \in [0, -\lambda \log \lambda]} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta) d\bar{\mathcal{H}}^{r-1}(\theta) \right) \cdot (\lambda - \lambda^2) \right)^{1/2}$$

$$= \frac{C_g}{C_{\mathcal{D}}^2} \cdot \sup_{x \in [0, -\lambda \log \lambda]} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta) d\bar{\mathcal{H}}^{r-1}(\theta) \right) \cdot \sqrt{\lambda} + o(\sqrt{\lambda})$$

Since $\sup_{x\in[0,-\lambda\log\lambda]}\left(\int_{v_{\mathcal{D}}^{-1}(x)}\frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right)$ is a decreasing function in λ, it follows that

$$\left|\mathbb{E}[g(\theta)|\theta\in\mathcal{D}]-\mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right|=O(\sqrt{\lambda})$$

as $\lambda\to 0^+$. □

B Proofs from Section 3.2

Proof. Recall that we have two densities. The first is the fully constrained density for $\theta\in\mathcal{D}$.

$$\pi_{\mathcal{D}}(\theta)=\frac{1}{m_0}\frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))}\mathbb{1}_{\mathcal{D}}(\theta)$$

where the normalizing constant m_0 is calculated w.r.t. Hausdorff measure

$$m_0=\int_{\mathcal{R}}\frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))}\mathbb{1}_{\mathcal{D}}(\theta)d\bar{\mathcal{H}}^{r-s}(\theta).$$

Secondly, we have the relaxed distribution

$$\tilde{\pi}_{\mathcal{D}}(\theta)=\frac{1}{m_\lambda}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{||\nu(\theta)||_1}{\lambda}\right)$$

where the normalizing constant is calculated w.r.t. Lesbesgue measure on $\mathcal{R}$, denoted by $\mu_{\mathcal{R}}$,

$$m_\lambda=\int_{\mathcal{R}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{||\nu(\theta)||_1}{\lambda}\right)d\mu_{\mathcal{R}}(\theta).$$

For a given function, $g:\mathcal{R}\to\mathbb{R}$, we can define the exact and approximate expectations of g, respectively $\mathbb{E}_\Pi$ and $\mathbb{E}_{\tilde{\Pi}}$, as

$$\begin{aligned}
\mathbb{E}_\Pi[g(\theta)]=\mathbb{E}[g(\theta)|\theta\in\mathcal{D}]&=\int_{\mathcal{R}}\frac{g(\theta)}{m_0}\frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))}\mathbb{1}_{\mathcal{D}}(\theta)d\bar{\mathcal{H}}^{r-s}(\theta)\\
&=\int_{\mathcal{D}}\frac{g(\theta)}{m_0}\frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))}d\bar{\mathcal{H}}^{r-s}(\theta)\\
\mathbb{E}_{\tilde{\Pi}}[g(\theta)]&=\int_{\mathcal{R}}\frac{g(\theta)}{m_\lambda}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{||\nu(\theta)||_1}{\lambda}\right)d\mu_{\mathcal{R}}(\theta)\\
&=\int_{\mathbb{R}^s}\frac{1}{m_\lambda}\int_{\nu^{-1}(x)}g(\theta)\frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y)}{J(\nu(\theta))}\exp\left(-\frac{||\nu(\theta)||_1}{\lambda}\right)d\bar{\mathcal{H}}^{r-s}(\theta)d\mu_{\mathbb{R}^s}\\
&=\int_{\mathbb{R}^s}\frac{\exp\left(-\frac{||x||_1}{\lambda}\right)}{m_\lambda}\int_{\nu^{-1}(x)}g(\theta)\frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y)}{J(\nu(\theta))}d\bar{\mathcal{H}}^{r-s}(\theta)d\mu_{\mathbb{R}^s}
\end{aligned}$$

Let,

$$m(x) = m^{r-s}(x) = \int_{\nu^{-1}(x)} \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta).$$

By construction, $m(x) > 0$ for $\mu_{\mathbb{R}^s}$-a.e. $x \in \text{Range}(\nu)$. In particular, $m_0 = m(0) > 0$. By Theorem 1,

$$\mathbb{E}[g(\theta)|\nu(\theta) = x] = \frac{1}{m(x)} \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta). \tag{21}$$

As such, we may may express $\mathbb{E}_{\tilde{\Pi}}[g(\theta)]$ as

$$\mathbb{E}_{\tilde{\Pi}}[g(\theta)] = \int_{\mathbb{R}^s} \frac{m(x)}{m_\lambda} \exp\left(- \frac{||x||_1}{\lambda} \right) \mathbb{E}\big[g(\theta)|\nu(\theta) = x\big] d\mu_{\mathbb{R}^s}(x). \tag{22}$$

Let us first consider the small λ behavior of m_λ. We begin by re-expressing m_λ in terms of $m(x)$ through the co-area formula.

$$\begin{aligned}
m_\lambda &= \int_{\mathcal{R}} \pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y) \exp\left(- \frac{||\nu(\theta)||_1}{\lambda} \right) d\mu_{\mathcal{R}}(\theta) \\
&= \int_{\mathbb{R}^s} \exp\left(- \frac{||x||_1}{\lambda} \right) \int_{\nu^{-1}(x)} \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y)}{J(v(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x) \\
&= \int_{\mathbb{R}^s} m(x) \exp\left(- \frac{||x||_1}{\lambda} \right) d\mu_{\mathbb{R}^s}(x)
\end{aligned}$$

Split the above integral into two regions: the interior and exterior of $B_1(0; \lambda|\log(\lambda^{s+1})|)$. Note that outside of B_1, $\exp(-||x||_1/\lambda) \leq \lambda^{s+1}$.

$$\begin{aligned}
m_\lambda &= \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \exp\left(- \frac{||x||_1}{\lambda} \right) d\mu_{\mathbb{R}^s}(x) + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \exp\left(- \frac{||x||_1}{\lambda} \right) d\mu_{\mathbb{R}^s}(x) \\
&= O\bigg(\lambda^{s+1} \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) d\mu_{\mathbb{R}^s}(x) \bigg) \\
&\qquad + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \bigg[1 + O\bigg(\frac{1}{\lambda} \exp\left(- \frac{||x||_1}{\lambda} \right) \bigg) \bigg] d\mu_{\mathbb{R}^s}(x) \\
&= O\bigg(\lambda^{s+1} \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) d\mu_{\mathbb{R}^s}(x) \bigg) \\
&\qquad + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \Big[1 + O(\lambda^s) \Big] d\mu_{\mathbb{R}^s}(x)
\end{aligned}$$

Since $m(x)$ is continuous on an open neighborhood containing the origin, we may choose λ small enough so

that $m(x)$ is uniformly continuous on $B_1(0;\lambda|\log\lambda^{s+1}|)$. Then,

$$m_\lambda = O\Big(\lambda^{s+1}\Big) + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} [m(0)+o(1)][1+O(\lambda^s)] d\mu_{\mathbb{R}^s}(x)$$

$$= O(\lambda^{s+1}) + [m(0)+o(1)][1+O(\lambda^s)] \underbrace{\frac{|2(s+1)\lambda\log\lambda|^s}{\Gamma(s+1)}}_{Vol(B_1(0;\lambda|\log(\lambda^{s+1})|)}$$

$$= m(0)\frac{|2(s+1)\lambda\log\lambda|^s}{\Gamma(s+1)} + o(|\lambda\log\lambda|^s)$$

at leading order as $\lambda \to 0^+$.

We now turn to the small λ behavior of $\tilde{\mathbb{E}}[g(\theta)]$. Again, we may choose λ sufficient small so that both

$$m(x) = \int_{\nu^{(-1)}(x)} \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta)$$

$$G(x) = \int_{\nu^{(-1)}(x)} g(\theta)\frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) = m(x)\mathbb{E}[g|\nu(\theta)=x]$$

are continuous on $B_1(0;\lambda|\log\lambda^{s+1}|)$ and hence uniformly continuous at $x=0$.

Similar to the study of m_λ, separate the $\mathbb{E}_{\tilde{\Pi}}[g(\theta)]$ into integrals over the interior and exterior of $B_1(0,\lambda|\log(\lambda)^{s+1}|)$. Again, we assume λ is taken to be sufficiently small so that both $m(x)$ and $G(x)$ are uniformly continuous on B_1. Then

$$\mathbb{E}_{\tilde{\Pi}}[g(\theta)] = \int_{\mathbb{R}^s} \frac{m(x)}{m_\lambda} \exp\Big(-\frac{||x||_1}{\lambda}\Big)\mathbb{E}[g(\theta)|\nu(\theta)=x] d\mu_{\mathbb{R}^s}(x)$$

$$= \int_{\mathbb{R}^s} \frac{1}{m_\lambda} \exp\Big(-\frac{||x||_1}{\lambda}\Big) \int_{\nu^{(-1)}(x)} \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x)$$

$$= \int_{\mathbb{R}^s\setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} \frac{1}{m_\lambda} \exp\Big(-\frac{||x||_1}{\lambda}\Big) \int_{\nu^{(-1)}(x)} \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x)$$

$$+ \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} \frac{1}{m_\lambda} \exp\Big(-\frac{||x||_1}{\lambda}\Big) \int_{\nu^{(-1)}(x)} \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x)$$

$$= O\Big(\frac{\lambda^{s+1}}{m_\lambda}\Big) + \int_{B_1} \frac{m(0)+o(1)}{m_\lambda}(1+O(\lambda^s))\Big(\mathbb{E}[g(\theta)|\nu(\theta)=0] + o(1)\Big) d\mu_{\mathbb{R}^s}(x)$$

$$= O\Big(C\mathbb{E}|g|\frac{\lambda}{|\log\lambda|^s}\Big) + \mathbb{E}[g(\theta)|\theta\in\mathcal{D}] + o(1).$$

And we may conclude that

$$\Big|\mathbb{E}[g|\theta\in\mathcal{D}] - \mathbb{E}_{\tilde{\Pi}}[g]\Big| \to 0 \text{ as } \lambda\to 0^+.$$

The proof of the corollary follows by changing the $o(1)$ correction within the integrals over $B_1(0;\lambda|\log\lambda^{s+1}|)$

to $O(\lambda|\log\lambda^{s+1}|)$ corrections. As a result, the leading order approximation error is then $O(\lambda|\log\lambda|^{-s})$ as $\lambda \to 0^+$. □

C Approximation Error of von–Mises Fisher distribution

We test $\lambda = 10^{-3}$, 10^{-4} and 10^{-5} for CORE-HMC. Table 3 shows the effective sample size per 1000 iterations, the effective 'violation' $|v(\theta)| = |\theta_1^2+\theta_2^2-1|$ and the $\left|\mathbb{E}_{\Pi}[\sum_j \theta_j]-\mathbb{E}_{\tilde{\Pi}}[\sum_j \theta_j]\right|$ as the approximation error. As the approximation error is numerically computed, to provide a baseline error, we also compare two independent samples from the same exact distribution. The approximation error based on $\lambda = 10^{-5}$ approximation is indistinguishable from this low numerical error, while the other approximations have slightly larger error but more effective samples.

	HMC based on CORE			Exact
	$\lambda = 10^{-3}$	$\lambda = 10^{-4}$	$\lambda = 10^{-5}$	
$\left\|\mathbb{E}_{\Pi}[\sum_j \theta_j] - \mathbb{E}_{\tilde{\Pi}}[\sum_j \theta_j]\right\|$	0.025 (0.014, 0.065)	0.016 (0.012, 0.019)	0.008 (0.006,0.015)	0.009 (0.007,0.015)
$\lvert v_{\mathcal{D}}(\theta)\rvert$	9×10^{-4} $(2.6 \cdot 10^{-5}, 3.3 \cdot 10^{-3})$	9×10^{-5} $(2.0 \cdot 10^{-6}, 3.4 \cdot 10^{-4})$	9×10^{-6} $(2.7 \cdot 10^{-7}, 3.5 \cdot 10^{-5})$	0
ESS /1000 Iterations	751.48	260.54	57.10	788.30

Table 3: Benchmark of constraint relaxation methods on sampling von–Mises Fisher distribution on a unit circle. For each CORE, average approximation error (with 95% credible interval, out of 10 repeated experiments) is computed, and numeric error is shown under column 'exact' as comparing two independent copies from the exact distribution. Effective sample size shows CORE with relatively large λ have high computing efficiency.