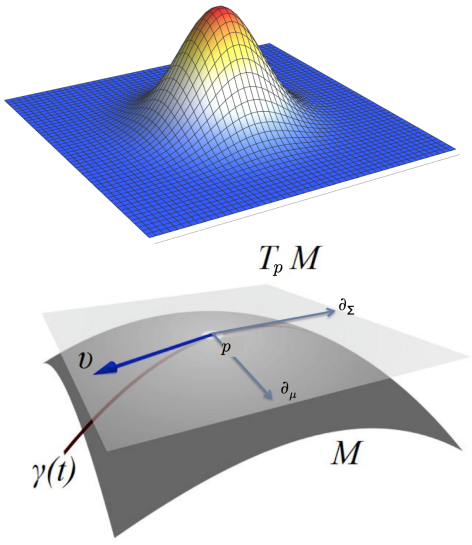
Brownian Motion on the Riemannian Manifold of Univariate Gaussian Distributions

Nicholas Wisniewski

Introduction

The central limit theorem is associated with Brownian motion, which can be illustrated by a random walk throught the parameter space. The asymptotic sampling distribution of estimators can be identified with the density of random walk endpoints over the parameter space. The variance of this distribution is inversely proportional to the sample size, though the central limit theorem is only valid when the sample size is very large, and consequently when the individual steps or fluctuations are very small.

In this sense, the asymptotic distribution results from a Euclidean geometry on the parameter space, even if the parameter space is truly curved and not flat, since tiny fluctuations are confined to a locally flat region (the tangent plane) as sample size becomes large. However, as fluctuations become large due to small sample size, the central limit theorem does not hold, and neither does the Euclidean geometry. Consequently, non-asymptotic behavior is much less understood for most distributions except the Gaussian.



In information geometry, the parameter space is understood to be a Riemannian manifold with curvature, and not a flat Euclidean space. This structure was first noted by Rao in 1945, but the statistical implications of curvature are still unclear. We postulate that large fluctuations might be better under-

stood within this setting, but calculation is prohibitively difficult. However, the definition of Brownian motion readily extends to Riemannian manifolds, allowing for numerical simulation.

Furthermore, the effect of Riemannian curvature on diffusion is known: negative curvature acts to accelerate diffusion at larger radius, and positive diffusion acts to decelerate it. The location-scale families of probability distributions, such as the Gaussian (both univariate and multivariate), all have constant negative curvature. Thus, it is possible that non-asymptotic behavior might have some explanation in terms of the Riemannian curvature, which would allow us to extend the diffusion analogy beyond the central limit theorem. We therefore ask: if Euclidean diffusion provides a physical interpretation of the central limit theorem, does Riemannian diffusion provide a non-asymptotic generalization? To answer this, we numerically simulate Brownian motion on the manifold of Gaussian distributions, and compare the results for small sample size to analytical results.

Information Geometry

We begin by reviewing the basic information geometric concepts related to statistical inference. Information geometry is a very rich mathematical framework with applications to many fields of scientific inquiry. But because the tools of differential geometry are unfamiliar to most scientists and statisticians, the field has consequently received little attention. Fortunately, many of the important concepts can be understood without much exposition. In this section, we review the elementary connections between geometry and inference, and provide well-known results for the family of Gaussian distributions.

The space of probability distributions $p(x; \theta)$ forms a Riemannian manifold, with coordinate system given by the parametrization θ, and metric tensor $g(\theta)$ given by the Fisher information. This unique geometric framework provides the necessary structure by which to generalize the notion of a *distance* between two statistical models, which can be used as a test statistic. Furthermore, this distance is invariant under reparametrization of the model $\theta \to \xi$, so that inferences made using that distance (as in hypothesis testing) do not depend on any particular parameterization.

It is natural to understand this geometry in the context of maximum likelihood estimation, where we are interested in maximizing the log-likelihood function to find the best-fit parameters θ:

$$\hat{\theta}_{\text{MLE}} = \text{argmax}_{\theta} \left[\log p(x; \theta)\right]. \tag{1}$$

We would like to know how much a change in one parameter affects the estimation of the other parameters. To do this, we can consider infinitesimal changes in the log-likelihood function, $\partial_{\theta} \log p(x; \theta)$. Since the covariance between vectors looks like the expectation value of their inner product, we can write the covariance between parameters θ_i and θ_j as

$$g_{\text{ij}}(\theta) = E_x\left[\partial_{\theta_i} \log p \cdot \partial_{\theta_j} \log p\right]. \tag{2}$$

This expression is the Fisher information, and can be understood geometrically by considering the infinitesimal vectors $\partial_{\theta_i} \log p$ as basis vectors in the plane tangent to the coordinate θ on the manifold. Since $g(\theta)$ is a Riemannian metric, it is symmetric positive definite, and infinitesimal distances in parameter space are given by the quadratic form

$$\text{ds}^2 = g_{\text{ij}}(\theta)\, d\hat{\theta}^i\, d\hat{\theta}^j. \tag{3}$$

This quadratic distance measure is generally recognized in statistics as the Mahalanobis distance, which can be formulated using any covariance matrix. The special case here, using the Fisher information matrix as the covariance, leads to an information distance. This information distance specifies the amount of information lost by misestimating a model. For instance, if the estimate looks nothing like the true model, then the two models will be far apart on the manifold since a lot of information has been lost. Likewise, if the estimate is very close to the model, then little information has been lost. The invari-

ance under reparameterization implies that information defined in this way deals strictly with the inference process (sampling), and not the parameterization. Furthermore, the invariance is unique to the Fisher metric; this is the only framework of statistical inference that does not depend on parameterization.

The role of Fisher information and the associated information distance is therefore fundamental to statistics, though many approximations are used that are much more familiar in practice. In frequentist statistics, the most common approximations to the information distance are the likelihood ratio and Wald statistic. In information theory, the most common approximations are the Kullback-Leibler divergence, and it's symmetrized form, the Jensen divergence. However, the approximate nature of these distance measures is readily apparent in that they don't actually satisfy all the axioms of distance measures (symmetry, positivity, and triangle inequality); none of them satisfy the triangle inequality. The Riemannian information distance is therefore seen as the most fundamental, since it is the only measure that satisfies the axioms.

Another important role of the Fisher information is the computation of standard errors for parameters. In our explanation of the Fisher information, we showed how it is derived as the covariance of parameters during estimation. This notion is formalized by the Cramer-Rao inequality, which specifies the maximum accuracy attainable in inference on a finite number of samples. Specifically, it shows that the uncertainty of an estimate (i.e. the variance of the sampling distribution, which is the square of the standard error) is greater than or equal to the inverse Fisher information,

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{n}\, g(\theta)^{-1}. \tag{4}$$

Asymptotically, this is the variance of the normally distributed estimates guaranteed by the central limit theorem. Thus, the asymptotic sampling distribution can be expressed using the Fisher information as the inverse covariance matrix in a multivariate Gaussian distribution on parameter space,

$$p\left(\hat{\theta}; \theta\right) \sim \exp\left\{-\frac{1}{2}\, n\, g_{\mathrm{ij}}(\theta)\, d\,\hat{\theta}^{\,i}\, d\,\hat{\theta}^{j}\right\}. \tag{5}$$

So, the asymptotic sampling distribution can be written simply using the infinitesimal distance,

$$p\left(\hat{\theta}; \theta\right) \sim \exp\left\{-\frac{1}{2}\, n\, \mathrm{ds}^2\right\}, \tag{6}$$

This allows us to think of a surface of equal-likelihood as simply a sphere of radius $\sqrt{n}$ ds.

Geodesic Equation

However, the infinitesimal distance ds is only defined in the tangent plane. For larger distances, it is necessary to extend the notion of distance to the manifold. This is done using geodesic curves. We can define a geodesic as any curve along which tangent vectors remain parallel when transported. In order to calculate the length of a geodesic curve, one chooses the affine connection to be the Levi-Civita connection because it is, in general, the only metric-compatible connection; it preserves the angles between tangent vectors, as well as the length of tangent vectors under parallel transport. The general formulation of the Levi-Civita connection, dependent only upon the metric tensor, is

$$\Gamma_{\mathrm{ij}}{}^{k} = \frac{1}{2}\, g^{\mathrm{km}}(\partial_j g_{\mathrm{mi}} + \partial_i g_{\mathrm{mj}} - \partial_m g_{\mathrm{ij}}) \tag{7}$$

This connection is what allows us to write covariant derivatives on the manifold,

$$\nabla = \partial + \Gamma \tag{8}$$

Now let $\theta(t)$ be a smooth curve connecting two points on the manifold, and $t \in \mathbb{R}$ parameterize the curve so that $\theta(t_1) = \theta_1$ and $\theta(t_2) = \theta_2$. For this curve to be a "straight line", the tangent vector $v = \dot{\theta}(t)$ to the

curve must not change as we move along the curve; the directional derivative of the tangent vector along the curve must vanish.

$$\nabla_v v = 0 \tag{9}$$

In other words, the curve is geodesic if and only if it satisfies the ordinary differential equation (the geodesic equation, or Euler-Lagrange equation):

$$\ddot{\theta}^k + \Gamma_{ij}{}^k \dot{\theta}^i \dot{\theta}^j = 0 \tag{10}$$

where the usual dot notation represents differentiation with respect to t. The shortest curve connecting two points on a manifold will be a geodesic. In this way we see that distances between points on a Riemannian manifold are found to be infima of arc lengths over all curves connecting the points, which are more readily computed as solutions to a set of nonlinear differential equations. The information distance between θ_1 and θ_2 is then defined as the length of the shortest curve between the two probability distributions:

$$s(\theta_1, \theta_2) = \left| \int_{t_1}^{t_2} \sqrt{g_{ij}(\theta(t))\, \dot{\theta}^i(t)\, \dot{\theta}^j(t)}\ dt \right|. \tag{11}$$

Gaussian Family

With these definitions, we can carry out the Riemannian calculations specifically for the Gaussian manifold.

Metric

The metric tensor is straightforward to calculate once a coordinate system is chosen. Since the metric tensor is by coordinate covariant, coordinate systems can be changed at any time by the tensor transformation rules, and the final results will remain unchanged. We will therefore begin with the most commonly used coordinate system, $\theta = \{\mu, \sigma\}$.

In this coordinate system, the metric is

$$ds^2 = \frac{1}{\sigma^2}\left(d\mu^2 + 2\, d\sigma^2\right), \tag{12}$$

or in matrix notation

$$g_{ij} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \tag{13}$$

Geodesic Equation

Computation of the Levi-Civita connection yields:

$$\Gamma_{ij}{}^k = \begin{pmatrix} \left\{0, \frac{1}{2\sigma}\right\} & \left\{-\frac{1}{\sigma}, 0\right\} \\ \left\{-\frac{1}{\sigma}, 0\right\} & \left\{0, -\frac{1}{\sigma}\right\} \end{pmatrix}. \tag{14}$$

From this we arrive at the geodesic equations:

$$\begin{aligned} 0 &= \mu'' - \frac{2\,\mu'\,\sigma'}{\sigma} \\ 0 &= \sigma'' + \frac{\mu'^2 - 2\,\sigma'^2}{2\,\sigma} \end{aligned} \tag{15}$$

Geodesic Length

The general solution to this equation is,

$$\mu(t) = \frac{\mp 4\, {c_2}^3}{c_1\left({c_2}^2 + 2\,{c_3}^2\, e^{\pm 2\, c_2\, t}\right)} + c_4$$
$$\sigma(t) = \frac{4\,{c_2}^2\, c_3\, e^{\pm c_2\, t}}{c_1\left({c_2}^2 + 2\,{c_3}^2\, e^{\pm 2\, c_2\, t}\right)} \tag{16}$$

Solving for the constants in terms of the boundary conditions $\theta_{t=0} = \{\mu_0, \sigma_0\}$, $\theta_{t=1} = \{\mu_1, \sigma_0\}$, yields

$$\mu(t) = \mu_0 + \left(\sqrt{2}\; e^{\frac{s}{\sqrt{2}}} \left(-1 + e^{\sqrt{2}\; s\, t}\right) \sigma_0 \sqrt{-\frac{s^2 \left(\sigma_0 - e^{-\frac{s}{\sqrt{2}}}\, \sigma_1\right)}{\sigma_0 - e^{\frac{s}{\sqrt{2}}}\, \sigma_1}} \left(-\sigma_0 + e^{\frac{s}{\sqrt{2}}}\, \sigma_1\right) \right) \Big/$$
$$\left(s \left(e^{\frac{s}{\sqrt{2}}} \left(-1 + e^{\sqrt{2}\; s\, t}\right) \sigma_0 + \left(e^{\sqrt{2}\; s} - e^{\sqrt{2}\; s\, t} \right) \sigma_1 \right) \right) \tag{17}$$

$$\sigma(t) = \frac{e^{\frac{s\, t}{\sqrt{2}}} \left(-1 + e^{\sqrt{2}\; s}\right) \sigma_0\, \sigma_1}{e^{\frac{s}{\sqrt{2}}} \left(-1 + e^{\sqrt{2}\; s\, t}\right) \sigma_0 + \left(e^{\sqrt{2}\; s} - e^{\sqrt{2}\; s\, t} \right) \sigma_1}$$

Where s is defined as the geodesic length, the minimal length on the manifold between ξ_0 and ξ_1.

$$s = \sqrt{2}\; \cosh^{-1}\left[\frac{(\mu_1 - \mu_0)^2 + 2\left(\sigma_0^2 + \sigma_1^2\right)}{4\, \sigma_0\, \sigma_1} \right] \tag{18}$$

Multivariate Generalization

In the multivariate case, we write the condition for a curve to be a geodesic as

$$\frac{d^2\, \mu}{\mathrm{dt}^2} = \left(\frac{\mathrm{d}\Sigma}{\mathrm{dt}}\right) \Sigma^{-1} \left(\frac{\mathrm{d}\mu}{\mathrm{dt}}\right) \tag{19}$$

$$\frac{d^2\, \Sigma}{\mathrm{dt}^2} = \left(\frac{\mathrm{d}\Sigma}{\mathrm{dt}}\right) \Sigma^{-1} \left(\frac{\mathrm{d}\Sigma}{\mathrm{dt}}\right) - \left(\frac{\mathrm{d}\mu}{\mathrm{dt}}\right) \left(\frac{\mathrm{d}\mu}{\mathrm{dt}}\right), \tag{20}$$

This is the more general form we will use in our Brownian motion algorithm.

Riemannian Curvature

The Riemann curvature tensor contains all the information about the curvature of a manifold, and is given by,

$$R^l{}_{\mathrm{ijk}} = \partial_j \Gamma_{\mathrm{ik}}{}^l - \partial_k \Gamma_{\mathrm{ij}}{}^l + \left(\Gamma_{\mathrm{js}}{}^l\, \Gamma_{\mathrm{ik}}{}^s - \Gamma_{\mathrm{ks}}{}^l\, \Gamma_{\mathrm{ij}}{}^s \right) \tag{21}$$

For the Gaussian distribution

$$R^{l}{}_{ijk} = \begin{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & \frac{-1}{\sigma^2} \\ \frac{1}{\sigma^2} & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & \frac{1}{2\sigma^2} \\ \frac{-1}{2\sigma^2} & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \end{pmatrix}$$

The Ricci curvature tensor is a contraction of the Riemann curvature tensor,

$$R_{ij} = R^{l}{}_{ijj} = \begin{pmatrix} \frac{-1}{2\sigma^2} & 0 \\ 0 & \frac{-1}{\sigma^2} \end{pmatrix} \tag{23}$$

Finally, the scalar curvature is a contraction of the Ricci tensor with the metric tensor, giving the trace of the Ricci curvature tensor,

$$R = g^{ij} R_{ij} = -1 \tag{24}$$

The space is constant negative curvature, or hyperbolic.

Brownian Motion

We simulate Brownian motion on the Riemannian manifold of univariate Gaussian distributions by implementing a random walk that numerically integrates the geodesic equation using Euler's method. By using the geodesic equation, we intend to capture the acceleration experienced due to the constant negative curvature of the manifold. It is known that negative curvature acts to accelerate diffusion outward, causing advanced diffusion at farther distances. This effect should alter the sampling distribution of any statistic, causing overdispersion compared to the central limit theorem. This effect should be pronounced for extremely small sample size, and tend to vanish for extremely large sample size.

The basis vectors

We use the basis defined by Skovgaard for the multivariate Gaussian distribution in order to be able to generalize our results.

In the multivariate case, the basis vectors are

$$\partial_{\mu} = e_i, \ \partial_{\sigma_{ij}} = E_{ij} \tag{25}$$

where

$$\begin{aligned} e_i &= 1_i \\ E_{i=j} &= 1_{i,i} \\ E_{i\neq j} &= 1_{i,j} + 1_{j,i} \end{aligned} \tag{26}$$

and the differential forms are

$$\begin{aligned} e_i{}^{*} &= 1_i \\ E^{*}{}_{i=j} &= 1_{i,i} \\ E^{*}{}_{i\neq j} &= \frac{1}{2}(1_{i,j} + 1_{j,i}) \end{aligned} \tag{27}$$

so that a duality exists

$$\langle A, B \rangle = \mathrm{tr}(\mathrm{AB}); \ A \in S, \ \mathrm{B} \in S^{*} \tag{28}$$

and the norm of a basis vector is thereby 1.

```
A = {{0, 1}, {1, 0}};
B = 1/2 {{0, 1}, {1, 0}};
A.B // MatrixForm
Tr[A.B]
```

(29)

$$\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \tag{30}$$

$$1 \tag{31}$$

This result implies that we can begin our random walk by drawing a random step from $N\left(0, \sqrt{\frac{g^{ij}(\theta)}{n}}\right)$.

Our results are greatly simplified by always working from the center point, the standard normal $N(0, I)$, because the Fisher metric is orthogonal there. For example, in the bivariate case $N\left(\mu = \{\mu_1, \mu_2\}, \Sigma = \begin{pmatrix} \sigma 1 & \sigma 3 \\ \sigma 3 & \sigma 2 \end{pmatrix}\right)$, the Fisher information matrix is

$$g_{ij}(\mu_1, \mu_2, \sigma_1, \sigma_2, \sigma_3) = \begin{pmatrix} \frac{\sigma 2}{\sigma 1\,\sigma 2 - \sigma 3^2} & \frac{\sigma 3}{-\sigma 1\,\sigma 2 + \sigma 3^2} & 0 & 0 & 0 \\ \frac{\sigma 3}{-\sigma 1\,\sigma 2 + \sigma 3^2} & \frac{\sigma 1}{\sigma 1\,\sigma 2 - \sigma 3^2} & 0 & 0 & 0 \\ 0 & 0 & \frac{\sigma 2^2}{2\left(-\sigma 1\,\sigma 2 + \sigma 3^2\right)^2} & \frac{\sigma 3^2}{2\left(-\sigma 1\,\sigma 2 + \sigma 3^2\right)^2} & -\frac{\sigma 2\,\sigma 3}{\left(-\sigma 1\,\sigma 2 + \sigma 3^2\right)^2} \\ 0 & 0 & \frac{\sigma 3^2}{2\left(-\sigma 1\,\sigma 2 + \sigma 3^2\right)^2} & \frac{\sigma 1^2}{2\left(-\sigma 1\,\sigma 2 + \sigma 3^2\right)^2} & -\frac{\sigma 1\,\sigma 3}{\left(-\sigma 1\,\sigma 2 + \sigma 3^2\right)^2} \\ 0 & 0 & -\frac{\sigma 2\,\sigma 3}{\left(-\sigma 1\,\sigma 2 + \sigma 3^2\right)^2} & -\frac{\sigma 1\,\sigma 3}{\left(-\sigma 1\,\sigma 2 + \sigma 3^2\right)^2} & \frac{\sigma 1\,\sigma 2 + \sigma 3^2}{\left(-\sigma 1\,\sigma 2 + \sigma 3^2\right)^2} \end{pmatrix} \tag{32}$$

and its inverse is

$$g^{ij} = \begin{pmatrix} \sigma_1 & \sigma_3 & 0 & 0 & 0 \\ \sigma_3 & \sigma_2 & 0 & 0 & 0 \\ 0 & 0 & 2\,\sigma_1{}^2 & 2\,\sigma_3{}^2 & 2\,\sigma_1\,\sigma_3 \\ 0 & 0 & 2\,\sigma_3{}^2 & 2\,\sigma_2{}^2 & 2\,\sigma_2\,\sigma_3 \\ 0 & 0 & 2\,\sigma_1\,\sigma_3 & 2\,\sigma_2\,\sigma_3 & \sigma_1\,\sigma_2 + \sigma_3{}^2 \end{pmatrix},$$

So the inverse metric tensor at the standard multivariate normal $N\left(\mu = \{0, 0\}, \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$ is then simply

$$g^{ij}(0, 0, 1, 1, 0) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

The step size will be proportional to the inverse of this matrix, according to the Cramer-Rao inequality. Thus, in the dμ directions, we can take a step of size $N\left(0, \sqrt{1/n}\right)$. In directions where $\Sigma_{i=j}$ the step should be $N\left(0, \sqrt{2/n}\right)$, and the off-diagonals $\Sigma_{i\neq j}$ have steps $N\left(0, \sqrt{1/n}\right)$. This result generalizes to higher dimensions, and simplifies computation because we can generate univariate Gaussian random numbers without having to compute the Fisher information matrix explicitly.

Next, we use this tangent vector as the starting point of the geodesic equation computation. Using Euler's method, and keeping the second-order term that specifies the acceleration due to the Levi-Civita connection, we numerically integrate by taking k steps. Upon reaching the endpoint, we return the mean

and standard deviation (or the lower triangular matrix specified by the Cholesky decomposition $\Sigma = LL^T$, which is a particular example of the square root of a covariance matrix).

The Cholesky decomposition is commonly used for the numerical solution of linear equations $\Sigma x = b$. Using the Cholesky decomposition, we first solve $L y = b$ for y by forward substitution, and then solve $L x = y$ for x by back substitution. It is also common in Monte Carlo simulation, where uncorrelated samples u are generated, and then transformed $L u$ to induce the desired covariance.

In our application, we note that a random variable tranforms according to

$$y \leftarrow L x + b \tag{33}$$

Then the mean of x transforms according to the same

$$\mu_y \leftarrow L \mu_x + b \tag{34}$$

and the variance of x transforms according to

$$\Sigma_y \leftarrow L \Sigma_x L^T \tag{35}$$

In the first leg of the walk, we move from $\{m_0 = 0,\ L_0 = I\}$ to $\{m_1, L_1\}$. The update to the current point then looks like

$$\begin{aligned} m_1 &\leftarrow L_1 m_0 + m_1 \\ L_1 &\leftarrow L_1 L_0 \end{aligned} \tag{36}$$

In the second leg, we move from $\{m_1, L_1\}$ to $\{m_2 L_2\}$. But to avoid recalculating the Fisher information, we will do the walk starting from $N(0, I)$, and then transform it into place. We again begin at $\{m_0, L_0\}$, and end up at $\{m_1'', L_1''\}$

$$\begin{aligned} m_2 &\leftarrow L_1' m_1 + m_1' \\ L_2 &\leftarrow L_1' L_1 \end{aligned} \tag{37}$$

In the third leg, we again begin at $\{m_0, L_0\}$, and end up at $\{m_1'', L_1''\}$

$$\begin{aligned} m_3 &\leftarrow L_1'' m_2 + m_1'' \\ L_3 &\leftarrow L_1'' L_1' L_1 \end{aligned} \tag{38}$$

And so forth until we complete the specified number of legs in the random walk.

Pseudocode

```
GaussianWalk = function[d dimensions, n samplesize, npoints in output, nlegs in
RandomWalk, k steps in ODE]
    for each point in Ensemble:
        Let m = m₀, S = Σ₀ = I_{d×d}
        for each leg in RandomWalk:
            {dμ, dS} ~ N(0, g^{ii}(0,I) / (n×nleg))
            {m₁, L} = EulerAndCholesky[m₀, Σ₀, dμ, dS]
            m ← L·(m + Solve_x[L·x = m₁])
            S ← L·S
        μ_point = m
        Σ_point = S S^T
    return μ, Σ
```

```
EulerAndCholesky = function[{μ, Σ}, {dμ, dΣ}] #Euler's Method
    p = {μ, Σ}
    dp = {dμ, dΣ}
    ddp = Acceleration[p, dp]
    p ← p + k^-1 dp
    for (step in 1:(k - 1)):
        dp ← dp + k^-1 ddp
        ddp = Acceleration[p, dp]
        p ← p + k^-1dp
    {μ, Σ} = p
```

```
    return[μ, CholeskyLT[Σ]]
```

```
Acceleration = function[{μ, Σ}, {dμ, dΣ}] #From Euler-Lagrange
     d² μ = (dΣ · Σ⁻¹ · dμ)
     d² Σ = (dΣ · Σ⁻¹ · dΣ) - (dμ · dμᵀ)
     return[{d² μ, d² Σ}]
```

Mathematica Code

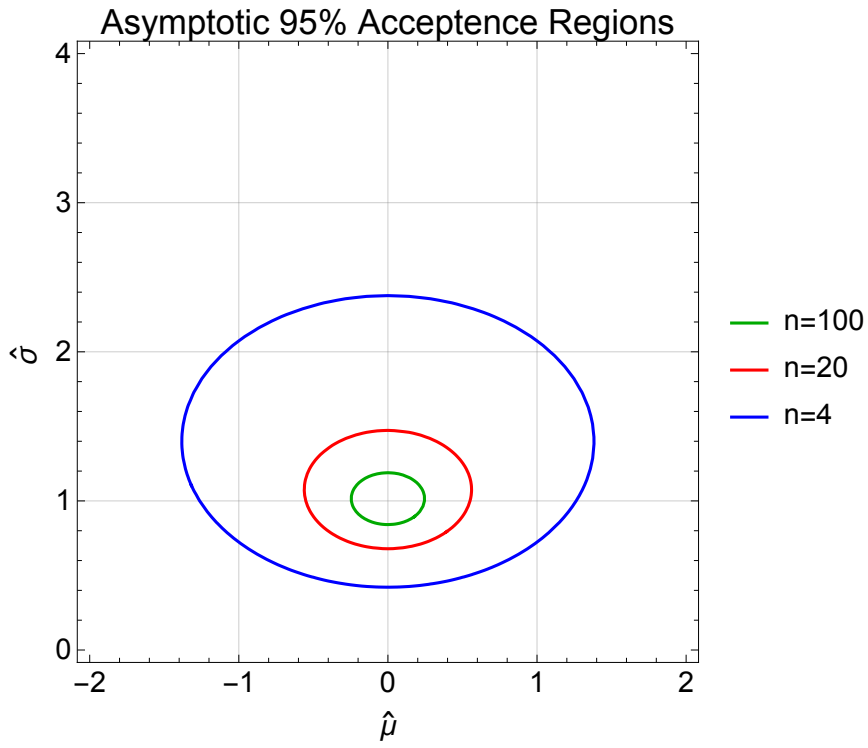
Simulation

Manifold Visualization

Acceptance Region

We would like to compare circles of equal statistical significance. Since the squared information distance is asymptotically χ^2-distributed, $n\,s^2 \sim \chi^2$, we should like to compare circles of equal distance $n\,s^2$. By setting this equal to the critical statistic $n\,s^2 = \chi_c^2$, we can find the circle that defines the asymptotic acceptance region. Using a lookup table, the critical statistic for $\alpha = 0.05$ is $\chi_c^2 = 0.5991$.

Out[93]=



Code

Figure 1

Here we run the simulation of Brownian motion, starting at the standard normal $N(0, 1)$. We generate 1000 estimates, using simulations of several sample sizes. We then compare this to Monte Carlo

maximum likelihood estimates, found using a Gaussian random number generator to generate samples. We also compare to a bootstrap simulation, where an original sample is resampled, and each resample gets treated with random Gaussian noise $N\left(0, 1\Big/\sqrt{n}\right)$. This essentially represents Brownian motion in the data space, rather than the model space.

The Brownian motion fails to replicate the Monte Carlo distribution. The Brownian motion result is consistent with the concentric geodesic circles as $\alpha = 0.05$ acceptance regions, with few points lying outside the boundaries. In this way, it more or less represents an extension of the central limit theorem. The Monte Carlo result, on the other hand, drifts towards lower estimates of standard deviation, in a clear departure from the central limit theorem.

We must now more fully characterize how Brownian motion fails to reproduce the Monte Carlo. Perhaps we can arrive at a modification of Brownian motion that replicates the non-asymptotic distribution found by Monte Carlo.

Figure 1a: Brownian Motion

Out[99]=

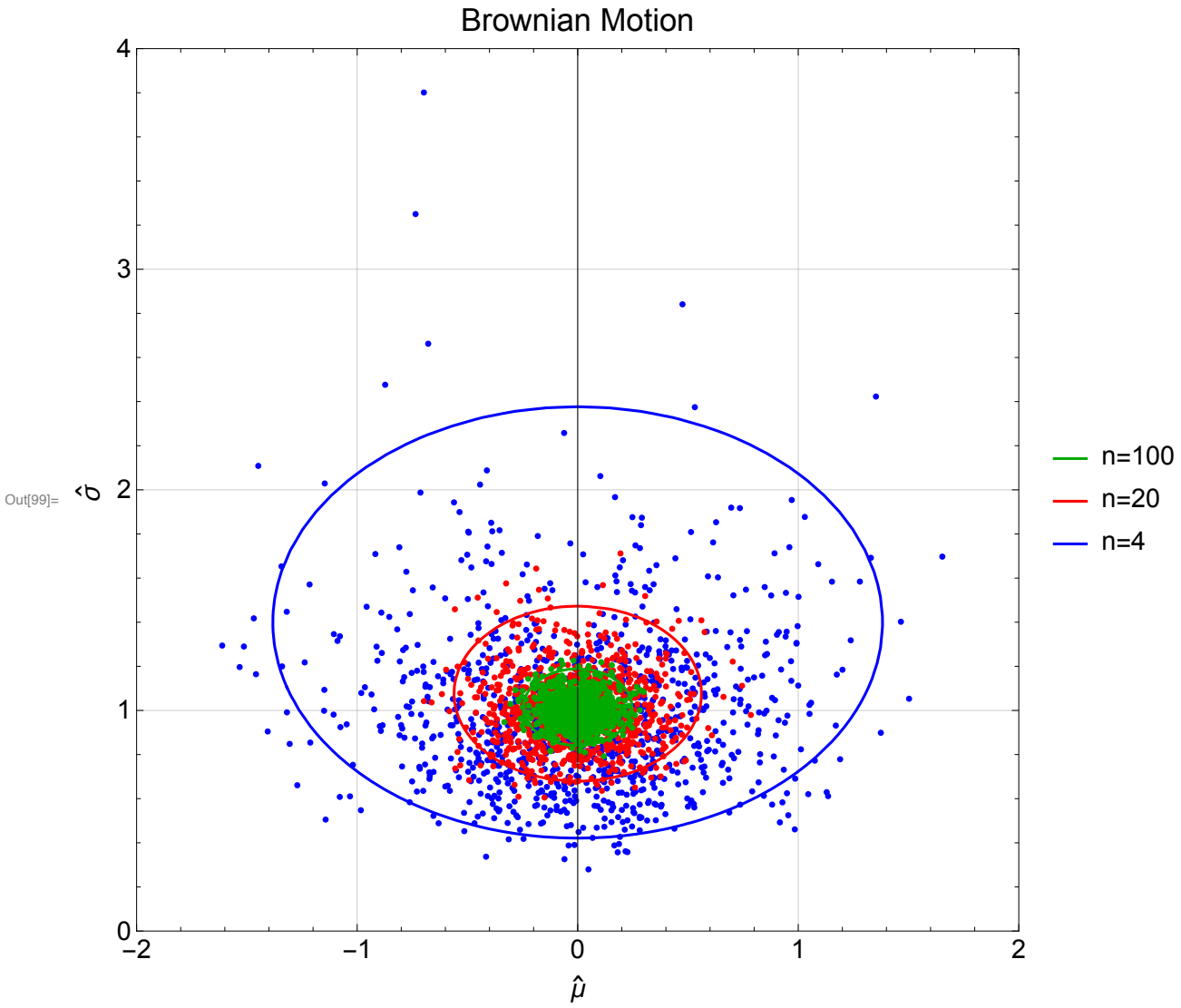


Figure 1b: Monte Carlo

In[100]:=
```
Show[plot0012]
```

Out[100]=

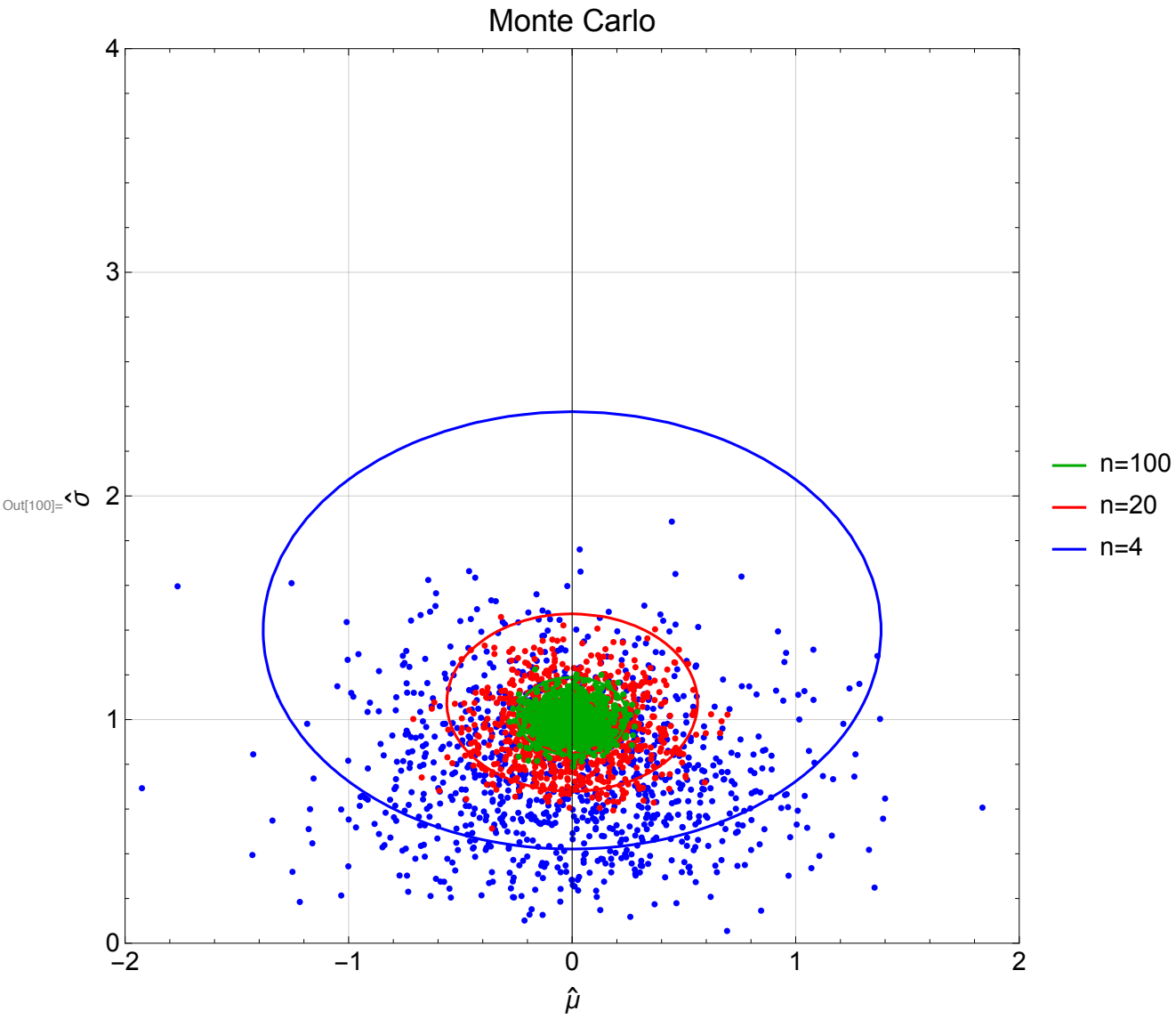
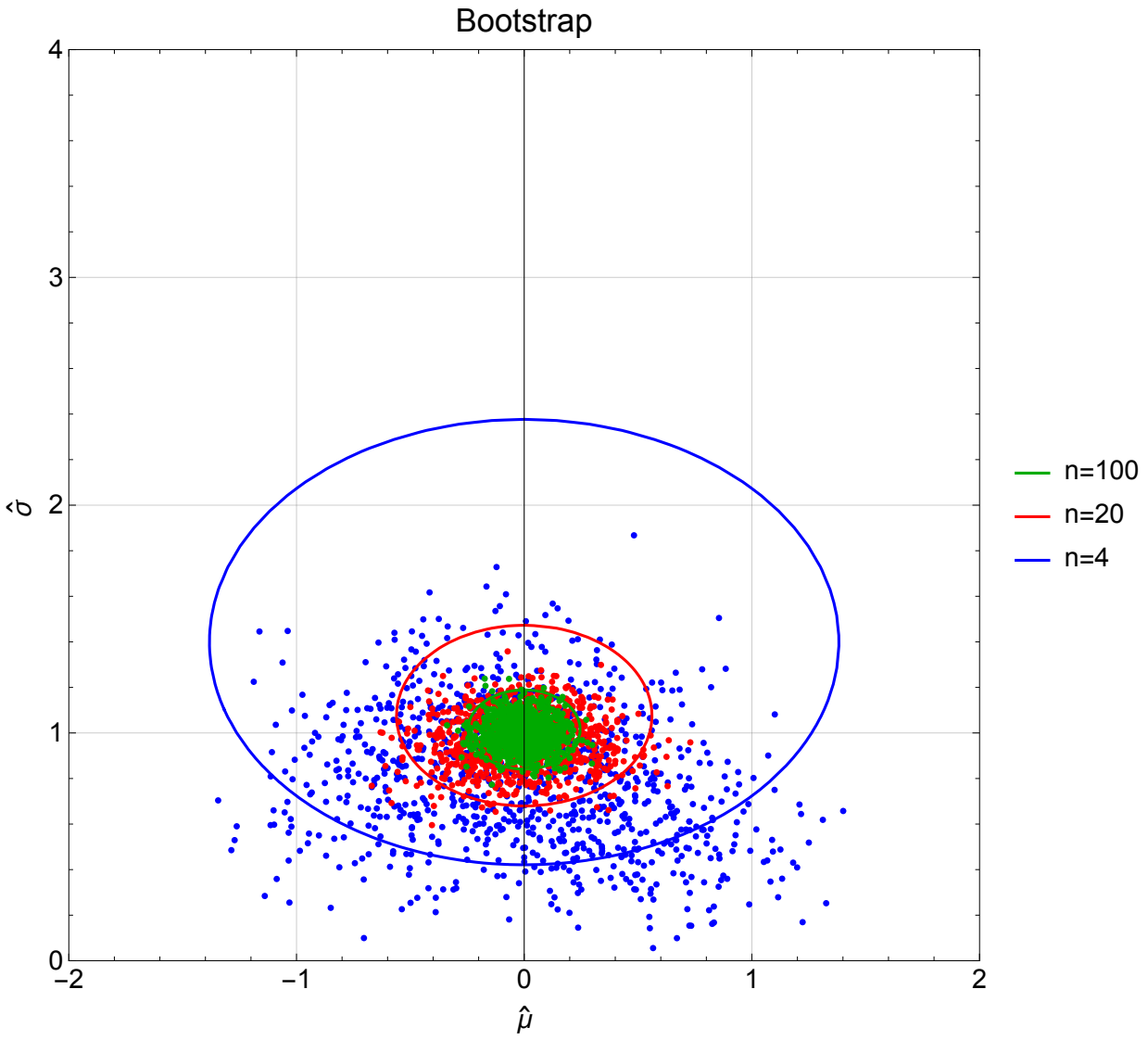


Figure 1c: Bootstrap

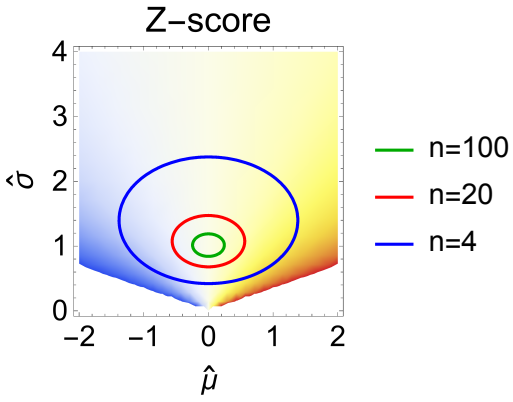
Out[275]=



Z-score

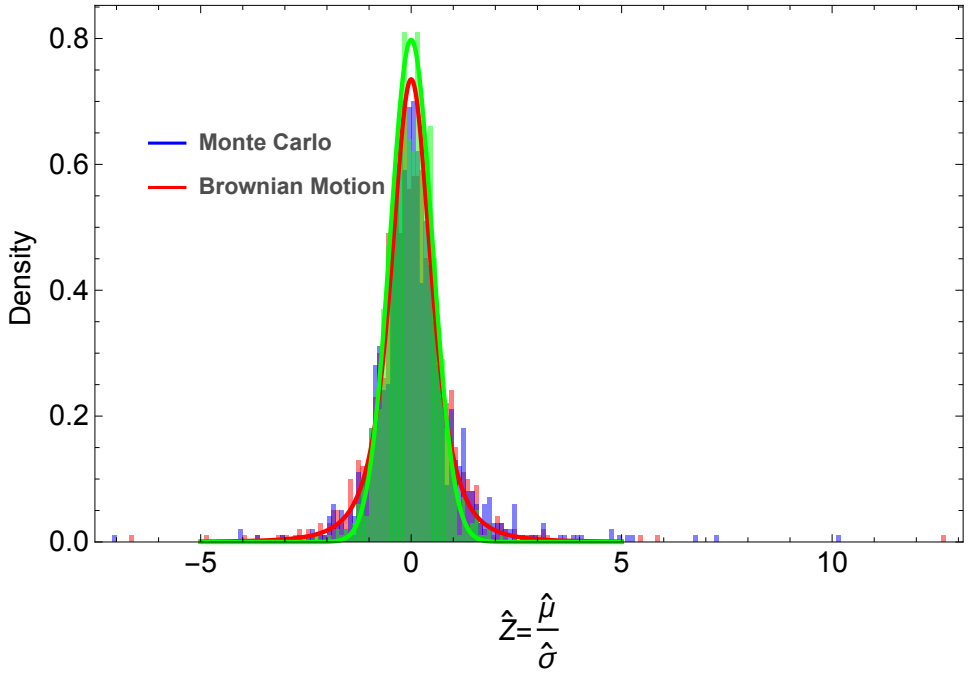
We can instead label each coordinate (μ, σ) by a single number formed from the ratio, $z = \mu / \sigma$.

Out[102]=



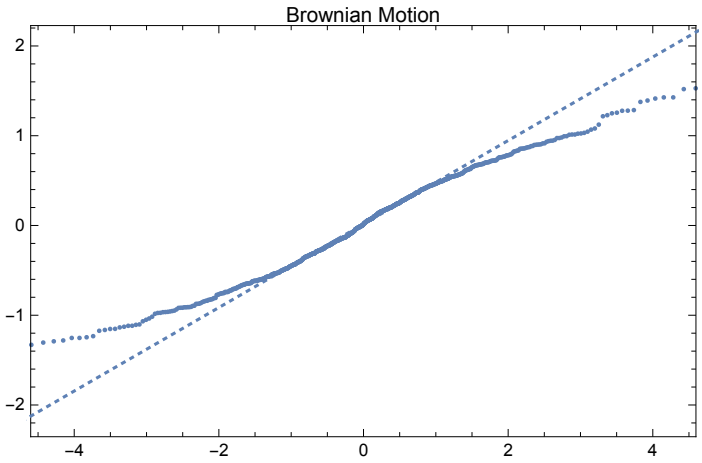
According to the central limit theorem, $\sqrt{n}\ \frac{(\hat{\mu}-\mu)}{\sigma} \sim N(0, 1)$ asymptotically. We can see the difference between Brownian motion and Monte Carlo for the small sample size $n = 4$: the Brownian motion is unable to reproduce the t – distribution like the Monte Carlo. Instead, it remains consistent with the central limit theorem.

Out[291]=

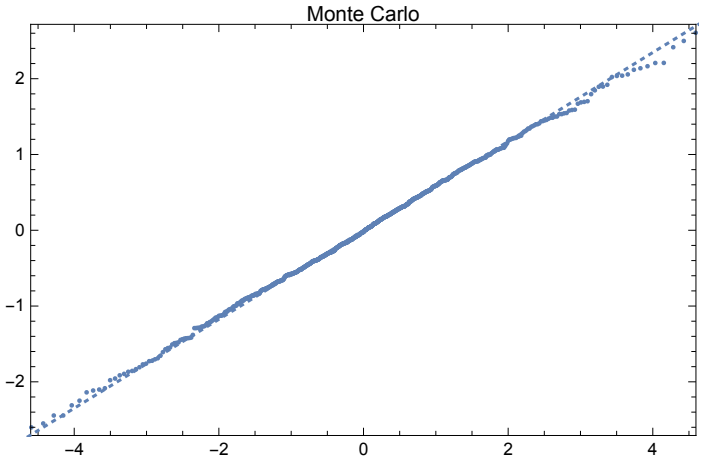


Looking at the Q-Q plots, we see how the Monte Carlo perfectly reproduces the t-distribution. We also notice that the bootstrap distribution is skewed, which is inherited from the original small sample.

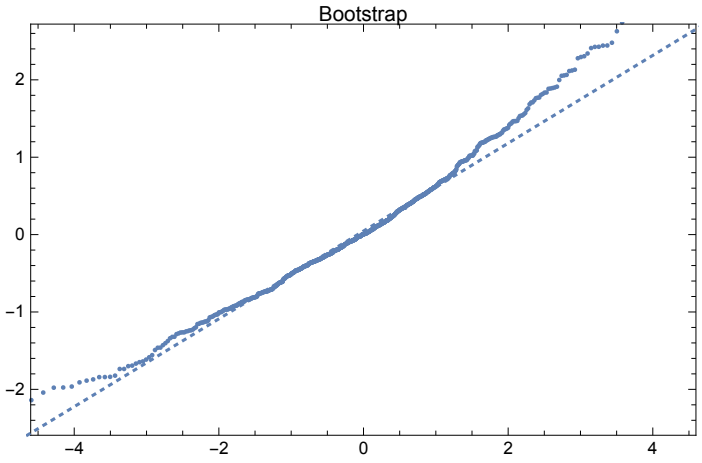
Out[300]=



Out[301]=



Out[302]=



Marginal Distributions

Code

Figure 2

To better understand where the Brownian motion fails to replicate the Monte Carlo, we examine the marginals of the sampling distribution.

Here we compare the sampling distribution of the mean, and the sampling distribution of the standard deviation, across the Brownian motion simulation, the Monte Carlo simulation, and the bootstrap simulation. The theoretical distributions are shown using solid curves.

The sampling distribution of the means are identical and in agreement with the theoretical distribution. However, Brownian motion fails to reproduce the sampling distribution of the standard deviation. It produces much fewer events at very small standard deviation, and more frequent events at very large standard deviations.

The reason for this is that the sampling distribution of the variance is actually a χ^2 distribution, which becomes Gaussian asymptotically. Thus, starting from the asymptotic Gaussian, we cannot recover the χ^2 characteristic from just the Riemannian geometry.

Figure 2a: Brownian Motion

Out[126]=

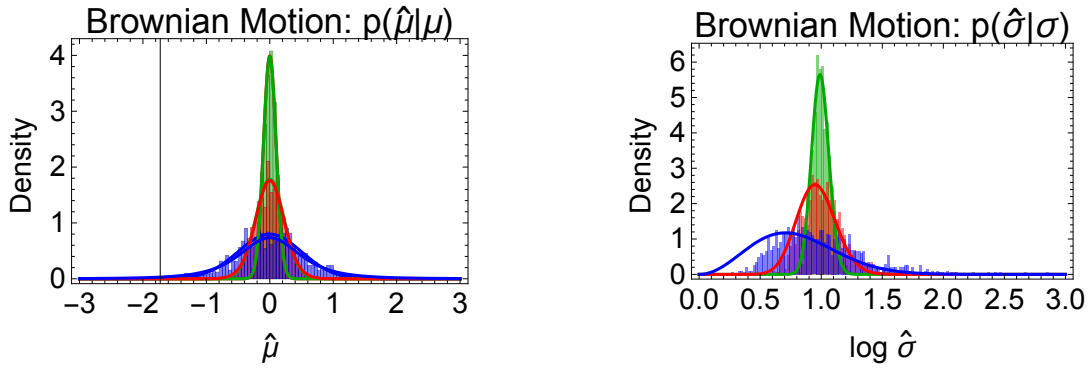


Figure 2b: Monte Carlo

Out[127]=

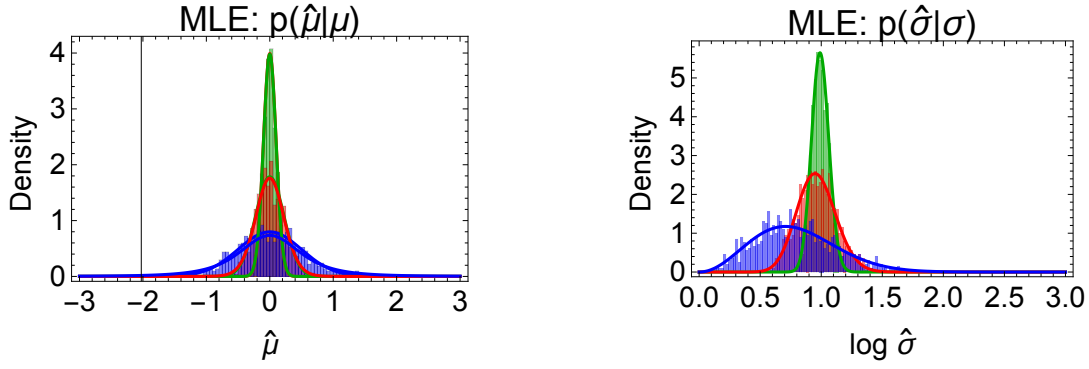
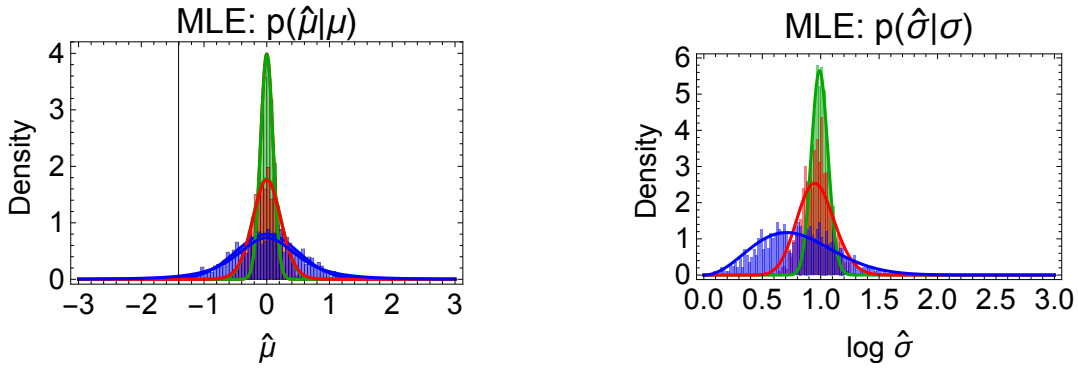


Figure 2c: Bootstrap

Out[285]=



Distribution of Distances

Brownian Motion

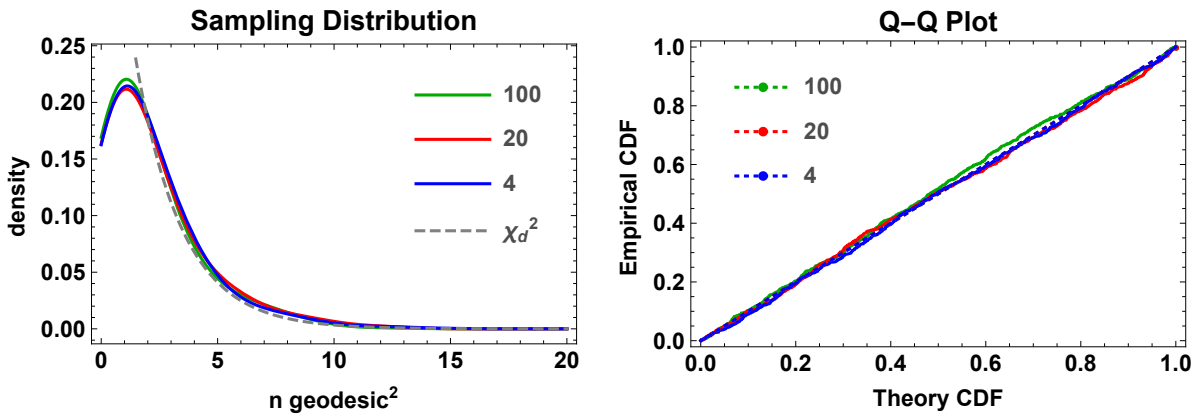
Figure 3a

Now we will compute the sampling distribution of the information distance measures using Brownian motion. Asymptotically, a number of distance measures follow a χ^2distribution, including the likelihood ratio and Wald statistic. Here, we compute the sampling distribution of the geodesic distance, (the Euclidean distance $\left(d\mu^2 + 2\,d\sigma^2\right)$, the Kullback-Leibler divergence, and the symmetrized $J-$divergence not shown).

On the left we show a kernel density estimate of the sampling distribution compared to the theoretical χ^2. On the right, we compare the quantiles of the simulated distribution and the χ^2 distribution.

The Brownian motion keeps close agreement with the asymptotic distribution. If the hyperbolic curvature had an effect on the Brownian motion, we would expect to see deviations here, but we do not.

Out[161]=

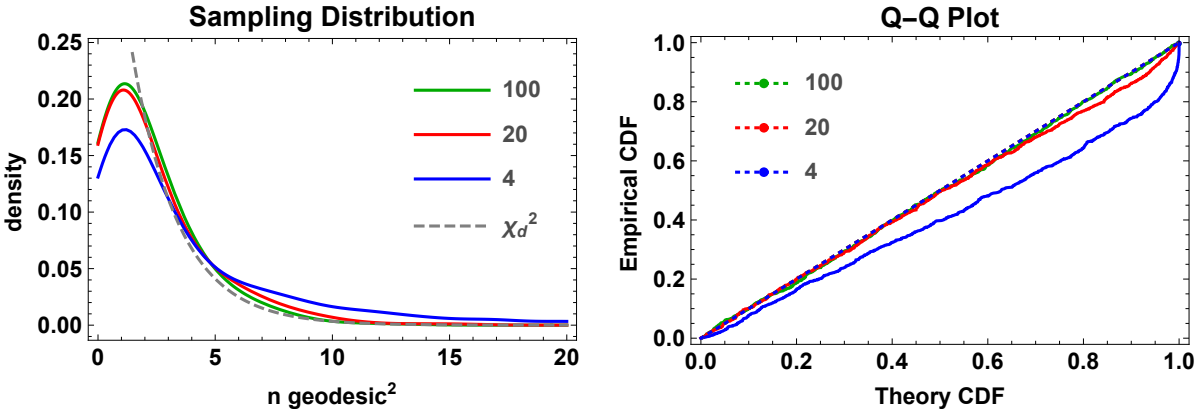


Maximum Likelihood Estimates from Monte Carlo

Figure 3b

Here we generate the same comparisons for the Monte Carlo simulations. We can clearly see the deviations from asymptotic theory. (Interestingly, the symmetrized J – divergence is least affected, not shown).

Out[195]=

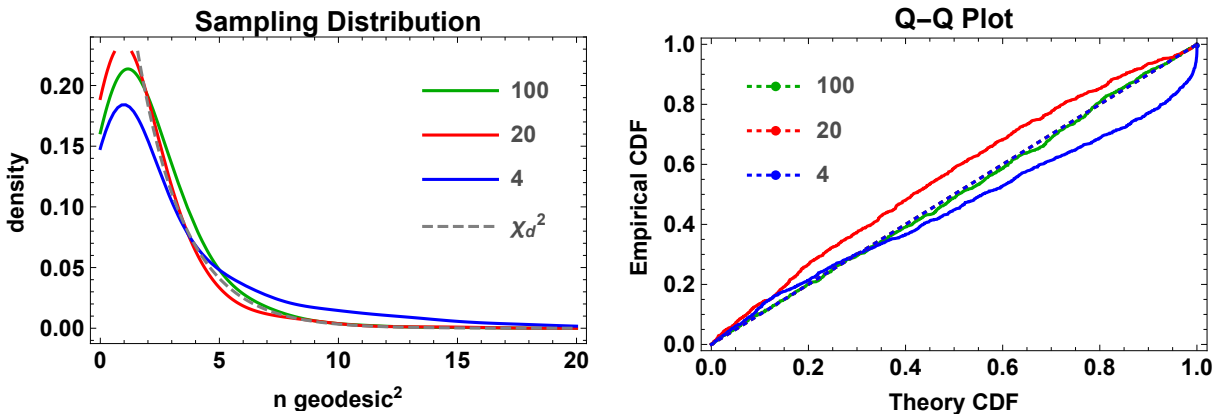


Maximum Likelihood Estimates from Bootstrap

Figure 3c

Here we generate the same comparisons for the bootstrap simulations. We see clear deviations that can be explained by properties of the original sample.

Out[336]=



Conclusion

It looks like the non-asymptotic distribution does not arise due to the Riemannian curvature of the manifold. In fact, it appeared to have no discernable effect. Rather, the addition of a negative drift term in the Brownian motion, inversely proportional to the sample size, helps move towards the Monte Carlo distribution. This is because the sampling distribution of the variance is χ^2distributed for small sample sizes, thus shifting the mean variance downward. We also found that doing Brownian motion in the data space, through a smoothed bootstrap, can give similar results as the Monte Carlo, even at small sample sizes, although the results inherit any peculiarities in the original sample.