Thicken, Sample, Project: Monte Carlo Computation for Implicit Statistical Models

David Kahle*

Jonathan D. Hauenstein†

Abstract

Implicit polynomial constraints arise throughout statistics, but practical computation on the resulting model spaces becomes difficult when those spaces are curved, positive dimensional, singular, or disconnected. In this paper, we study a workflow that turns such implicitly defined model spaces into objects that can be explored with modern Monte Carlo tools. The central idea is to thicken the constraint set into the ambient space, sample the resulting distribution with Hamiltonian Monte Carlo, and project sampled points back to the constraint set when exact feasibility is required. The resulting workflow supports exploratory sampling, constrained objective-function evaluation, synthetic benchmark generation for topological data analysis, and component discovery without requiring an explicit parameterization of the model space. Since projected draws do not inherit a canonical probability law on the variety, we clarify which computational tasks are still well served by the workflow and which require additional justification. We develop the variety-normal construction, discuss normalizability and tuning issues, and illustrate the approach on a contingency-table independence model, a persistent-homology benchmark built from a Lissajous curve, and a higher-dimensional multimodal polynomial system. Our emphasis is statistical computation: what the construction targets, how it can be implemented in practice, and where truncation, initialization, and projection materially affect performance.

1 Introduction

Statistical models defined by implicit polynomial constraints are common throughout statistics but rarely come with convenient computational tools for exploration. Independence models for contingency tables, conditional independence constraints in Gaussian graphical models, identifiability conditions for mixture models, and broader classes of algebraic statistical models [Sullivant, 2018, Drton et al., 2008] are all specified by polynomial equalities. In these cases analysts face computational tasks that go well beyond just root finding, including exploring the feasible region, evaluating an objective function over it, diagnosing disconnected components, generating representative constrained points, or producing synthetic point clouds near a structured implicit set for downstream procedures such as topological summaries or pattern recognition.

Standard computational tools struggle with these tasks since the space of interest is specified only implicitly, not as a parameterized surface or simple box. If $g_1, \ldots, g_m \in \mathbb{R}[\boldsymbol{x}]$, the polynomial ring in the n variables $x_1, \ldots, x_n$ with real coefficients, the corresponding real variety is

$$\mathcal{V}(g_1, \ldots, g_m) = \{\boldsymbol{x} \in \mathbb{R}^n : g_1(\boldsymbol{x}) = \cdots = g_m(\boldsymbol{x}) = 0\}.$$

Using vector notation we can express the same constructs as $\boldsymbol{g} \in \mathbb{R}[\boldsymbol{x}]^m$ and $\mathcal{V}(\boldsymbol{g}) = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0}_m\}$; this is a nonlinear algebraic system with m equations in n unknowns. Notice that over the reals such

*Associate Professor, Department of Statistical Science, Baylor University. david_kahle@baylor.edu.

†Robert and Sara Lumpkins Collegiate Professor, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame. hauenstein@nd.edu.

systems can be combined into a single polynomial via $g = \boldsymbol{g}'\boldsymbol{g} = \sum_{i=1}^{m} g_i^2$, so that $\mathcal{V}(g) = \mathcal{V}(g_1, \ldots, g_m)$. Real varieties are surprisingly very flexible geometric objects: they may be curved, positive dimensional, singular, or disconnected—precisely the geometric features that make generic constrained samplers and optimizers unreliable.

Several existing computational traditions address parts of this problem. Symbolic and numerical algebraic geometry methods, e.g., [Sturmfels, 2002, Sommese and Wampler, 2005, Bates et al., 2013, Breiding and Marigliano, 2020], are powerful tools for solving polynomial systems, certifying roots, and decomposing solution sets, but are less naturally suited to repeated stochastic exploration of positive-dimensional or disconnected feasible regions. Manifold MCMC methods, e.g., [Byrne and Girolami, 2013, Girolami and Calderhead, 2011, Lelièvre et al., 2019], can sample constrained spaces by working in local coordinates or projecting proposals onto level sets, but typically require an explicit parameterization or a well-behaved constraint Jacobian at every iterate. Bayesian computation has developed mature tools for sampling difficult unnormalized densities, especially using Gibbs sampling and Hamiltonian Monte Carlo (HMC) strategies [Lunn et al., 2000, Duane et al., 1987, Neal, 2011].

Our goal is to connect these traditions: convert an implicit polynomial description into an ambient-space density that HMC can sample directly, then recover constrained points by numerical projection when needed. The central idea is a three-step computational workflow: *thicken* the hard equality-constrained space into an ambient-space density, *sample* that density with HMC, and *project* sampled points back to the variety when exact feasibility is required. This thicken–sample–project workflow yields a practical pipeline for implicit model spaces even when explicit coordinates are unavailable or inconvenient, and it requires only the ability to evaluate the defining polynomials and their gradients, a relatively low bar.

An important caveat: the paper studies this workflow rather than a canonical measure on the variety itself. The distributions introduced here are designed to concentrate near the target variety and to be computationally tractable. When projected points are used for optimization or constrained exploration, the method is often already useful. However, projected draws should not automatically be interpreted as samples from a canonical "uniform" law on the variety, and inferential claims that depend on such a law require separate justification.

Our contributions are fourfold.

1. **Construction.** We formalize variety-normal and induced variety distributions that convert an implicit polynomial description into an ambient-space probability model amenable to standard samplers.
2. **Sampling and projection.** We show how HMC can sample these distributions and how endgame-style numerical projection can recover points on the constrained set when exact feasibility is needed.
3. **Practical diagnostics.** We identify computational issues that govern the workflow in practice—normalizability, truncation, near-root behavior, initialization, and the interaction between projection and additional statistical constraints such as positivity—and provide guidance for practitioners.
4. **Worked examples.** We demonstrate the workflow on three representative tasks: constrained sampling and objective evaluation on a contingency-table independence model, synthetic benchmark generation for topological data analysis, and component discovery on a higher-dimensional multimodal variety.

Our focus throughout is statistical computing: what the constructions target, how they can be implemented in practice, and where tuning decisions materially affect performance.

The remainder of the paper is organized as follows. Section 2 introduces the variety-normal construction and the geometric intuition behind it. Section 3 broadens the construction to induced variety distributions. Section 4 describes rejection sampling, HMC, and projection to the variety. Section 5 presents computational examples. Section 6 closes with limitations, implementation guidance, and directions for further work.

2 Variety-Normal Constructions

In this section, we construct a family of probability distributions that focus their mass on a variety of interest based on the normal and multivariate normal distributions. We begin with an observation about

the univariate normal distribution that is later generalized greatly.

2.1 The variety normal distribution of a single polynomial

The univariate normal distribution is characterized by a probability density function (PDF) with respect to the Lebesgue measure on $\mathbb{R}$ of the form $p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are given quantities. Disregarding normalization factors, the normal density is proportional to the exponential of a negative quadratic form, written $p(x|\mu,\sigma^2) \propto \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$ or $\widetilde{p}(x|\mu,\sigma^2) = \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$, where the tilde is used to express an un-normalized density, which we sometimes refer to as a pseudodensity.

The normal distribution focuses its mass on μ, not merely in the sense that if $X \sim p(x|\mu,\sigma^2)$ the expected value $\mathbb{E}[X] = \mu$, but also and perhaps more fundamentally in the sense that the mode of the distribution of X is μ and the density exponentially and symmetrically decays as one moves away from μ. The dispersion parameter σ^2, the variance of X, describes the extent to which the distribution is concentrated on μ: the closer that σ^2 is to 0, the more focused the distribution is on μ. Importantly, the normal distribution focuses its mass where the quantity in the exponent vanishes. In the case of the univariate normal, this is the root of the linear polynomial $g = x - \mu \in \mathbb{R}[x]$. The variety is zero-dimensional: $\mathcal{V}(g) = \{\mu\}$.

We note in passing that the $\frac{1}{2}$ can be absorbed into the polynomial $p(x|\mu,\sigma^2) \propto \exp\{-\frac{1}{\sigma^2}\left(\frac{1}{\sqrt{2}}x - \frac{1}{\sqrt{2}}\mu\right)^2\}$ so that the coefficient vector of the polynomial lies on the unit sphere (here circle). This might have some benefit in providing a canonical representation of the polynomial; however, as we are not here interested in the parameter space but rather the individual distribution itself given g, and since the representation using the standard normal density has benefits later on, we leave the description as-is.

The observation that the normal distribution concentrates its mass near the variety of the linear polynomial $g = x - \mu \in \mathbb{R}[x]$ suggests that the same is true for an arbitrary multivariate polynomial $g(\boldsymbol{x}|\boldsymbol{\beta}) \in \mathbb{R}[\boldsymbol{x}]$: if one wishes to construct a distribution that places its mass near $\mathcal{V}(g)$, simply swap $x - \mu$ for g. This brings us to the heteroskedastic variety normal distribution.

Definition 1. *A random vector $\boldsymbol{X} \in \mathbb{R}^n$ is said to have the heteroskedastic variety normal (HVN) distribution, denoted $\boldsymbol{X} \sim \mathfrak{N}_n\left(g,\sigma^2\right)$, if it admits a density*

$$p(\boldsymbol{x}|g,\sigma^2) \propto \widetilde{p}(\boldsymbol{x}|g,\sigma^2) := \exp\left\{-\frac{g(\boldsymbol{x}|\boldsymbol{\beta})^2}{2\sigma^2}\right\} \tag{1}$$

with respect to the Lebesgue measure on $\mathbb{R}^n$ for some $g(\boldsymbol{x}|\boldsymbol{\beta}) \in \mathbb{R}[\boldsymbol{x}]$ and $\sigma^2 > 0$. $\boldsymbol{X}$ is said to have the truncated HVN distribution, denoted $\boldsymbol{X} \sim \mathfrak{N}_{n,\boldsymbol{\mathcal{X}}}\left(g,\sigma^2\right)$, if (1) only holds on some non-null $\boldsymbol{\mathcal{X}} \subset \mathbb{R}^n$, outside of which p vanishes.

The conditional notation $|g$, indicating g is given, is unconventional since it incorporates both the coefficients $\boldsymbol{\beta}$ of the polynomial, which are parameters, and the variates $\boldsymbol{x}$ themselves. Nevertheless, for the present purposes, the notation $|g$ comports a clarity that the vector of coefficients does not, so we find it to be a useful device to communicate intent.

Just as the univariate normal distribution focuses its mass on μ, the HVN focuses its mass on the variety $\mathcal{V}(g)$: the $-g^2$ in the exponent demands that the pseudodensity is maximized at the variety (if nonempty) and from there decreases–locally, at least–so that the variety is the mode of the distribution. Indeed, the HVN can be thought of as a blurred version of a variety where σ^2 governs the amount of blurring. Probabilistically, σ^2 is intended to gauge the likelihood of points falling a given distance away from the variety. In the simple univariate normal case, the variety is linear and zero-dimensional so that the empirical rule helps us interpret $\sigma = \sqrt[+]{\sigma^2}$. The general case is more complex. An example of the bivariate nonlinear case is illustrated in Figure 1 with $g = x^2 + y^2 - 1 \in \mathbb{R}[x,y]$. Figure 1 suggests the scheme works perfectly; however, a few considerations meter this enthusiasm: normalizability and the role of σ^2. We start with the role of σ^2 and defer the discussion concerning normalizability to Section 2.2.

Coming from the normal distribution, it feels natural to want to interpret σ^2 as gauging the amount of variability uniformly across the variety in the sense that any two points in $\mathbb{R}^n$ at the same distance

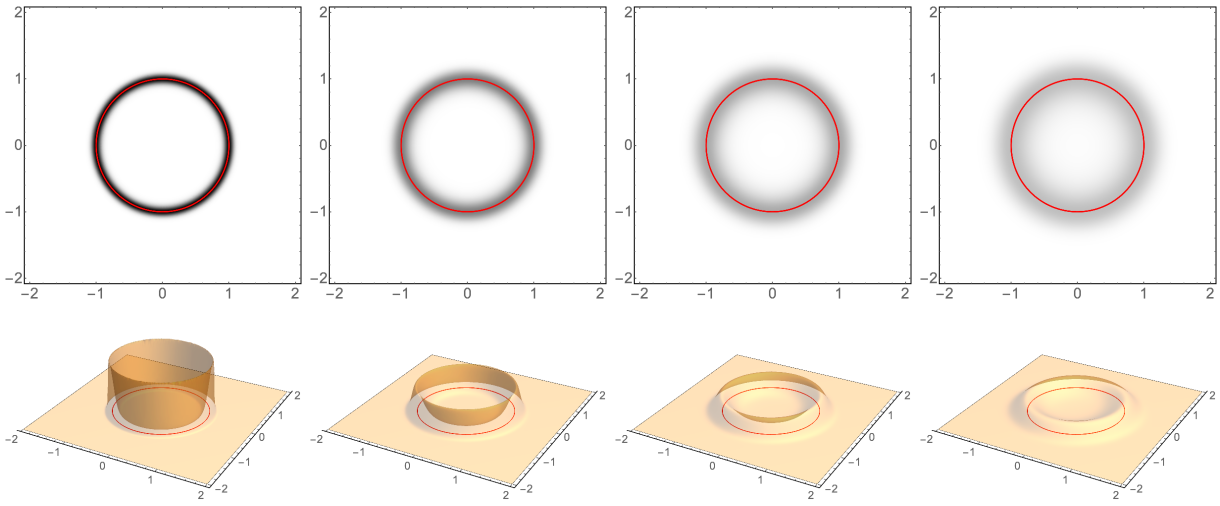

Figure 1: (Left to right.) $\mathfrak{N}_2\left(x^2+y^2-1, \sigma^2\right)$ from (1) with $\sigma = .1$, .2, .3, and .4; $\mathcal{V}(g)$ in red.

from the variety should have the same likelihood of occurring, at least approximately. By distance here we simply mean the Euclidean distance from the points to the nearest point(s) on the variety. Unfortunately, however, the HVN distribution does not obey this basic intuition, as is easily seen with the polynomial $g = y^2 - (x^3 + x^2) \in \mathbb{R}[x, y]$, whose variety is the alpha curve displayed in red in Figure 2. This explains the adjective heteroskedastic, which does not here refer to variability changing on account of another variable *per se* but across the variety itself. It is the phenomenon where two points, similarly situated with respect to the variety, can have quite different likelihoods of occurring.

Why does this happen? Recalling that varieties are zero level sets, the HVN distribution can be thought of as shifting the graph of g, a surface in $\mathbb{R}^3$, up or down by a random amount following a $\mathcal{N}\left(0, \sigma^2\right)$ distribution and selecting a point on the resulting zero level set. However, the shifting has varying effects across the variety: equal amounts of shifting do not result in equal amounts of movement of the zero level set across the intersection with the xy-plane. Figure 2 illustrates this with five equally spaced level sets of g.

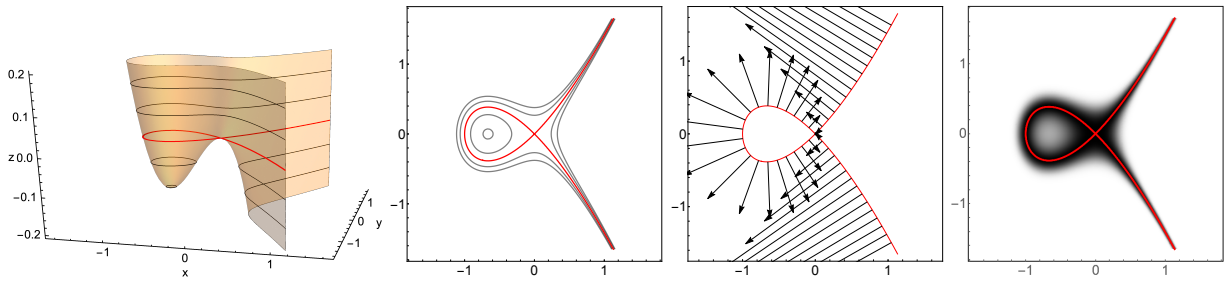

Figure 2: Equidistant level sets of $g = y^2 - (x^3 + x^2)$ do not result equidistant contours. From left to right: The graph of g with equally spaced slices in z; those intersections projected down into the xy-plane; the gradient vector field on the variety; and the density plot of the HVN with $\sigma = .1$.

This is in fact a general feature of level sets related to the inverse function theorem: parts of the variety where the function changes rapidly exhibit small changes, and places where it changes gradually exhibit large changes. This basic phenomenon is easily seen in univariate linear and quadratic polynomials; see Figure 3. In one dimension, for a nonconstant univariate linear polynomial $g = \beta_0 + \beta_1 x \in \mathbb{R}[x]$, $\mathcal{V}(g) = \{-\beta_0/\beta_1\}$.

Shifting the graph of g up by ϵ moves the root from $-\beta_0/\beta_1$ to $-\beta_0/\beta_1 - \epsilon/\beta_1$, i.e. an amount proportional to β_1^{-1} and in the direction of the opposite of the sign of β_1. Shifting it down by ϵ moves it the same amount, but in the other direction–the direction of the sign of β_1, the derivative/slope of g.

In general, root mobility is inversely related to the magnitude of the gradient and in the opposite direction if the shift is up $(+\epsilon)$ or the same direction if the shift is down $(-\epsilon)$. This is because the gradient $\nabla_{\boldsymbol{x}} g$ points in the direction of steepest ascent. Hence, near the variety where g is 0, shifting the graph up requires a negative offset to balance to 0, and this occurs in the direction of the negative gradient, assuming simple roots of course–otherwise the gradient vanishes. Similarly, shifting the graph down moves the root in the direction of the gradient, the direction that would be required to move the value of the function g from 0 to a sufficiently positive value to offset the shift down. Curvature of the graph, a departure from linearity, affects an asymmetry in the magnitude of root mobility in same-sized upward and downward shifts, but this asymmetry is negligible for sufficiently small shifts up and down. In the current setting, this phenomenon can be observed at different points on the same variety, e.g., the images of Figure 2. The varying amount of change in the gradient's magnitude results in a disproportionate amount of the distribution's mass being placed on areas where the surface near the variety is relatively flat.

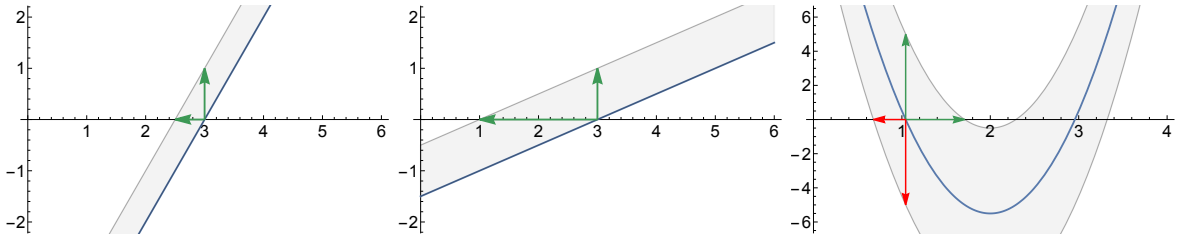

Figure 3: For a fixed amount of shifting up/down, varieties of rapidly changing functions exhibit less change than those of slowly changing functions. Curvature modulates this change in different directions.

The chief problem with the HVN distribution is that since $g(\boldsymbol{x}|\boldsymbol{\beta})$ changes at different rates across the variety, the volume traversed in the ambient space $\mathbb{R}^n$ by the surface as it is shifted up and down is uneven across the variety, and so the HVN distribution does not faithfully represent the notion that σ^2 gauges proximity to the variety in a constant way across the whole of the variety. Depending on the context and use case, this may be considered a feature of the HVN; however, for our purposes we consider it a defect.

One solution to this problem is to locally linearize the surface to a constant scale by normalizing g by the size of its gradient; that is, to use $\overline{g} = \overline{g}(\boldsymbol{x}|\boldsymbol{\beta}) = \frac{g}{||\nabla_{\boldsymbol{x}} g||}$ instead of g in place of the linear form $x - \mu$ in the normal density. (Norms in this article are assumed to be the 2-norm.) This can be justified by expanding g in a Taylor series about a point $\boldsymbol{x} \in \mathcal{V}(g) \subset \mathbb{R}^n$. Since $g(\boldsymbol{x}) = 0$,

$$g(\boldsymbol{x} + \boldsymbol{h}) = \nabla g(\boldsymbol{x})' \boldsymbol{h} + o(||\boldsymbol{h}||) = ||\nabla g(\boldsymbol{x})||\, ||\boldsymbol{h}|| \cos\theta + o(||\boldsymbol{h}||),$$

where θ is the angle between the gradient $\nabla g(\boldsymbol{x})$ and $\boldsymbol{h}$. Dividing by the norm of the gradient provides

$$\frac{g(\boldsymbol{x} + \boldsymbol{h})}{||\nabla g(\boldsymbol{x})||} = ||\boldsymbol{h}|| \cos\theta + o(||\boldsymbol{h}||),$$

where the o term is unaffected since $||\nabla g(\boldsymbol{x})||$ is a constant with respect to $\boldsymbol{h}$. This of course assumes the gradient is nonzero. Note two facts: 1) where $||\nabla g|| \neq 0$, the zero locus of $\overline{g}$ is precisely that of g, and 2) $||\nabla g|| = 0$ on a set of Lebesgue measure 0, since $||\nabla_{\boldsymbol{x}} g|| = 0 \iff ||\nabla_{\boldsymbol{x}} g||^2 = 0$, and the latter only occurs on a set of measure 0 since $||\nabla_{\boldsymbol{x}} g||^2$ is a nonzero polynomial (i.e., a nonzero polynomial vanishes on at most a set of Lebesgue measure zero).[1] As PDFs are only defined up to equivalence classes that allow for modifications on sets of measure zero, the un-normalized density that uses the normalized expression $\overline{g}$

[1] In this article, g is always nonconstant.

is, up to normalizability at least, properly defined. Analogues to Figure 2 using the normalized version of g are presented in Figure 4.

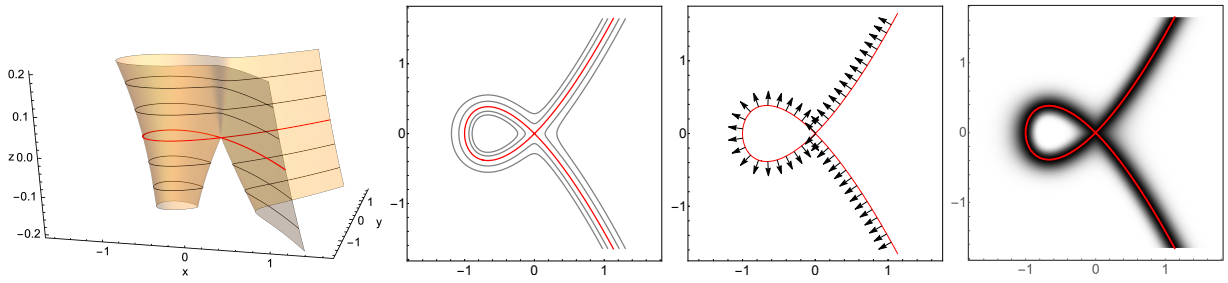

Figure 4: Equidistant level sets of $\overline{g} = g/\left|\left|\nabla_{\boldsymbol{x}} g\right|\right|$ result in approximately equidistant contours ($\sigma = .1$). Vectors drawn with 2σ magnitude.

With these things in mind, we are now able to present the definition we adopt in this work as the variety normal distribution.

Definition 2. *A random vector $\boldsymbol{X} \in \mathbb{R}^n$ is said to have the (homoskedastic) variety normal (VN) distribution, denoted $\boldsymbol{X} \sim \mathcal{N}_n\left(g, \sigma^2\right)$, if it admits a density*

$$p(\boldsymbol{x}|g,\sigma^2) \propto \widetilde{p}(\boldsymbol{x}|g,\sigma^2) := \exp\left\{-\frac{\overline{g}(\boldsymbol{x}|\boldsymbol{\beta})^2}{2\sigma^2}\right\} = \exp\left\{-\frac{1}{2\sigma^2}\left(\frac{g(\boldsymbol{x}|\boldsymbol{\beta})}{||\nabla_{\boldsymbol{x}} g(\boldsymbol{x}|\boldsymbol{\beta})||}\right)^2\right\} \tag{2}$$

with respect to the Lebesgue measure on $\mathbb{R}^n$ for some $g(\boldsymbol{x}|\boldsymbol{\beta}) \in \mathbb{R}[\boldsymbol{x}]$ such that $\overline{g}(\boldsymbol{x}|\boldsymbol{\beta}) = \frac{g(\boldsymbol{x}|\boldsymbol{\beta})}{||\nabla_{\boldsymbol{x}} g(\boldsymbol{x}|\boldsymbol{\beta})||}$ and $\sigma^2 > 0$. $\boldsymbol{X}$ is said to have the truncated (homoskedastic) variety normal distribution, which is denoted $\boldsymbol{X} \sim \mathcal{N}_{n,\boldsymbol{\mathcal{X}}}\left(g, \sigma^2\right)$, if (2) only holds on some non-null $\boldsymbol{\mathcal{X}} \subset \mathbb{R}^n$, outside of which p vanishes.

We present density plots for several bivariate variety normal distributions in Figure 5.

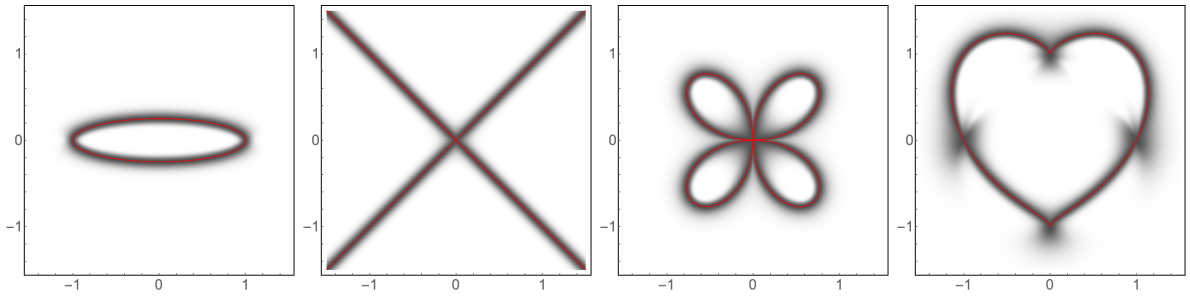

Figure 5: From left to right, density plots of $\mathcal{N}_{2,\boldsymbol{\mathcal{X}}}\left(\overline{g}, \sigma^2\right)$, truncated to the window, with $g = x^2 + (4y)^2 - 1$, $(y - x)(y + x)$, $(x^2 + y^2)^3 - 4x^2y^2$, and $(x^2 + y^2 - 1)^3 - x^2y^3$ and $\sigma = .05$. Aesthetic scales are consistent across the images. Note abnormalities at singularities in the last image.

A helpful way to think of these distributions is as probabilistic thickenings of the variety into the ambient space $\mathbb{R}^n$. The thickening expands all components of the variety, regardless of dimension, into the full dimension of the ambient space. Moreover, the thickening is locally uniform across the variety in the sense that, away from singularities, two points equidistant from the variety have arbitrarily close probability density provided σ^2 is selected sufficiently small. By construction, probability density decreases exponentially as one moves orthogonally away from the variety–in the direction parallel to the gradient. It is also approximately

symmetric, with asymmetries arising from the curvature in $g(\boldsymbol{x}|\boldsymbol{\beta})$ relative to the size of σ^2, the same effect as seen in Figure 3 (right). A three-dimensional example using the Whitney umbrella, as well as several geometric perspectives on the VN distribution as intersections of random and fixed surfaces, are presented in Supplement Sections S1 and S2. In Section 3, we describe how different kinds of thickenings can be performed.

2.2 Normalizability

Do the expressions defined in Definitions 1 and 2 always refer to actual probability distributions? Not necessarily. If $g = y - x \in \mathbb{R}[x, y]$, the variety $\mathcal{V}(y - x)$ is unbounded, the pseudodensity is an infinitely long ridge over $y = x$, and the integral diverges. In general, unbounded varieties require truncation or tapering.

In most cases of practical interest, normalizability is easy to guarantee. Truncating the support to a compact region $\boldsymbol{\mathcal{X}} \subset \mathbb{R}^n$ immediately yields a proper distribution, and more generally multiplying the pseudodensity by any Lebesgue PDF $f(\boldsymbol{x})$ produces a normalizable pseudodensity since $\widetilde{p}(\boldsymbol{x}|g, \sigma^2) \leq 1$ everywhere (see Supplement Section S4 for the proof). On the other hand, even with a bounded variety, normalizability on all of $\mathbb{R}^n$ can fail due to persistent near zeros of $\overline{g}$; a detailed example involving a sum-of-squares polynomial and the role of coercivity in providing sufficient conditions is given in Supplement Section S4. Additional discussion of near roots and algebraic mixtures is provided in Supplement Section S3.

2.3 The multivariety normal distribution

It is natural at this point to consider the relationship between the VN distribution and the multivariate normal distribution: both are multivariate distributions and both generalize the univariate normal distribution. One key difference is that, like the univariate normal distribution, the multivariate normal distribution always concentrates its mass around a single point, the zero-dimensional variety $\mathcal{V}(\boldsymbol{x} - \boldsymbol{\mu}) = \{\boldsymbol{\mu}\}$, whereas the VN distribution can support a positive-dimensional variety. A second difference can be found in the variability about the variety: in the VN distribution, σ^2 gauges deviations from the variety globally and isotropically, whereas the multivariate normal distribution can support differing amounts of variability anisotropically. In fact, the VN distribution can be generalized to support this kind of variability as well. This leads us to a more general formulation of the VN distribution.

Definition 3. *A random vector $\boldsymbol{X} \in \mathbb{R}^n$ is said to have the multivariety normal (MVN) distribution, or simply the variety normal distribution, denoted $\boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{g}, \boldsymbol{\Sigma})$, if it admits a density*

$$p(\boldsymbol{x}|\boldsymbol{g}, \boldsymbol{\Sigma}) \propto \exp\left\{-\frac{1}{2}\overline{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{B})'\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{B})\right\} \tag{3}$$

$$= \exp\left\{-\frac{1}{2}\big(\boldsymbol{J}_{\boldsymbol{x}}^{+}\boldsymbol{g}(\boldsymbol{x}|\boldsymbol{B})\big)'\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{J}_{\boldsymbol{x}}^{+}\boldsymbol{g}(\boldsymbol{x}|\boldsymbol{B})\big)\right\} \tag{4}$$

with respect to the Lebesgue measure on $\mathbb{R}^n$ for some $\boldsymbol{g}(\boldsymbol{x}|\boldsymbol{B}) \in \mathbb{R}[\boldsymbol{x}]^m$ such that $\overline{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{B}) = \boldsymbol{J}_{\boldsymbol{x}}^{+}\boldsymbol{g}(\boldsymbol{x}|\boldsymbol{B})$, where $\boldsymbol{J}_{\boldsymbol{x}}^{+}$ is the $n \times m$ pseudoinverse of the $m \times n$ $\boldsymbol{x}$-Jacobian matrix of $\boldsymbol{g}(\boldsymbol{x}|\boldsymbol{B})$, and $\boldsymbol{\Sigma} \in \mathbb{S}_{+}^{n}$, the cone of symmetric positive definite $n \times n$ matrices.

Here, $\overline{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{B}) = \boldsymbol{J}_{\boldsymbol{x}}^{+}\boldsymbol{g}(\boldsymbol{x}|\boldsymbol{B})$ generalizes the notion of normalizing by the gradient. Considering $\boldsymbol{g}$ as the vector field $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$, the local linear approximation to points on the variety (properly translated) is given by $\boldsymbol{J}_{\boldsymbol{x}} \in \mathbb{R}^{m \times n}$, which is "inverted" by $\boldsymbol{J}_{\boldsymbol{x}}^{+} \in \mathbb{R}^{n \times m}$. In the square case, at smooth points the inverse function theorem guarantees that $\overline{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{B})$ will be properly normalized so that $\boldsymbol{\Sigma}$ gauges "covariance" globally across $\mathcal{V}(\boldsymbol{g})$. Several such square MVN distributions over zero-dimensional varieties are illustrated in Figure 6. If there are more variables than equations ($n > m$), the Jacobian represents an underdetermined system corresponding to a positive dimensional variety; $\boldsymbol{J}_{\boldsymbol{x}}$ is full row rank for almost all $\boldsymbol{x} \in \mathbb{R}^n$, and $\boldsymbol{J}_{\boldsymbol{x}}^{+} = \boldsymbol{J}_{\boldsymbol{x}}'(\boldsymbol{J}_{\boldsymbol{x}}\boldsymbol{J}_{\boldsymbol{x}}')^{-1}$. Previous examples of the variety normal, a single equation in many variables, are in fact examples of this situation. If there are more equations than variables ($m > n$) but the variety is nonempty, the Jacobian represents an overdetermined system; $\boldsymbol{J}_{\boldsymbol{x}}$ is full rank for almost every $\boldsymbol{x} \in \mathbb{R}^n$; and

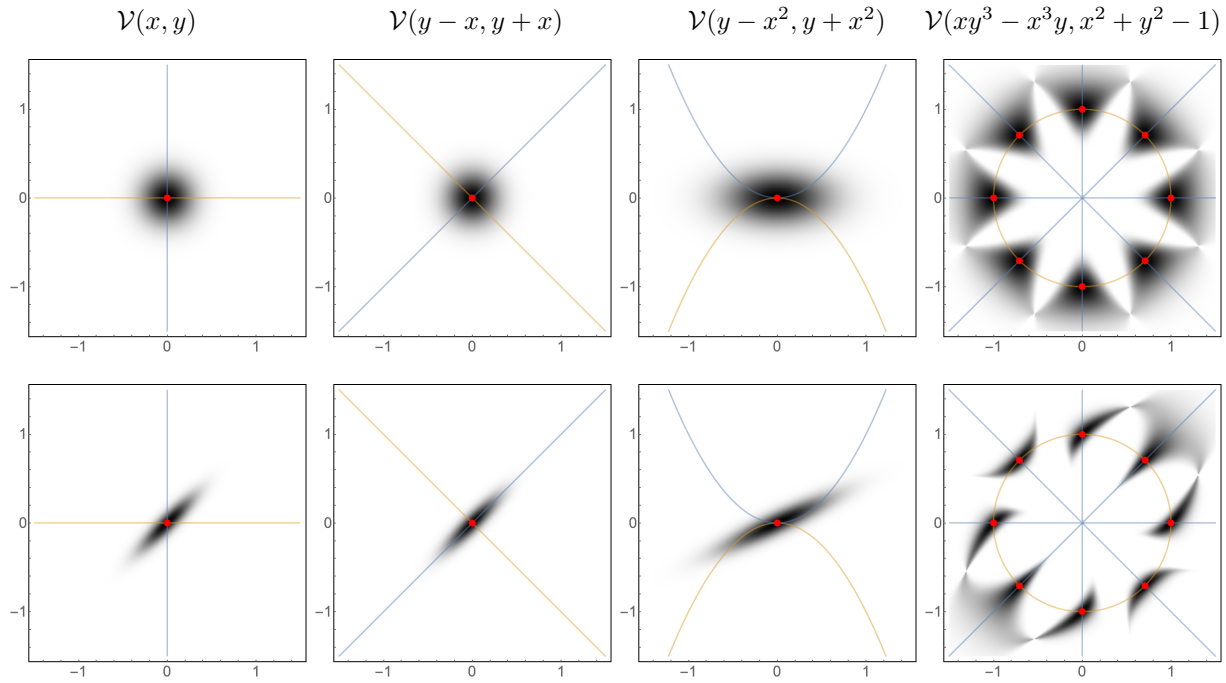

Figure 6: Multivariety normal distributions over zero-dimensional varieties and their associated systems with no "correlation" (top) and positive "correlation" ($\rho = .9$, bottom). In each case, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$. While $\sigma^2 = .2^2$ in each case, the density color scales and z-axes ranges vary across graphics.

$\boldsymbol{J}_{\boldsymbol{x}}^{+} = (\boldsymbol{J}_{\boldsymbol{x}}' \boldsymbol{J}_{\boldsymbol{x}})^{-1} \boldsymbol{J}_{\boldsymbol{x}}'$. Either way, $\boldsymbol{J}_{\boldsymbol{x}}^{+}$ transforms the original m-dimensional $\boldsymbol{g}$ into an n-dimensional vector $\overline{\boldsymbol{g}}$ compatible with $\boldsymbol{\Sigma}$.

As one might expect, the multivariety normal family of distributions subsumes the variety normal and multivariate normal families. Moreover, since the multivariety normal subsumes the multivariate normal, it subsumes the matrix-variate normal and tensor-variate normal (also called the multilinear normal) families as well [Ohlson et al., 2013].

If $m = 1$ and $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$, the MVN reduces to the VN distribution (2) (see Supplement Section S5 for the derivation). More generally, if the system is not overdetermined, $\boldsymbol{J}_{\boldsymbol{x}}$ typically has m independent rows and so if $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$,

$$p(\boldsymbol{x} | \boldsymbol{g}, \boldsymbol{\Sigma}) \quad \propto \quad \exp\left\{ -\frac{1}{2\sigma^2} \boldsymbol{g}' (\boldsymbol{J}_{\boldsymbol{x}} \boldsymbol{J}_{\boldsymbol{x}}')^{-1} \boldsymbol{g} \right\}. \tag{5}$$

If $\boldsymbol{g} = \boldsymbol{x} - \boldsymbol{\mu}$, $m = n$ and $\boldsymbol{J}_{\boldsymbol{x}} = \boldsymbol{I}_n$, and the distribution reduces to the multivariate normal. The individual varieties of the g_i's are the n hyperplanes of dimension $n - 1$ that run parallel to the coordinate axes and intersect at the point $\boldsymbol{\mu}$.

An added nuance of the MVN distribution not present in the VN distribution is that of covariance. Off-diagonal elements of $\boldsymbol{\Sigma}$ introduce "correlation" that orients the thickening anisotropically, and these can be easily specified via the decomposition $\boldsymbol{\Sigma} = \boldsymbol{DRD}$, with $\boldsymbol{D}$ a diagonal matrix of "standard deviationss" and $\boldsymbol{R}$ a "correlation" matrix . Additional examples illustrating this effect for both 2D and 3D varieties are provided in Supplement Section S6.

Understanding a variety tends to be easiest when correlations are zero, so that $\boldsymbol{\Sigma}$ is diagonal, and the magnitude of the thickening (variability) in each coordinate axis is uniform, so that $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$. In principle, the smaller the σ^2, the better; however, if σ^2 is chosen to be too small, the sampling schemes described in Section 4 tend to not perform as well. We revisit this topic there and in Section 5.3.

3 Induced Variety Distributions

The variety normal construction uses a Gaussian kernel to concentrate mass near the variety, but there are practical reasons a practitioner might want a different shape. For example, a distribution with compact support can enforce a hard bound on how far sampled points stray from the variety, avoiding tails that extend into irrelevant regions of the ambient space. A heavier-tailed base distribution can improve robustness when the gradient normalization introduces local distortions. As we will discuss in Section 3.1, choosing a one-sided base distribution is a stepping stone toward distributions on basic semi-algebraic sets defined by polynomial equalities and *inequalities*.

The generalization itself is straightforward. In the VN case, instead of thinking of $\overline{g}(\boldsymbol{x}|\boldsymbol{\beta})$ as replacing $x - \mu$ in the normal kernel, one can think of taking any distribution, applying a location transformation so that the mode is zero, and then substituting $\overline{g}(\boldsymbol{x}|\boldsymbol{\beta})$ for x; it just so happens that in the normal case this location transformation resets μ to 0. The same is true for the MVN case by replacing $\boldsymbol{x} - \boldsymbol{\mu}$ with $\overline{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{B})$. This same process applies generally:

1. Select a univariate or multivariate distribution with features of interest, e.g., unimodality,
2. Shift the distribution's location to the origin, rescaling as desired, and
3. Substitute $\overline{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{B}) = \boldsymbol{J}_{\boldsymbol{x}}^{+}\boldsymbol{g}$ for $\boldsymbol{x}$ in the resulting expression.

We refer to these distributions as *induced variety distributions*. It is helpful to illustrate the process on a univariate family. Suppose again we are interested in the alpha curve, the variety of $g = y^2 - (x^3 + x^2)$, but want to use a uniform distribution, a member of the beta family, instead of a normal. The PDF of the beta distribution is $p(x|\alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$ on $0 < x < 1$ for $\alpha, \beta > 0$; the uniform occurs when $\alpha = 1$ and $\beta = 1$. If $X \sim p(x|\alpha, \beta)$, it is a basic fact that if $\sigma > 0$ and $\mu \in \mathbb{R}$, the distribution of $\sigma X + \mu$ is proportional to $p(\frac{x-\mu}{\sigma}|\alpha, \beta)$ for $x \in (\mu, \mu + \sigma)$, so choosing $\mu = -\frac{\sigma}{2}$ is reasonable to center the distribution on zero. The resulting bivariate distribution described by the method is therefore $p(x, y|\alpha, \beta, \mu, \sigma) \propto p(\frac{\overline{g}-\mu}{\sigma}\Big|\alpha, \beta)$ with $\overline{g} = (y^2 - (x^3 + x^2))/\sqrt{(3x^2 + 2x)^2 + (2y)^2}$. As $\mathcal{V}(g)$ is not compact, we truncate the distribution to $\boldsymbol{\mathcal{X}} = [-1.5, 1.5]^2$. This is illustrated in Figure 7 with $\sigma = 1/2$ and various settings of α and β. The slight asymmetry observed is an artifact of selecting σ too large relative to the linear approximation produced by the gradient normalization.

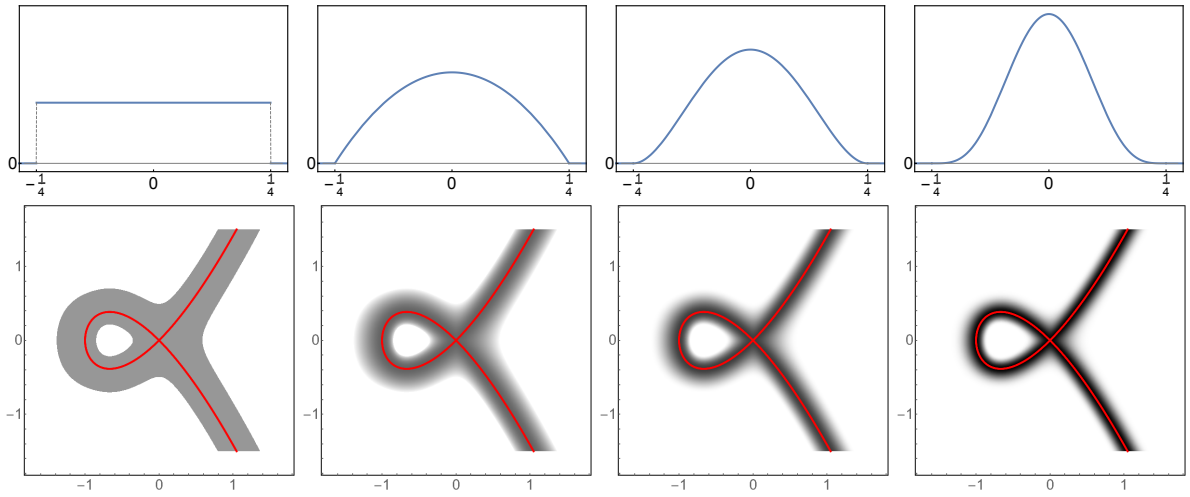

Figure 7: The induced variety beta distribution on the alpha curve with $\sigma = 1/2$, $\mu = -\sigma/2$, and $(\alpha, \beta) = (1,1), (2,2), (3,3)$, and $(5,5)$, from left to right.

While this strategy works, it does not in itself offer a clear advantage over the variety normal for sampling *about* a variety. Compact-support bases additionally introduce boundary constraints that can complicate HMC (e.g., initialization and feasible proposals). The real payoff of induced distributions comes when one moves from varieties to semi-algebraic sets, the topic of the next subsection.

3.1 Semi-algebraic distributions

Although the specific definition varies slightly, varieties are commonly referred to as algebraic sets in the literature. Semi-algebraic sets are generalizations of algebraic sets that allow for inequalities.

Definition 4. *A (basic)*[2] *semi-algebraic set is a set of the form* $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{g}(\boldsymbol{x}) = \mathbf{0}_m \text{ and } \boldsymbol{h}(\boldsymbol{x}) > \mathbf{0}_l\}$ *for* $\boldsymbol{g}(\boldsymbol{x}) \in \mathbb{R}[\boldsymbol{x}]^m$ *and* $\boldsymbol{h}(\boldsymbol{x}) \in \mathbb{R}[\boldsymbol{x}]^l$.

A few minor modifications to the strategies described above allow us to create distributions that focus their mass on semi-algebraic sets in the same way that variety distributions focus their mass on varieties. We note a quick trick at the outset, however, which is both common and useful in practice: if a single variable needs to be non-negative, it suffices to swap the variable with the square of another variable. As the square is always non-negative (over the reals), one simply retains the draws of the squared variable.

There are two basic approaches one can take to creating semi-algebraic distributions.

1. The first strategy is similar to the induced variety approach, but instead of picking a distribution that centers its mass on 0, pick a distribution that borders 0 by piling all its mass up against one side. We call the resulting distribution a border (induced variety) distribution. While this strategy seems like a good fit, it suffers from two drawbacks. First, while it works for varieties generated by a single polynomial, it is not clear how it could be conveniently applied to semi-algebraic sets in general, i.e., sets specified with several polynomials and a mix of equalities and inequalities. Second, the usefulness of the resulting distribution for understanding the semi-algebraic set is sensitive to the selection of the base distribution, often placing mass in the immediate vicinity of the boundary of the semi-algebraic set and almost entirely missing its interior.

2. The second strategy is to eliminate inequalities by converting them to equalities by introducing slack variables, placing a distribution about the corresponding variety in the higher dimensional space, and marginalizing out the slack variables. Geometrically, this corresponds to creating a variety in higher dimensional space whose projection is the semi-algebraic set of interest. This strategy appears to largely alleviate both of the challenges faced by the first strategy.

As a simple illustration of these concepts, consider sampling from the unit disc $\left\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\right\}$, which in canonical form is $\left\{(x, y) \in \mathbb{R}^2 : h(x, y) > 0\right\}$ with $h(x, y) = 1 - (x^2 + y^2)$:

1. For a border distribution, we may select the uniform distribution as the base distribution: $p(x|0, \beta) = 1[0 \leq x \leq \beta]$. Following the approach of the previous section, no shifting of the distribution is required, since it comes up against 0 from above, and we can simply set the distribution of interest to be $p(x, y|0, \beta) \propto 1[0 \leq \overline{h}(x, y) \leq 1]$, where $\overline{h}(x, y) = \frac{1-(x^2+y^2)}{2\sqrt{x^2+y^2}}$. This is illustrated in Figure 8.

   Notice that the resulting distribution always exhibits a hole in the middle, resulting in a distribution that for practical purposes is situated on an annulus instead of the disk. While this is alleviated by forcing β to be large, it always induces the artificial hole. As an alternative, one may choose to use a base distribution without an upper bound such as the exponential distribution, whose density is $p(x|\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$. This is also illustrated in Figure 8. Unfortunately this, too, results in a noticeably non-uniform distribution. As a uniform distribution cannot be placed on all of $\mathbb{R}_+$, this strategy generally appears unsatisfying.

[2] In real algebraic geometry, a semi-algebraic set is a finite union of such basic semi-algebraic sets defined by polynomial equations and inequalities. For simplicity, we focus on the basic ones.

2. By contrast, the second strategy converts each inequality into an equality by introducing a slack variable. The key observation is that $h(x, y) > 0$ if and only if there exists a real number $s \neq 0$ such that $h(x, y) - s^2 = 0$. Writing $g(x, y, s) = h(x, y) - s^2$, this is a polynomial equality in $\mathbb{R}^3$, so we can apply any variety distribution from the previous sections to the variety $\mathcal{V}(g) \subset \mathbb{R}^3$ and then marginalize out the slack variable s to obtain a distribution over the original (x, y) space.

   Concretely, with $h(x, y) = 1 - x^2 - y^2$, the lifted polynomial is $g(x, y, s) = 1 - x^2 - y^2 - s^2$, whose variety is the unit sphere $\mathcal{S}^2 \subset \mathbb{R}^3$. A variety normal distribution on the sphere concentrates mass near it in $\mathbb{R}^3$; marginalizing out s yields a distribution over the open unit disc in $\mathbb{R}^2$. In a sampler, one simply includes s as an additional variable and then discards it from the output. This is illustrated in Figure 9.

For the reasons described above, the second strategy generally seems preferable to the first, and so we recommend it when presented with semi-algebraic sets. Nevertheless, it too is imperfect and more work is

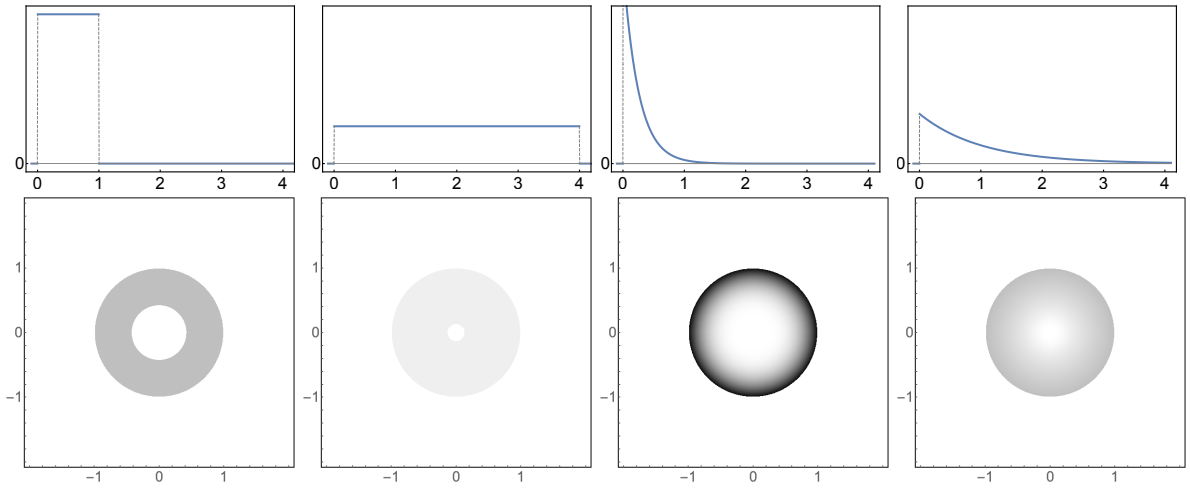

Figure 8: Distributions about semi-algebraic sets can be constructed by bordering distributions on one side of 0. From left to right, this is illustrated using the Unif$(0, 1)$, Unif$(0, 4)$, Exp(4), and Exp(1) distributions.

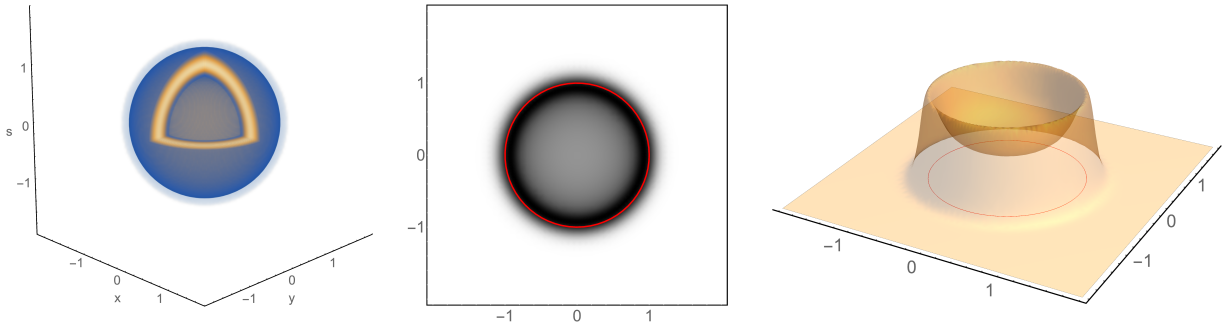

Figure 9: Distributions about semi-algebraic sets can be constructed as marginal distributions of variety distributions. (Left) A variety normal density on $\mathcal{S}^2$ along with a cutaway to visualize variability. (Middle) A density plot of the marginalized distribution over the semi-algebraic set of the open unit disc. (Right) The graph of the joint PDF of the distribution of (X, Y).

needed to develop a more satisfactory solution. Experimentally, the strategy presents some boundary effects where the distribution is slightly more concentrated near the boundary of the projected region; however, these appear to be relatively small in practice. It should also be noted that alternative algebraic formulations can be used to represent the semi-algebraic set of interest. For example, instead of moving from $h(x, y) > 0$ to $g(x, y, s) = h(x, y) - s^2 = 0$ in the above, we might have opted for $g(x, y, s) = s^2 h(x, y) - 1 = 0$. Unfortunately, the representation matters, and a bad representation may not work at all from a sampling perspective since the resulting varieties may be unbounded.

In general, if a semi-algebraic set is presented as above, it can be represented as the projection of the variety defined by $g_1(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x}), h_1(\boldsymbol{x}) - s_1^2, \ldots, h_l(\boldsymbol{x}) - s_l^2 \in \mathbb{R}[\boldsymbol{x}, \boldsymbol{s}]$ in $\mathbb{R}^{p+l}$ [Bochnak et al., 1991]. In fact, Motzkin has shown that this can always be done by introducing only one slack variable [Motzkin, 1970, Bochnak et al., 1991]; however, as the current strategy is simple and effective, we saw no need to pursue it further. In practice, introducing one slack variable per inequality simplifies both implementation and interpretation. A similar yet simpler strategy is provided in [Pecker, 1990]. It should be noted that semi-algebraic sets represent a very diverse collection of subsets of $\mathbb{R}^p$. In particular, they can allow puncturing of sets and, more generally, the removal of smaller dimensional geometries from larger dimensional sets. These cannot be faithfully represented by the present construction since variety distributions always thicken the variety into the fullness of the ambient space.

4 Sampling and Projection

In this section, we describe practical strategies to sample from the distributions presented in the previous sections to generate points near the variety as well as methods to move those points to the variety. To ease the exposition and accentuate how nice the situation can be, we focus our presentation on the variety normal distribution and note relevant differences for other induced variety distributions where key features stick out.

4.1 Sampling

There are many strategies one can use to sample from a probability distribution [Gentle, 2006, Robert and Casella, 2013]. For a given situation, the best sampler depends heavily on features of the distribution of interest, the target distribution. Perhaps most importantly from a sampling perspective, all the distributions described in this work are continuous on $\mathbb{R}^n$ and only known up to some constant of proportionality; this limits the strategies that can be used. Fortunately, there is already a wealth of knowledge on sampling from un-normalized distributions since these distributions form the foundation of Bayesian statistics. As such, Bayesian statisticians and others have developed very powerful general purpose tools to perform such sampling, both theoretically and in the form of freely accessible computer implementations. Before progressing to these more complex techniques, however, it is worth noting a simple one that works in many low-dimensional settings: rejection sampling.

4.1.1 Rejection sampling

Rejection sampling relies on the following basic fact referred to by Robert and Casella [2013] as the Fundamental Theorem of Simulation (FTS): Given a (potentially un-normalized) PDF $\widetilde{p}(\boldsymbol{x})$, sampling from $\widetilde{p}(\boldsymbol{x})$ is equivalent to sampling from the uniform distribution on the region below its graph,

$$\left\{(\boldsymbol{x}, u) \in \mathbb{R}^{n+1} : 0 \leq u \leq \widetilde{p}(\boldsymbol{x})\right\}.$$

Rejection sampling refers to any algorithm that obtains such uniform draws by sampling uniformly from the region below a proposal distribution $q(\boldsymbol{x})$ for which $Mq(\boldsymbol{x}) \geq \widetilde{p}(\boldsymbol{x})$ for some $M \in \mathbb{R}$ and then checking if the draws are also under the graph of $\widetilde{p}(\boldsymbol{x})$. Specifically: 1) sample $\boldsymbol{x}' \sim q(\boldsymbol{x})$, 2) sample $u \sim \text{Unif}(0, Mq(\boldsymbol{x}'))$, and 3) if $u \leq \widetilde{p}(\boldsymbol{x}')$, $\boldsymbol{x}' \sim p(\boldsymbol{x})$ otherwise repeat the procedure. The resulting retained/accepted draws constitute an independent and identically distributed sample from $p(\boldsymbol{x})$, the normalized distribution corresponding to $\widetilde{p}(\boldsymbol{x})$. The efficiency of the sampler depends on how similar $q(\boldsymbol{x})$ is to $p(\boldsymbol{x})$ and how small M is.

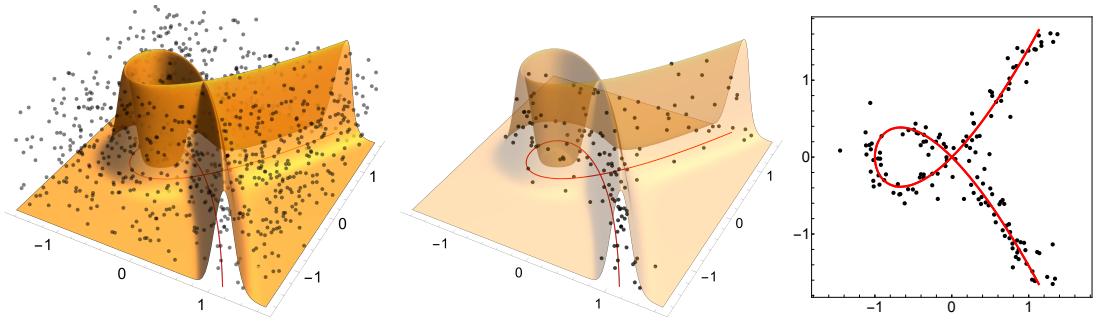

Figure 10: Rejection sampling is a simple algorithm that can be used to sample from low-dimensional variety distributions. Here we sample from the $\mathcal{N}_{2,\boldsymbol{\mathcal{X}}}\left(y^2-(x^3-x^2),\frac{1}{10^2}\right)$ distribution using a uniform proposal distribution over $\boldsymbol{\mathcal{X}}=[-1.5,1.5]^2$. We begin with 10,000 points (left), of which only 1,948 were found to be under $\widetilde{p}(x,y|g,\sigma^2)$, or about 1 in 5 draws (middle). The projected values, the (x,y) coordinates of the draws, are then retained (right). Probabilistically this corresponds to marginalizing the height variable u.

By construction, the un-normalized (multi)variety normal distributions' PDFs (2) and (3) are bounded above by 1, which is tight on the variety. If a spatial extent (range of the $\boldsymbol{x}$ variables) is known a priori, one can therefore simply apply a rejection sampling algorithm over that extent using the uniform distribution.

As an example, consider sampling from the variety normal on the alpha curve, $\mathcal{N}_{2,\boldsymbol{\mathcal{X}}}\left(y^2-(x^3+x^2),\frac{1}{10^2}\right)$ with $\boldsymbol{\mathcal{X}}=[-1.5,1.5]^2$, whose normalized polynomial was already provided in the previous section, namely $\overline{g}=(y^2-(x^3+x^2))/\sqrt{(3x^2+2x)^2+(2y)^2}$. From (2), its un-normalized PDF is $\widetilde{p}(\boldsymbol{x}'|g,\sigma^2)=\exp\left\{-50\overline{g}^2\right\}$. Thus, rejection sampling from the distribution would take first take two draws $x,y\sim\text{Unif}(-1.5,1.5)$ and one $u\sim\text{Unif}(0,1)$, and then check if $u\leq\widetilde{p}(\boldsymbol{x}'|g,\sigma^2)$. If so, we retain (x,y) as one draw from $p(x,y|g,\sigma^2)$, otherwise we repeat the procedure. This is illustrated in Figure 10.

Using this simple rejection scheme on the induced variety distribution formed by plugging $\overline{g}$ into the PDF of a $\text{Unif}(-\epsilon,\epsilon)$ PDF for small ϵ provides an interesting, intuitive procedure. The PDF of the uniform distribution is always proportional to a constant over its support and zero elsewhere, so that $\widetilde{p}(x,y|g,\epsilon)=1[-\epsilon\leq\overline{g}\leq\epsilon]$. If we use the same proposal $q(\boldsymbol{x})$ that is uniform over the square $\boldsymbol{\mathcal{X}}$, notice that we can skip the uniform sampling of the auxiliary variable u entirely since $\widetilde{p}$ is either 0 or 1. In this case, the algorithm reduces to this: 1) sample $x,y\sim\text{Unif}(-1.5,1.5)$; 2) evaluate $\overline{g}(x,y)$; and 3) retain (x,y) if $|\overline{g}(x,y)|\leq\epsilon$. This is very similar to the naive technician's procedure that, when looking for where g is 0, simply evaluates g at a collection of random points and retains the ones that are close to zero, but with one distinction: the present procedure uses the normalized version $\overline{g}$, not g, and that constitutes a significant distinction. Figure 11 illustrates this distinction on a Lissajous polynomial [Merino, 2003].

Figures 10 and 11 suggest that the method works well, and it does, at least in low dimensions. It has several advantages: it produces independent draws, it is almost trivial to implement, and it is easy to monitor. Nonetheless, it has two important limitations. First, one needs to know *a priori* the spatial extent of the variety so that a bounding box can be specified for the proposal. The more serious limitation is scaling. For varieties in higher dimensional ambient spaces, the variety distribution concentrates its mass near a vanishingly small region in that the acceptance rate of a uniform box proposal drops exponentially with dimension.

When to use rejection sampling. In practice, rejection sampling is attractive when the ambient dimension is small (roughly $n\leq 3$), the variety's spatial extent is known, and independent draws are desired. Beyond $n\approx 3$–4, the acceptance rate is generally too low to be practical, and one should switch to HMC as described next.

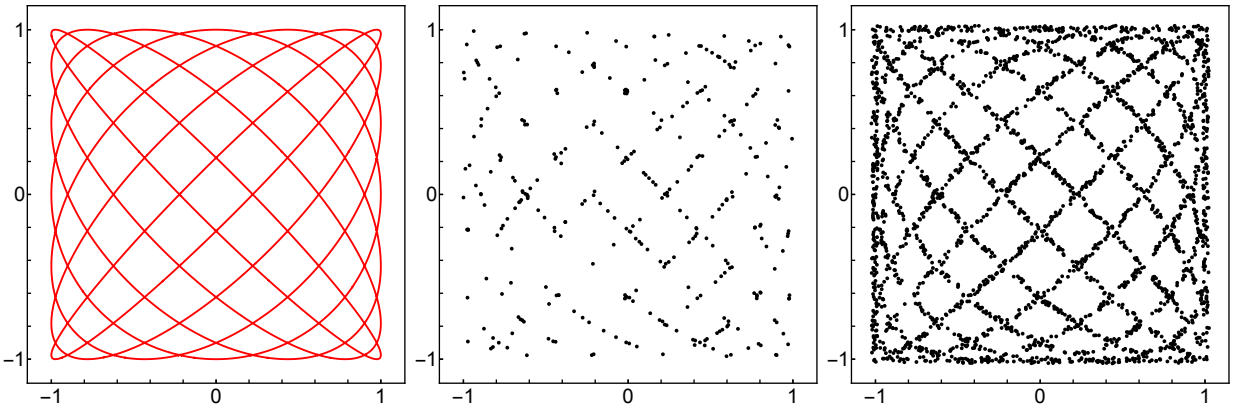

Figure 11: The rejection sampling algorithm on the induced variety distribution $\text{Unif}(-\epsilon, \epsilon)$ on the degree-18 Lissajous polynomial's variety with $\epsilon = .15$. (Left) the variety, (middle) the induced distribution plugging in g, corresponding to randomly plugging in points and checking where the polynomial is $|g| \leq \epsilon$, and (right) the same inserting $\overline{g}$ instead of g. The acceptance rate changes from 3.4% to 28.3% by using $\overline{g}$. $N = 10,000$.

4.1.2 MCMC sampling via HMC

Markov chain Monte Carlo (MCMC) constructs a Markov chain whose stationary distribution is the target, requiring only evaluation of an un-normalized density. Since the Metropolis acceptance ratio involves a ratio of densities, the normalizing constant cancels and never needs to be computed—exactly the situation for the distributions of Sections 2 and 3.

Among MCMC schemes, Hamiltonian Monte Carlo (HMC) has proved to be the most robust general-purpose strategy for sampling continuous distributions in moderate-to-high dimensions [Duane et al., 1987, Neal, 2011]. HMC simulates Hamiltonian dynamics to generate proposals in high-probability regions. The physical analogy is a puck sliding on a frictionless surface whose height is $-\log \widetilde{p}(\boldsymbol{x})$: regions of high density are valleys, so the puck naturally stays in them even as it moves along curved ridges—precisely the geometry of variety distributions. After a simulated trajectory of some duration, the puck's position is proposed and accepted or rejected via the standard Metropolis probability. Under mild conditions, the resulting chain has the target as its stationary distribution. An authoritative, accessible, and freely available introduction is provided in [Neal, 2011]. The process is illustrated in Figure 12.

For the variety normal, the HMC potential energy is $U(\boldsymbol{x}) = -\log \widetilde{p}(\boldsymbol{x})$, which reduces to a quadratic form in the normalized polynomials. Since $\boldsymbol{g} \in \mathbb{R}[\boldsymbol{x}]^m$ is polynomial, its Jacobian $\boldsymbol{J}_{\boldsymbol{x}}$ is available in closed form and the gradient of U needed by the leapfrog integrator can be computed exactly. (The full derivation of Hamilton's equations for this case is given in Supplement S7.) Since our focus is not on novel or optimally efficient samplers, we use the No-U-Turn sampler (NUTS), an HMC variant that adaptively tunes trajectory length and step size [Hoffman and Gelman, 2014].

Recasting as a Bayesian posterior. Rather than implementing HMC from scratch, a simple reformulation lets us use any off-the-shelf Bayesian sampler. The key is to recognize the variety normal PDF as the posterior of an overparameterized Bayesian model with a flat prior.

The central result of Bayes' theorem is that in a Bayesian setting with a parameter $\boldsymbol{\theta}$ and collection of data $\boldsymbol{y}$, the posterior distribution over $\boldsymbol{\theta}$ is proportional to the likelihood times the prior: $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. When the prior is flat, the posterior is simply proportional to the likelihood: $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})$. In the variety normal, the roles of parameter and data are swapped relative to typical Bayesian usage so that the polynomial coefficients $\boldsymbol{\beta}$ are known–they play the role of data–and the indeterminate vector $\boldsymbol{x}$ is unknown–it plays the role of parameters. Since $\overline{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{B})'\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{B}) = \big(\boldsymbol{0}_n - \overline{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{B})\big)'\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{0}_n - \overline{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{B})\big)$, the variety normal PDF is exactly the posterior of a model with a multivariate normal likelihood having mean $\overline{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{\beta})$ and

known covariance $\boldsymbol{\Sigma}$, a flat prior on $\boldsymbol{x}$, and a single "observation" of $\mathbf{0}_n$. Another way of viewing this is that the variety normal distribution can be sampled by framing it as the posterior distribution of an overparameterized Bayesian model and dummy observation. While this is ordinarily considered a defect of a Bayesian model, here it is an advantage, and as a consequence any probabilistic programming language that accepts a user-specified log-likelihood can therefore sample variety distributions, at least in principle.

We use Stan, a free and open-source probabilistic programming language that implements NUTS with automatic differentiation [Carpenter et al., 2017, Stan Development Team, 2018]. Stan has interfaces in R, Python, Julia, and other environments; models are specified in a markup language, translated to C++, and compiled before sampling. A Stan program to sample from the multivariety normal distribution defined by $x^2 + y^2 + z^2 = 1$ and $z = x + y$ illustrates the pattern:

```
data { real<lower=0> si; }

parameters { real x; real y; real z; }

transformed parameters {
  vector[2] g = [x^2+y^2+z^2-1, -x-y+z]';
  matrix[2,3] J = [
      [2*x, 2*y, 2*z],
      [-1, -1, 1]
    ];
}

model { target += normal_lpdf(0.00 | J' * ((J*J') \ g), si); }
```

The `data` block declares the known scale parameter σ; the `parameters` block declares the variety's indeterminates as the unknowns to be sampled; the `transformed parameters` block evaluates the polynomial

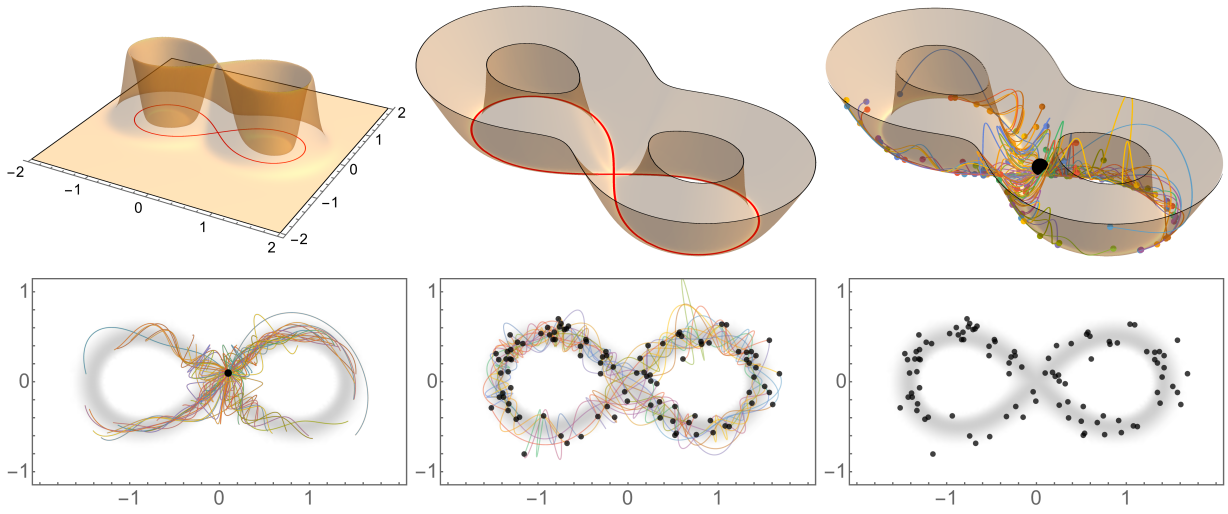

Figure 12: HMC samples distributions by simulating Hamiltonian dynamics on $-\log \widetilde{p}(\boldsymbol{x})$, which allows paths to efficiently explore the distribution. This display illustrates the process for the variety normal on the polynomial $g(x, y) = (x^2 + y^2)^2 - 2(x^2 - y^2)$, the lemniscate of Bernoulli, $\sigma = \frac{1}{10}$. Top: $\widetilde{p}(\boldsymbol{x})$, $-\log \widetilde{p}(\boldsymbol{x})$, and 100 different simulated paths originating from $(.1, .1)$. Bottom: The same 100 paths in the (x, y) plane, and a real HMC simulation using head-to-tail simulations with and without the transition paths.

system $\boldsymbol{g}$ and its Jacobian $\boldsymbol{J}$; and the `model` block contributes $\log p(\mathbf{0} \mid \boldsymbol{J}^{+}\boldsymbol{g}, \sigma)$ to Stan's target log-density. The expression `J' * ((J*J') \ g)` computes $\boldsymbol{J}^{+}\boldsymbol{g}$ using Stan's backslash linear solver, which avoids explicitly inverting $\boldsymbol{J}\boldsymbol{J}'$.

Since the polynomial coefficients function as data from the Bayesian perspective, a single compiled Stan binary can serve an entire family of variety normal distributions. The following program samples from any quadratic polynomial in indeterminates x and y; it is compiled once and then fed different coefficients at run-time via the data block. Box constraints $[-5, 5]^2$ are imposed directly in the parameter declarations:

```
data {
  real<lower=0> si;
  real b0; real bx; real by;
  real bx2; real bxy; real by2;
}

parameters {
  real<lower=-5,upper=5> x;
  real<lower=-5,upper=5> y;
}

model {
  real g = b0 + bx*x + bx2*x^2 + bxy*x*y + by*y + by2*y^2;
  real dgx = bx + 2*bx2*x + bxy*y;
  real dgy = by + 2*by2*y + bxy*x;
  real ndg = sqrt(dgx^2 + dgy^2);
  target += normal_lpdf(0.00 | g/ndg, si);
}
```

Of course, it bears repeating that not all combinations of parameters yield varieties over the reals, so we may need to check if an apparent component corresponds to actual roots. This verification can be done with endgames described next.

4.2 Endgames

In some applications, points *on* the variety are preferred to points merely near it. When σ is sufficiently small, the sampled points are already close, so the natural finishing step is to project each point onto the variety. Given a point $\boldsymbol{x}_0$ near the variety, this amounts to solving the nearest-point problem

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \mathcal{V}(\boldsymbol{g})}{\operatorname{argmin}} \, ||\boldsymbol{x} - \boldsymbol{x}_0||_2 = \underset{\boldsymbol{x} \in \mathcal{V}(\boldsymbol{g})}{\operatorname{argmin}} \, ||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 \,, \tag{6}$$

whose solution is unique for almost all $\boldsymbol{x}_0 \in \mathbb{R}^n$.

Why naive projection fails. Two natural approaches—minimizing $\sum g_i(\boldsymbol{x})^2$ via gradient descent initialized at $\boldsymbol{x}_0$, or solving the KKT system of the Lagrangian $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = ||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 + \boldsymbol{\lambda}'\boldsymbol{g}(\boldsymbol{x})$—can both find *a* point on the variety, but that point need not be $\boldsymbol{x}^*$. Newton-type methods are sensitive to initialization and can jump to a distant part of the variety, undoing the careful distributional exploration. This is illustrated in Figure 13.

Endgames from numerical algebraic geometry. Numerical algebraic geometry utilizes homotopy continuation methods that solve polynomial systems by smoothly deforming a system with known solutions into the target system and tracking the solutions along the deformation, e.g., see [Sommese and Wampler, 2005, Bates et al., 2013]. Near the end of each path, specialized finishing algorithms called *endgames* handle potential algebraic pathologies, e.g. singular roots where Newton's method loses quadratic convergence.

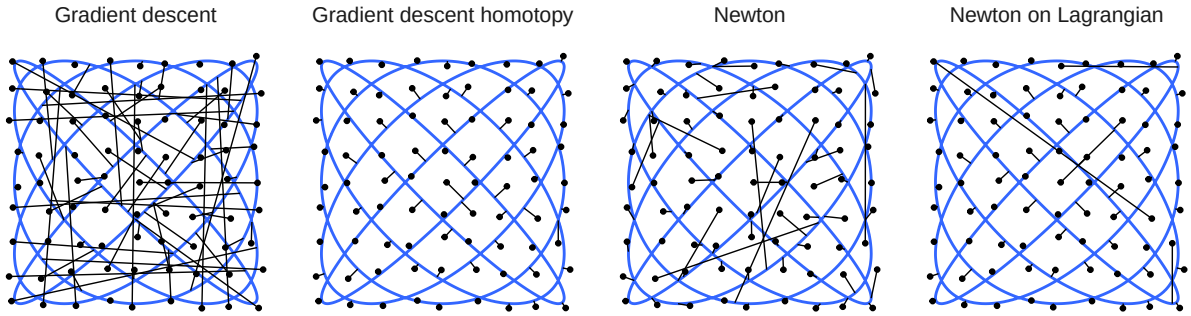

Figure 13: Points sampled from variety distributions can be projected onto them using endgames, special homotopy-based solvers that exploit the algebraic structure of the optimization problem. Naive strategies to move points to the variety can produce very unreliable results. In this image, a grid of points is projected onto the degree-18 Lissajous polynomial.

For our projection problem, endgames are ideal: they track a single path from the nearby point $\boldsymbol{x}_0$ to the variety, exploiting the algebraic structure of the problem and producing a result that is both on the variety and provably close to $\boldsymbol{x}^*$ as illustrated in Figure 13. The practitioner does not need to implement these methods from scratch; they are available in software packages such as Bertini [Bates et al., 2013] and HomotopyContinuation.jl [Breiding and Timme, 2018]. Concretely, the projection is formulated as a homotopy that continuously deforms the starting point $\boldsymbol{x}_0$ onto the variety [Griffin and Hauenstein, 2015, Brake et al., 2019]. Using Fritz John optimality conditions, the homotopy is

$$\boldsymbol{H}(\boldsymbol{x}, \boldsymbol{\lambda}; t) = \begin{bmatrix} \boldsymbol{g}(\boldsymbol{x}) - t\boldsymbol{g}(\boldsymbol{x}_0) \\ \lambda_0(\boldsymbol{x} - \boldsymbol{x}_0) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}) \end{bmatrix} = \boldsymbol{0}_{m+n}. \tag{7}$$

where $\boldsymbol{\lambda} \in \mathbb{P}^m$, projective space of dimension m. One starts the homotopy at $t = 1$ and start point $(\boldsymbol{x}, \boldsymbol{\lambda}) = (\boldsymbol{x}_0, [1, 0, \ldots, 0])$ and aims to track to $t = 0$. In our case, this is most easily done using the single polynomial $h(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x})'\boldsymbol{g}(\boldsymbol{x}) = \sum_{i=1}^{m} g_i(\boldsymbol{x})^2$ so that $m = 1$, and the computations can be done over an affine patch of $\mathbb{P}^1$, yielding the homotopy

$$\boldsymbol{H}^a(\boldsymbol{x}, \boldsymbol{\lambda}; t) = \begin{bmatrix} h(\boldsymbol{x}) - th(\boldsymbol{x}_0) \\ \lambda_0(\boldsymbol{x} - \boldsymbol{x}_0) + \lambda_1 \nabla h(\boldsymbol{x}) \\ \lambda_0 + \alpha_1 \lambda_1 - \alpha_0 \end{bmatrix} = \boldsymbol{0}_{n+2}, \tag{8}$$

where (α_0, α_1) can be selected randomly from the standard normal distribution.

Summary of the workflow. To obtain points on the variety given only its defining polynomial(s):

1. *Thicken.* Construct a variety normal (or induced) distribution that concentrates mass near the variety.
2. *Sample.* Draw from that distribution—rejection sampling for low dimensions ($n \leq 3$) and HMC via Stan for higher dimensions.
3. *Project.* If exact feasibility is required, project the draws onto the variety using endgames.

An illustration of this three-step procedure is included in Figure 14. Two examples in three dimensions are provided in Figure 15.

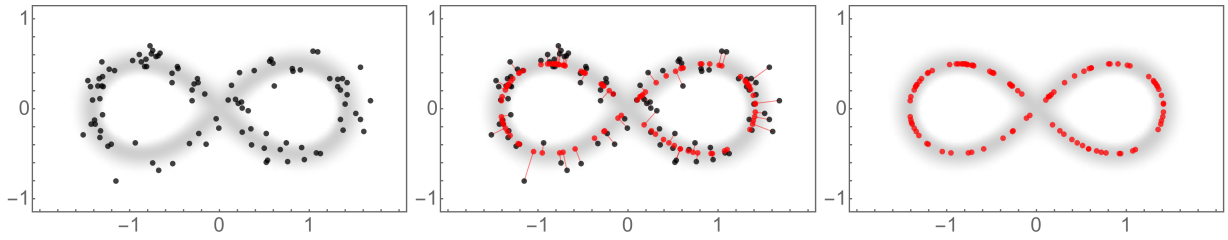

Figure 14: 100 draws from the lemniscate distribution in Figure 12 and their projections via endgames.

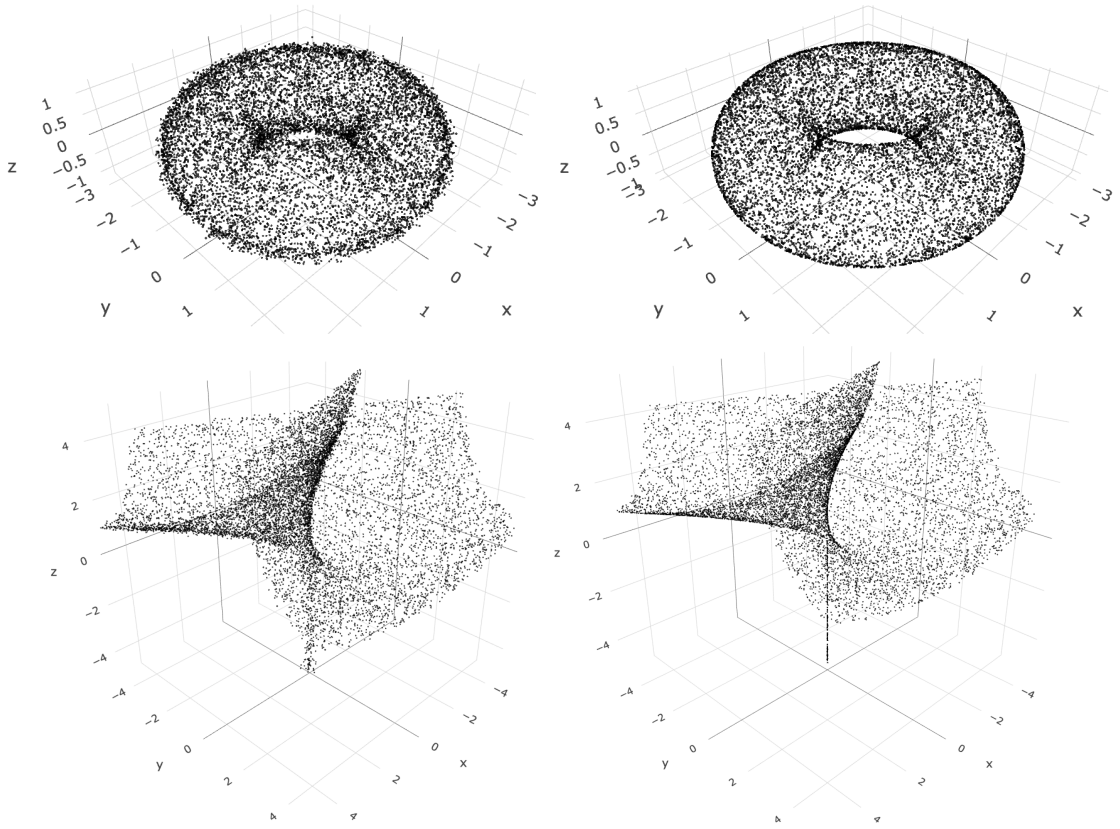

Figure 15: 10,000 draws from the torus $g = (x^2 + y^2 + z^2 + R^2 - r^2)^2 - 4R^2(x^2 + y^2)$ with $R = 2$ and $r = 1$ (top) and Whitney umbrella (bottom) variety normal distributions with $\sigma^2 = .25$ as drawn (left) and projected with endgames (right). Notice that draws are obtained on both the "handle" and the "canopy" of the umbrella.

5 Computational Examples

In this section, we illustrate the thicken–sample–project workflow on three tasks that are central to statistical computing on implicitly defined spaces. Each example is chosen to stress a different aspect of the pipeline:

- **Independence model** (Section 5.1): a low-dimensional setting where the workflow is transparent and rejection sampling provides a direct baseline, demonstrating constrained exploration and objective evaluation on a classical statistical model.

- **Lissajous benchmark** (Section 5.2): a planar curve with rich topology, illustrating the workflow as a tool for generating controlled synthetic benchmarks for topological data analysis.
- **Kuramoto model** (Section 5.3): a higher-dimensional system with both extended and isolated components, showing how multiple HMC chains can discover disconnected structure that is inaccessible to grid-based or rejection-based methods.

The examples are intended as workflow case studies that emphasize different computational roles rather than as head-to-head benchmarks. All computations were performed in R, often using the packages **mpoly** and **algstat** [R Core Team, 2023, Kahle, 2013, Kahle et al., 2017], with HMC implemented through Stan.

5.1 Sampling and objective evaluation on the independence model

A statistician's first question about the workflow is likely to be: "Can I use it on a model I already know well?" The 2×2 independence model is the simplest nontrivial algebraic statistical model, so it provides a transparent setting in which every step of thicken–sample–project can be inspected visually. Since the ambient space is four-dimensional and the model sits inside the probability simplex, rejection sampling over the unit cube is efficient and serves as a direct baseline.

A foundational perspective of algebraic statistics is that probabilistic independence is naturally encoded implicitly. This applies in several settings, but the most studied have been conditional independence models for multiway contingency table analysis and Gaussian graphical models [Sullivant, 2018, Drton et al., 2008]. In the latter setting, the covariance matrix is positive definite, marginal independence is encoded as zero-covariance constraints, and conditional independence is encoded by polynomial constraints on the inverse covariance matrix.

The most basic manifestation of independence is in a 2×2 table representing two binary random variables X and Y. In this setting $p_{ij} = P[X = i, Y = j]$ for $i, j = 0, 1$. By definition, independence demands four quadratic polynomial constraints $p_{ij} = p_{i+}p_{+j}$, where a $+$ in the subscript indicates summing the indeterminates over the index. These are taken together with the constraint $p_{++} = 1$ and the semi-algebraic constraints $p_{ij} \geq 0$. As is well known, the four independence constraints can be summarized into one: $p_{00}p_{11} = p_{01}p_{10}$. Ignoring the semi-algebraic constraints for the moment, the model is given by $\boldsymbol{g} = [p_{00}p_{11} - p_{01}p_{10}, p_{00} + p_{01} + p_{10} + p_{11} - 1] \in \mathbb{R}[\boldsymbol{p}]$.

In this low-dimensional example rejection sampling, over the unit cube is efficient so it provides a useful baseline. Figure 16 shows 10,000 draws from the corresponding variety normal distribution with $\sigma = .025$, displayed in barycentric coordinates. With rejection sampling on $[0, 1]^4$, the acceptance rate was approximately 1 in 1,000 reflecting the low codimension of the model relative to the ambient cube; generating the full sample required only a few seconds on a standard laptop.

As in most statistical settings, estimation on this model can be cast as an optimization problem. The feasible region is an algebraic surface, while the objective depends on the inferential target, for example likelihood or minimum chi-square [Kahle, 2011, Read and Cressie, 2012]. Variety sampling can serve as a Monte Carlo device for this optimization problem by evaluating the objective function at sampled points or, preferably, at projected points on the model. In that role, the method is exploratory rather than canonical: it helps reveal the shape of the objective over an implicitly described feasible set.

The right panel of Figure 16 illustrates exactly that use. The empirical distribution corresponding to counts $(8, 8, 1, 3)$ (a hypothetical observed table) is taken as a point in the simplex and the log-likelihood is evaluated on projected candidate points.

Two takeaways are immediate.

1. *Objective exploration without parameterization.* The method lets a practitioner visualize how a likelihood or loss surface behaves over an implicit model without deriving explicit coordinates.

2. *Algebraic projection ≠ statistical feasibility.* Since the semi-algebraic constraints were ignored, some projected points leave the simplex. In any application with positivity or boundedness constraints, projection must be followed by filtering, repair, or by incorporating the semi-algebraic structure into the projection routine.

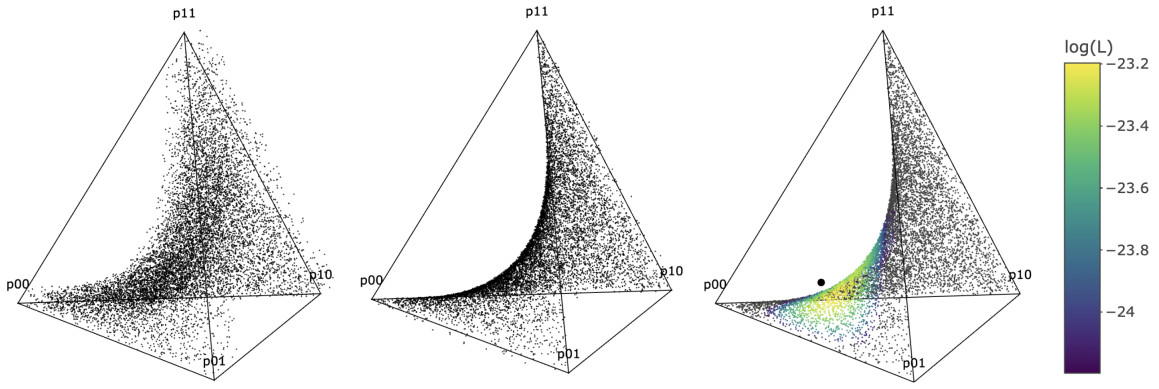

Figure 16: The independence model on a 2×2 contingency table is an implicitly defined surface in the probability simplex. Here 10,000 points were sampled near the model (left) and then projected to the variety (middle) using rejection sampling in the ambient cube. For those projected points that remain inside the simplex, the log-likelihood is evaluated with respect to the empirical distribution $(8, 8, 1, 3)$ (right), illustrating the workflow as a constrained objective-evaluation device.

5.2 Synthetic benchmarks for topological data analysis

The independence model shows how the workflow operates on a familiar statistical surface. We now turn to a complementary use case: generating controlled point clouds for downstream methodology, where the workflow serves as a benchmark-generation tool rather than as a statistical model per se.

Topological data analysis (TDA) studies qualitative structure in point clouds through summaries such as persistent homology and has become an increasingly statistical enterprise [Carlsson, 2009, Bubenik, 2015, Wasserman, 2018]. A practical bottleneck in this area is benchmark generation. Many synthetic examples are built from explicit parameterizations or regular meshes, which are unavailable for many implicit objects and can introduce coverage artifacts even when they are available. Variety distributions offer a different route: they generate point clouds near an implicitly defined shape directly in the ambient space, with a tunable noise level while not requiring a preferred parameterization. The contribution here is not a new TDA method, but a demonstration that the thicken–sample–project pipeline provides a flexible tool for constructing the test beds on which TDA methods can be evaluated.

Figure 17 shows 500 draws from a degree-6 Lissajous variety normal distribution in $\mathbb{R}^2$. Since this is a planar curve, rejection sampling on a bounding box was used. Vietoris–Rips persistent homology was computed using **ripserr**, which interfaces with Ripser [Wadhwa et al., 2020, Bauer, 2021]. The raw sample yields a noisy cloud around the target curve, while the projected sample sharpens the underlying geometry. In this example the dominant one-dimensional persistence signal reflects the thirteen bounded cells of the curve.

For our purposes, the important point is not that this construction supplies a canonical probability law on the variety. Rather, it provides a generative framework for controlled TDA experiments: one can vary σ^2, sample size, truncation, and whether draws are raw or projected to study how topological summaries respond to noise around an implicitly defined object. By reporting these choices explicitly, researchers can build reproducible and comparable benchmark suites for persistent homology and related summaries.

5.3 Disconnected and higher-dimensional constraint sets

The first two examples operate in low ambient dimension where rejection sampling is competitive. The real payoff of HMC-based variety sampling appears when the ambient dimension is high enough that rejection sampling becomes impractical and the variety itself has disconnected components that a single chain may fail to explore.

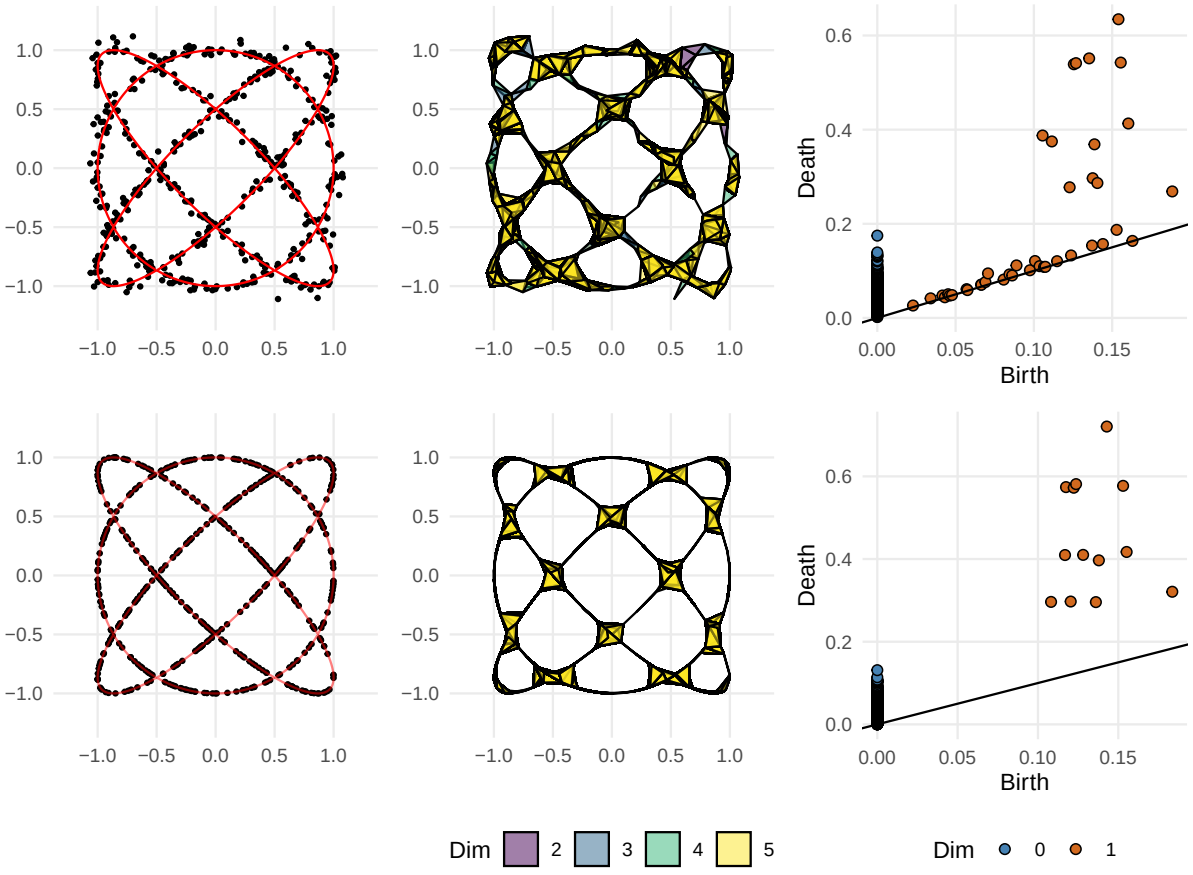

Figure 17: Variety distributions can serve as generative models for persistent-homology benchmarks. Here 500 draws were generated from the degree-6 Lissajous variety normal distribution $\mathcal{N}_2\left(9x^2 - 24x^4 + 16x^6 + 9y^2 - 24y^4 + 16y^6 - 1, \sigma = .025\right)$. The top row shows the raw point cloud, a Vietoris–Rips complex, and the resulting persistence diagram; the bottom row repeats the analysis after projecting the draws to the variety. The persistence diagram exhibits thirteen prominent one-dimensional features, matching the thirteen bounded cells of the curve.

We need a system whose solution set has both positive-dimensional components and isolated points so that the sampler must both explore a manifold and discover remote modes—a pattern that also arises in mixture identifiability problems and equilibrium analysis of reaction networks. The algebraic form of the Kuramoto synchronization model [Kuramoto, 1975, Mehta et al., 2015] provides exactly such a system while being well studied in its own right.

For $N = 5$ oscillators, the following polynomial system is obtained after fixing one oscillator to remove rotational symmetry and setting the natural frequencies to zero:

$$
\begin{aligned}
&5s_1 - 5c_1s_2 + 5c_2s_1 - 5c_1s_3 + 5c_3s_1 - 5c_1s_4 + 5c_4s_1, \quad 5s_2 + 5c_1s_2 - 5c_2s_1 - 5c_2s_3 + 5c_3s_2 - 5c_2s_4 + 5c_4s_2,\\
&5s_3 + 5c_1s_3 - 5c_3s_1 + 5c_2s_3 - 5c_3s_2 - 5c_3s_4 + 5c_4s_3, \quad 5s_4 + 5c_1s_4 - 5c_4s_1 + 5c_2s_4 - 5c_4s_2 + 5c_3s_4 - 5c_4s_3,\\
&c_1^2 + s_1^2 - 1, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad c_2^2 + s_2^2 - 1,\\
&c_3^2 + s_3^2 - 1, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad c_4^2 + s_4^2 - 1.
\end{aligned}
$$

The top four describe the interaction between the oscillators, while the bottom four restrict the oscillators to the unit circle. To remove rotational symmetry, the 5^{th} oscillator is fixed with $s_5 = 0$ and $c_5 = 1$.

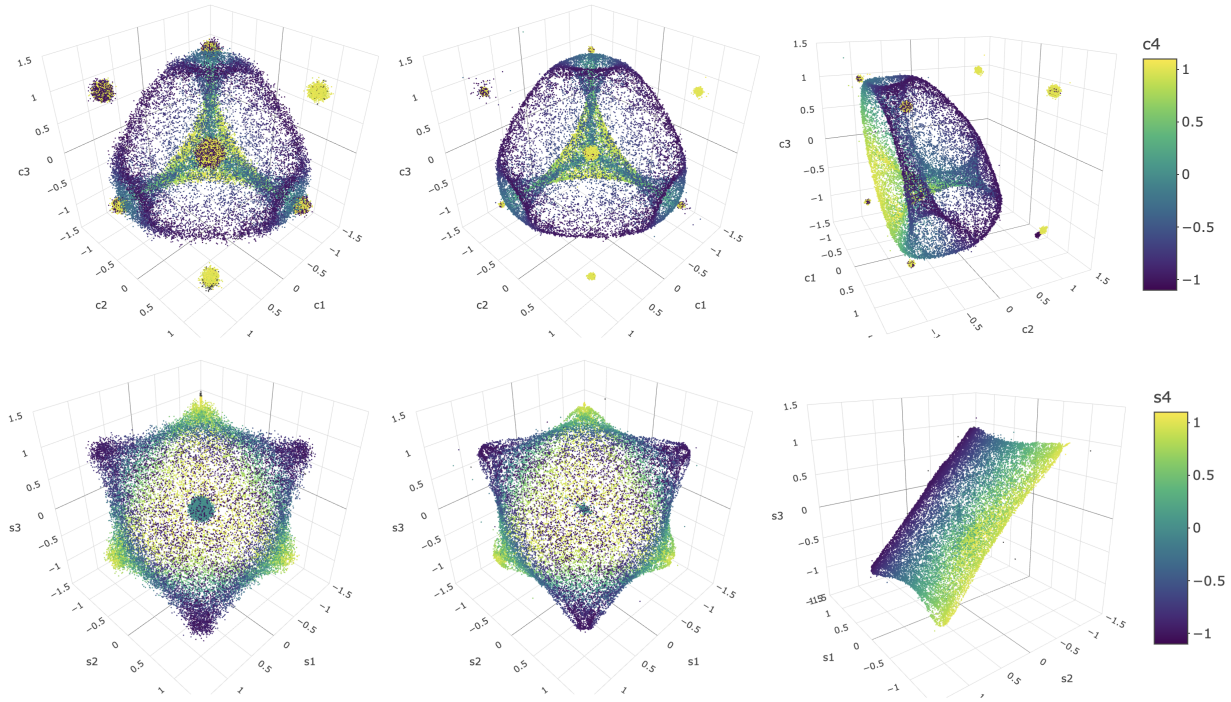

Figure 18: HMC with Stan can sample intricate variety distributions in higher-dimensional settings. Here 64,000 draws from the variety normal on the $N = 5$ Kuramoto model in 8 dimensions with $\sigma = .05$ were drawn using 64 HMC chains in parallel. The top row projects onto (c_1, c_2, c_3) and the bottom row onto (s_1, s_2, s_3). From left to right: raw draws, projected draws, and projected draws from a different perspective.

The corresponding real variety in $\mathbb{R}^8$ contains a two-dimensional component as well as $2^4 = 16$ isolated points such that $s_i = 0$ and $c_i^2 = 1$ for $i = 1, \ldots, 4$. We sampled from the corresponding variety normal distribution with $\sigma = .05$ using 64 HMC chains run in parallel through Stan, drawing 1,000 post-warmup iterations per chain for a total of 64,000 draws. Running many chains is essential here: since the variety has disconnected components, individual chains typically settle into a single mode, and multi-chain diagnostics (split-$\widehat{R}$ and between-chain variance) are the primary tool for detecting incomplete exploration. Projected draws were obtained by Newton-style projection. Figure 18 shows that HMC can recover both the positive-dimensional component and isolated solutions in a setting where rejection sampling, gridding, and marching-cubes-style approaches are no longer practical.

Three observations emerge from this example.

1. *Multi-chain initialization is essential, not optional.* As in all major MCMC schemes, HMC gets stuck in local modes when σ^2 is sufficiently small. Variety components correspond to modes of the variety normal. Thus, some of the 64 chains migrated to and explored the two-dimensional component while others moved to isolated solutions. As in Bayesian sampling, initializing many chains at random ("overdispersed") starting points [Lunn et al., 2012] is prudent in an effort to discover all components. A practical rule of thumb is to run at least as many chains as the expected number of components, though in exploratory mode one may need substantially more.

2. σ^2 *controls a bias–mixing trade-off.* Larger values make sampling easier and make it less likely that chains become trapped in disconnected components, but they also blur the geometry. Smaller values sharpen the representation of the target set but make mixing and projection harder. In this example, $\sigma = .05$ gave a useful balance; reducing σ to .01 caused several chains to produce divergent transitions

in Stan's diagnostic output, signaling that the target density was too sharply concentrated for the default leapfrog step size.

3. *Projection quality must be monitored.* The use of Newton-style projections did not work perfectly as illustrated by the stray points falling off some isolated solutions. Projection quality is therefore a substantive part of the computational method, not a cosmetic post-processing step, with the more stable endgame approach being advantages as illustrated in Figure 13. For statistical use, projected points should be checked against the ambient constraints of the problem—for example, by verifying that $\|\overline{\boldsymbol{g}}(\boldsymbol{x})\|$ is below a user-specified tolerance.

5.4 Summary of practical guidance

Taken together, the three examples support four concrete recommendations for practitioners adopting the thicken–sample–project workflow.

1. *Run multiple dispersed chains.* For disconnected or sharply curved spaces, single-chain behavior can be badly misleading. In the Kuramoto example, no single chain visited all components; multi-chain diagnostics were the only reliable signal that components had been missed.

2. *Treat σ^2 as a tuning parameter with diagnostics.* Smaller values sharpen fidelity to the target set but make mixing and component discovery harder. Standard HMC diagnostics—divergent transitions, effective sample size per second, and split-$\widehat{R}$—remain informative and should be used to guide the choice of σ^2.

3. *Report the full ambient construction.* For benchmark-generation problems, the ambient construction, truncation region, σ^2, and whether the analysis uses raw or projected draws should all be reported. These choices materially change the geometric and topological signal presented to downstream methods, as the TDA example illustrates.

4. *Separate algebraic from statistical feasibility.* Projection onto the variety need not preserve side constraints such as positivity or simplex membership when those conditions are ignored. Thus, such constraints must be utilized with the projection method or checked and enforced post-projection as the independence-model example demonstrates.

These points are familiar in spirit from broader Monte Carlo practice, but they become especially important when the target space is defined only implicitly and there is no parameterization to enforce the constraints automatically.

6 Discussion

This paper develops a statistical-computing workflow for implicitly defined polynomial model spaces. The central move is simple: replace a hard equality-constrained space by a probabilistic thickening in the ambient space, sample that thickened distribution with tools such as HMC, and project the resulting points back to the constraint set when exact feasibility is needed. The workflow is useful for at least three kinds of tasks illustrated here: constrained exploration and objective-function evaluation, synthetic benchmark construction for topology-aware methods, and component discovery in disconnected systems. In this section we discuss limitations of the approach, its relationship to existing methodology, and directions for future work.

6.1 Computational cost and scaling

The per-iteration cost of the HMC sampler is dominated by two operations: evaluating the log-density and its gradient, and performing the leapfrog integration that underlies each proposal. For the variety normal distribution on a system $\boldsymbol{g} \in \mathbb{R}[\boldsymbol{x}]^m$ in n variables, the log-density involves the Jacobian matrix $\boldsymbol{J}_{\boldsymbol{x}} \in \mathbb{R}^{m \times n}$,

the product $\boldsymbol{J}_{\boldsymbol{x}}\boldsymbol{J}_{\boldsymbol{x}}'$, and its inverse applied to $\boldsymbol{g}(\boldsymbol{x})$. Computing $\boldsymbol{J}_{\boldsymbol{x}}$ requires evaluating mn partial derivatives, each of which is a polynomial one degree lower than the corresponding g_i; for polynomials of degree d this is an $O(mn\binom{n+d-1}{n})$ operation in the worst case, though in practice sparsity reduces the cost considerably. The matrix inversion is $O(m^3)$, which is negligible when the number of defining equations m is small but becomes a bottleneck for larger systems. Since Stan computes these quantities through automatic differentiation, the user does not need to supply analytic derivatives, but the cost scales with the complexity of the computational graph defined by $\boldsymbol{g}$ and its Jacobian.

Leapfrog integration adds a multiplicative factor equal to the number of leapfrog steps per proposal, which NUTS determines adaptively [Hoffman and Gelman, 2014]. When σ^2 is small, the log-density surface $-\log\widetilde{p}(\boldsymbol{x})$ has steep, narrow ridges, and the leapfrog integrator requires shorter step sizes and more steps to track them without divergent transitions. This effect is the single largest practical bottleneck: as σ^2 decreases, the effective cost per independent sample can grow rapidly. Users should monitor the number of divergent transitions and the effective sample size per unit time as primary diagnostics.

The projection step adds a separate cost. Each endgame tracks a single homotopy path using predictor–corrector steps, whose cost depends on the degree and dimension of the system. For smooth points on the variety, this is typically fast; near singular points, the path may require more refined step control [Bates et al., 2013]. In our examples, projection constituted a small fraction of total computation time relative to the HMC sampling, but this balance could shift for higher-degree systems.

On the whole, the workflow scales most naturally to problems with moderate ambient dimension ($n \lesssim 20$), a small number of defining equations, and moderate polynomial degree (polynomials with degree at most 5–6, based on the examples studied here). The Kuramoto example in Section 5.3, with $n = 8$ and eight defining polynomials, represents the upper end of what we have tested systematically. Scaling to significantly larger systems will likely require exploitation of sparsity in the Jacobian, more efficient Stan model parameterizations, and possibly analytic gradient implementations that bypass general-purpose automatic differentiation.

6.2 Relationship to existing constrained-sampling methods

Several families of methods address the problem of sampling on or near constrained sets, and it is useful to position the variety-distribution workflow relative to them.

Manifold MCMC methods. Geodesic Monte Carlo [Byrne and Girolami, 2013], Riemann manifold Langevin and HMC [Girolami and Calderhead, 2011], and constrained HMC on submanifolds [Lelièvre et al., 2019] operate directly on the constraint surface by projecting proposals back to the manifold at each MCMC step, typically using the constraint Jacobian. These methods produce draws that satisfy the constraints exactly at every iteration and, under suitable regularity, sample a well-defined measure on the manifold. By contrast, the variety-distribution approach samples in the ambient space and only projects after the chain has run, which has three practical consequences. First, it avoids the need for a full-rank Jacobian at every iterate, so it can in principle operate on singular varieties where manifold MCMC methods require special handling. Second, it leverages existing Stan infrastructure without requiring a custom integrator, which substantially lowers the implementation burden. Third, since it does not constrain the chain to the variety at each step, it can more easily jump between disconnected components when σ^2 is not too small. The cost of these advantages is that the ambient-space chain does not sample a canonical surface measure, and the relationship between the ambient distribution and the geometry of the variety is mediated by both σ^2 and the projection map.

Hit-and-run and related polytope samplers. Hit-and-run sampling and its variants are effective for sampling convex bodies and polytopes [Lovász and Vempala, 2007]. They require a membership oracle for the feasible set, which is straightforward for linear constraints but more involved for nonlinear varieties. More fundamentally, hit-and-run methods are designed for full-dimensional feasible regions, whereas the varieties considered here are typically lower-dimensional subsets of $\mathbb{R}^n$. The thickening step in our workflow effectively creates a full-dimensional surrogate that can be sampled with standard tools, but does so via a smooth density rather than a hard membership constraint.

Algebraic sampling via Markov bases. In the algebraic statistics literature, the Diaconis–Sturmfels algorithm [Diaconis and Sturmfels, 1998] constructs Markov bases that enable exact sampling from conditional

distributions on discrete exponential families. This is a different problem setting—discrete rather than continuous, and focused on exact conditional tests—but shares the goal of using algebraic structure to enable statistical computation. The variety-distribution approach addresses continuous polynomial constraints and is closer in spirit to the computational geometry side of the problem.

Numerical algebraic geometry. Methods such as monodromy-based, distance-based, and gradient-based sampling [Breiding and Marigliano, 2020, Dufresne et al., 2019, Hauenstein, 2013, Harris et al., 2026] can produce points on varieties with well-understood coverage properties. These are powerful when the goal is to determine existence of real solutions and produce exactly feasible points by construction. The variety-distribution approach is complementary: it is better suited to stochastic exploration and objective evaluation over a variety, and it naturally produces a cloud of points whose spread is controlled by σ^2, which is useful for downstream tasks such as topological data analysis.

No single method dominates across all criteria. The variety-distribution workflow is most attractive when the user wants to explore an implicitly defined space with minimal implementation effort, when the variety may be singular or disconnected, and when approximate rather than exact feasibility is acceptable for the sampling phase. When exact on-manifold samples with known distributional properties are required, manifold MCMC or numerical algebraic geometry methods are preferable.

6.3 The Sigma-Squared Trade-off

The parameter σ^2 governs a genuine tension between geometric fidelity and computational tractability, and practical guidance for its selection is important.

When σ^2 is small, the variety normal distribution concentrates its mass in a narrow tube around the variety. This is desirable for geometric accuracy: sampled points lie close to the target set, and projection via endgames is fast and reliable because the starting points are already near the variety. However, the log-density surface becomes steep and sharply curved, which creates difficulties for HMC. The leapfrog integrator requires smaller step sizes to avoid divergent transitions, NUTS may take many leapfrog steps per proposal, and the effective sample size per unit of computation drops. Moreover, when the variety has multiple disconnected components, a small σ^2 creates deep energy barriers between them, making it unlikely that any single chain will visit more than one component.

When σ^2 is large, the distribution is broad and easy to sample. HMC mixes quickly, and chains initialized at dispersed locations can explore the ambient space freely. However, the sampled points may be far from the variety, which has two consequences. First, projection becomes less reliable: endgames initialized far from the variety may converge to the wrong critical point or fail to converge at all. Second, the relationship between the ambient density and the geometry of the variety becomes attenuated; features such as curvature, singular points, and narrow components may be poorly represented in the sample.

In practice, we recommend the following diagnostic strategy.

1. Begin with a moderate value of σ^2 and examine standard HMC diagnostics: divergent transitions, $\widehat{R}$ statistics, and effective sample size. If divergences are frequent, increase σ^2.

2. Evaluate the residual $\|\overline{\boldsymbol{g}}\|$ at each of the sampled points. These should be small relative to the scale of the polynomials. If not, σ^2 may be too large for the sampled points to be usefully close to the variety.

3. After projection, check the fraction of endgames that converge successfully and the distribution of distances $\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\|$ between raw and projected points. Large or highly variable distances suggest that σ^2 is too large for reliable projection.

4. For problems with various connected components, run multiple chains at several values of σ^2. A larger σ^2 is useful for initial component discovery; a smaller value can then be used to refine sampling within each component.

An annealing-style strategy—starting with large σ^2 and gradually decreasing it—is a natural extension that has not been systematically investigated. Automatic selection of σ^2 based on properties of the defin-

ing polynomials (degree, Jacobian condition number, distance between components) would be a valuable methodological contribution.

6.4 Projection and side constraints

When only using equality constraints, the endgame-based projection step moves sampled points to the algebraic variety $\mathcal{V}(\boldsymbol{g})$ and does not account for additional constraints that may be part of the statistical problem. In the independence-model example of Section 5.1, the model of interest is not the full algebraic variety but rather its intersection with the probability simplex: the projected point must satisfy $g(\boldsymbol{p}) = 0$ and $\sum p_{ij} = 1$ as well as $p_{ij} \geq 0$. The algebraic projection respects only the equality conditions and some projected points may violate the inequality constraints since they were not considered when projecting.

This limitation is not merely cosmetic. In applications where the feasible region is a semi-algebraic set $S = \mathcal{V}(\boldsymbol{g}) \cap K$ for some region K defined by polynomial inequalities, the projected points may fall on components of $\mathcal{V}(\boldsymbol{g})$ that lie entirely outside K, or may land on the correct component but at an infeasible point.

Several potential remedies exist, though none is fully satisfactory. *Post-projection rejection* simply discards projected points that violate the side constraints. This is the approach taken in Section 5.1 and is straightforward to implement, but it wastes computation and can introduce bias if the rejection rate varies systematically across the variety. *Barrier or penalty terms* can be added to the ambient log-density to discourage the HMC sampler from exploring regions that would project outside K. For example, adding a log-barrier term $-\mu \sum \log(p_{ij})$ to the log-density would keep sampled points in the positive orthant, at the cost of introducing an additional tuning parameter μ and modifying the effective distribution. *Constrained projection* would replace the unconstrained endgame with a constrained optimization that finds the nearest point on $\mathcal{V}(\boldsymbol{g}) \cap K$. This is the most principled approach but is also the most computationally demanding as it requires solving a semi-algebraic optimization problem for each sampled point. Developing efficient constrained endgames that respect common statistical constraints such as positivity and simplex membership is an important open problem.

6.5 Scope and intended use

It is important to be clear about what the variety-distribution workflow is designed for and what it is not. The method is a tool for *computational exploration* of implicitly defined polynomial spaces. It is well suited to the following tasks:

- *Exploration and visualization*: generating point clouds that reveal the geometry of a variety, including its dimension, curvature, connected components, and singular structure.
- *Objective evaluation*: evaluating a function (likelihood, loss, distance) over a constrained set without requiring an explicit parameterization of that set.
- *Component discovery*: identifying various connected components of a variety by running multiple chains and examining which components they visit.
- *Benchmark generation*: producing synthetic point clouds near structured implicit sets for use in topological data analysis, machine learning, or other downstream procedures.

The method is *not* a replacement for dedicated constrained optimization algorithms when a single global optimum is needed, nor is it a substitute for numerical algebraic geometry when a complete enumeration of solutions or a certified numerical irreducible decomposition is required. Most importantly, the projected draws should not be treated as samples from a canonical "uniform" measure on the variety without further justification. The thickened distribution is well defined in the ambient space, but the measure induced on the variety after projection depends on the projection map, on σ^2, and on the local geometry of the variety. Inferential conclusions that depend on a specific surface measure require either a separate theoretical analysis of the projection-induced measure or the use of methods that sample a known measure directly.

6.6 Future directions

Several lines of further work are suggested by the current methodology.

Characterizing the projection-induced measure. The most fundamental open question is: what measure does the thicken-sample-project workflow induce on the variety? For smooth varieties and small σ^2, one expects the induced measure to be related to the volume form weighted by the condition number of the Jacobian, but making this precise—and understanding how the relationship breaks down at singular points—requires further analysis. Such a characterization would clarify when projected draws can be used for inference and what reweighting, if any, is needed.

Automatic σ^2 selection. A principled rule for choosing σ^2 based on the defining polynomials would make the workflow more accessible. One natural approach is to link σ^2 to the condition number of the Jacobian or to the reach of the variety (the largest radius for which the nearest-point projection onto the variety is well defined). Adaptive schemes that adjust σ^2 during sampling, analogous to the step-size adaptation in NUTS, are also worth exploring.

Annealing and tempering. The tension between component discovery (large σ^2) and geometric fidelity (small σ^2) suggests that tempering strategies may be valuable. Parallel tempering, in which multiple chains run at different values of σ^2 and exchange states, could combine the exploratory power of large σ^2 with the geometric accuracy of small σ^2. Simulated annealing schedules that decrease σ^2 over the course of sampling are another natural option.

Extension to complex varieties. The distributions introduced here are defined over $\mathbb{R}^n$, but many algebraic objects of interest are naturally complex. Extending the variety normal construction to $\mathbb{C}^n$ would connect the workflow to a richer algebraic-geometric toolkit, including Bézout-type root counts and the full power of homotopy continuation over the complex numbers. The main challenge is that HMC requires a real-valued log-density, so a complex extension would, for example, need to work with the real and imaginary parts of the variables, effectively doubling the ambient dimension.

Scaling to larger systems. For systems with many variables or high-degree polynomials, the current approach may become prohibitively expensive. Exploiting sparsity in the Jacobian, using structured linear algebra for the matrix inversions in the log-density, and developing problem-specific Stan models that avoid redundant computation are all promising directions. Integration with GPU-accelerated automatic differentiation frameworks could also extend the practical reach of the method.

Connections to learning on varieties. The ability to generate point clouds near implicit sets has potential applications in machine learning, including training data generation for classifiers whose decision boundaries are algebraic, exploration of loss-function level sets, and parameterization learning via neural networks trained on variety samples. These connections are largely unexplored.

6.7 Closing remarks

Even with the limitations discussed above, the variety-distribution workflow fills a useful computational niche. Many implicit model spaces are easy to write down algebraically but hard to explore numerically. The tools available for these spaces—symbolic Gröbner basis computations, numerical homotopy continuation, semidefinite relaxations—are powerful but are designed primarily for solving and classifying rather than for the exploratory, stochastic, repeated-evaluation tasks that arise in statistical computing. Variety distributions, HMC, and projection together turn implicit polynomial spaces into objects that can be investigated with familiar statistical software. The workflow requires minimal problem-specific implementation: the user supplies the defining polynomials and a value of σ^2, and existing infrastructure handles the rest. We hope that this lowers the barrier between the algebraic geometry and statistical computing communities, and enables new computational approaches to problems defined by polynomial constraints.

Acknowledgments

JDH was supported in part by NSF CCF-2331400.

References

Daniel J. Bates, Jonathan D. Hauenstein, Andrew J. Sommese, and Charles W. Wampler. *Numerically solving polynomial systems with Bertini*, volume 25. SIAM, 2013.

Ulrich Bauer. Ripser: efficient computation of vietoris–rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, 2021.

Jacek Bochnak, Michel Coste, and Marie-François Roy. *Real Algebraic Geometry*, volume 36. Springer-Verlag, 1991.

Danielle A. Brake, Noah S. Daleo, Jonathan D. Hauenstein, and Samantha N. Sherman. Solving critical point conditions for the hamming and taxicab distances to solution sets of polynomial equations. In *2019 21st International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, pages 51–58, Timisoara, Romania, 2019, .

Paul Breiding and Orlando Marigliano. Random points on an algebraic manifold. *SIAM Journal on Mathematics of Data Science*, 2(3):683–704, 2020.

Paul Breiding and Sascha Timme. HomotopyContinuation.jl: A package for homotopy continuation in Julia. In *International Congress on Mathematical Software*, pages 458–465. Springer, 2018.

Peter Bubenik. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102, 2015.

Simon Byrne and Mark Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017.

Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26(1):363–397, 1998.

Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on algebraic statistics*, volume 39. Springer Science & Business Media, 2008.

Simon Duane, Anthony D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.

Emilie Dufresne, Parker B. Edwards, Heather A. Harrington, and Jonathan D. Hauenstein. Sampling real algebraic varieties for topological data analysis. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1531–1536, 2019.

James E. Gentle. *Random number generation and Monte Carlo methods*. Springer Science & Business Media, 2006.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Zachary A. Griffin and Jonathan D. Hauenstein. Real solutions to systems of polynomial equations and parameter continuation. *Adv. Geom.*, 15(2):173–187, 2015.

Katherine E. Harris, Jonathan D. Hauenstein, and Hoon Hong. Sampling smooth points on real algebraic sets using perturbations. *Journal of Symbolic Computation*, 137:102584, 2026.

Jonathan D. Hauenstein. Numerically computing real points on algebraic sets. *Acta Appl. Math.*, 125: 105–119, 2013.

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

David Kahle. *Minimum distance estimation in categorical conditional independence models*. PhD thesis, Rice University, 2011.

David Kahle. mpoly: Multivariate polynomials in R. *The R Journal*, 5(1):162–170, 2013. URL https://journal.r-project.org/archive/2013-1/kahle.pdf.

David Kahle, Luis Garcia-Puente, and Ruriko Yoshida. *algstat: Algebraic Statistics in R*, 2017. URL https://github.com/dkahle/algstat. R package version 0.1.1.

Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics (Kyoto Univ., Kyoto, 1975)*, volume 39 of *Lecture Notes in Phys.*, pages 420–422. Springer, Berlin, 1975.

Tony Lelièvre, Mathias Rousset, and Gabriel Stoltz. Hybrid Monte Carlo methods for sampling probability measures on submanifolds. *Numerische Mathematik*, 143:379–421, 2019.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

David Lunn, Chris Jackson, Nicky Best, Andrew Thomas, and David Spiegelhalter. *The BUGS book: A practical introduction to Bayesian analysis*. CRC press, 2012.

David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. WinBUGS-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, 2000.

Dhagash Mehta, Noah S. Daleo, Florian Dörfler, and Jonathan D. Hauenstein. Algebraic geometrization of the kuramoto model: Equilibria and stability analysis. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(5), 2015. 053103.

Julio Castineira Merino. Lissajous figures and chebyshev polynomials. *The College Mathematics Journal*, 34(2):122–127, 2003.

Theodore Samuel Motzkin. The real solution set of a system of algebraic inequalities is the projection of a hypersurface in one more dimension. *Inequalities, II (Proceedings of the Second Symposium, US Air Force Academy, Colorado, 1967)*, pages 251–254, 1970.

Radford M Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galen L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, Handbooks of Modern Statistical Methods, chapter 5, pages 113–162. CRC Press, Boca Raton, 2011.

Martin Ohlson, M. Rauf Ahmad, and Dietrich Von Rosen. The multilinear normal distribution: Introduction and some basic properties. *Journal of Multivariate Analysis*, 113:37–47, 2013.

Daniel Pecker. On the elimination of algebraic inequalities. *Pacific Journal of Mathematics*, 146(2):305–314, 1990.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

Timothy R.C. Read and Noel A.C. Cressie. *Goodness-of-fit statistics for discrete multivariate data*. Springer Science & Business Media, 2012.

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

Andrew J. Sommese and Charles W. Wampler. *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science*. World Scientific, 2005.

Stan Development Team. RStan: the R interface to Stan, 2018. URL http://mc-stan.org/. R package version 2.18.2.

Bernd Sturmfels. *Solving systems of polynomial equations*. Number 97 in CBMS Regional Conference Series in Mathematics. American Mathematical Society, 2002.

Seth Sullivant. *Algebraic statistics*, volume 194. American Mathematical Society, 2018.

Raoul Wadhwa, Matt Piekenbrock, and Jacob Scott. **ripserr***: Calculate Persistent Homology with Ripser-Based Engines*, 2020. URL https://CRAN.R-project.org/package=ripserr. R package version 0.1.1.

Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2018.