Applications of Intrinsic Gaussian Distributions on Symmetric Spaces

Dustin Pluta

UCI, Dept. of Statistics

Summary of Projects

1. **Adaptive Mantel Test for Association Testing in Imaging Genetics Data.** *(resubmitting)*
   Zhaoxia Yu, Hernando Ombao (KAUST), Chuansheng Chen (UCI), Gui Xue (Beijing Normal University)
2. **Latent Factor Gaussian Process Model for Non-stationary Time Series.** *(submitted to NeurIPS 2019)*
   Lingge Li, Babak Shahbaba, Norbert Fortin (UCI), Pierre Baldi
3. **Random Effects Mediation Model for Imaging Genetics Studies.** *(in progress)*
   Zhaoxia Yu
4. **Statistical Applications of Harmonic Analysis on Symmetric Spaces.** *(in progress)*
   Motivated by discussions with Moo Chung (Univ. of Wisconsin), Hernando Ombao, Andrew Holbrook.

LFGP Model

- The central idea of the Latent Factor GP model is to map the covariance process into Euclidean space using the matrix logarithm.

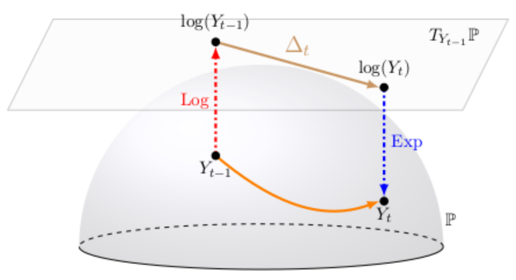


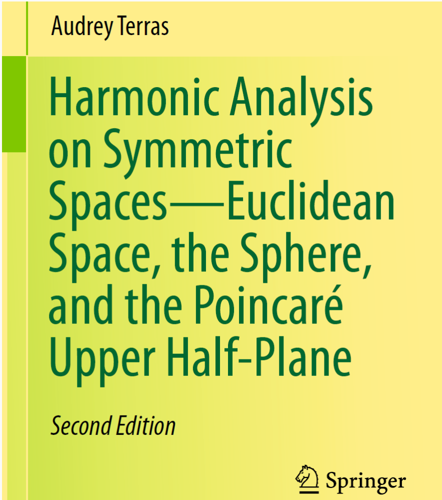
Figure 1: The matrix logarithm maps the space of symmetric positive definite (SPD) matrices to the tangent space at the identity.

Overview

1. Definition of the intrinsic Gaussian distribution as the solution to the heat equation.
2. The intrinsic Gaussian on hyperbolic space.
3. The differences of the Log-Euclidean Gaussian and the intrinsic Gaussian for SPD matrices.

Reference

Essentially all of the theoretical results discussed here are presented in *Harmonic Analysis on Symmetric Spaces, A. Terras 2013/2016.*



Symmetric Spaces

- *Symmetric spaces* are an important class of Riemannian manifolds with a special structure
- Examples of symmetric spaces are
  - $\mathbb{R}^n$: Euclidean space
  - S^{n-1}: Sphere
  - H: Poincaré upper half plane
  - $\mathcal{P}_n$: space of $n \times n$ symmetric positive definite matrices

The Heat Equation on $\mathbb{R}^n$

Definition: The Heat Equation on $\mathbb{R}^n$

$$\begin{cases} \frac{\partial E}{\partial t} - a^2 \Delta E = \delta(x, t) \qquad t > 0, x \in \mathbb{R}^n \\ \Delta = \sum_{j=1}^{n} \frac{\partial^2}{\partial x_j^2} \end{cases}$$

- Δ is the Laplacian operation on $\mathbb{R}^n$,
- a is a constant related to the diffusion rate.

Definition: Fourier Transform

$$\hat{f}(t) = \int_{-\infty}^{\infty} \exp\{-2\pi i t x\} f(x)\, dx.$$

Theorem. Properties of the Fourier Transform

1. $D^a \hat{f} = \widehat{(-2\pi i x)}^a f$
2. $\widehat{D^a} f = (2\pi i x)^a \hat{f}$
3. Convolution: $\widehat{f * g} = \hat{f} \cdot \hat{g}$
4. Fourier Inversion: $\hat{\hat{f}}(x) = f(-x)$

The Heat Equation on $\mathbb{R}^n$

To solve the heat equation, apply the Fourier transform to change the problem into an ODE:

$$\frac{d\hat{E}}{dt} + (2\pi a\|x\|)^2\hat{E} = 1$$
$$\hat{E}(t) = \exp\{-(2\pi a x)^2 t\} \cdot \mathbf{1}[t > 0]$$
$$E(x,t) = \hat{\hat{E}}(-x,t) = \exp\{-\widehat{(2\pi a\|x\|})^2 t\} \cdot \mathbf{1}[t > 0]$$
$$E(x,t) = (2a\sqrt{\pi t})^{-n} \exp\left\{-\frac{1}{2}\frac{\|x\|^2}{4a^2 t}\right\} \cdot \mathbf{1}[t > 0]$$

Fundamental Solution to the Heat Equation on $\mathbb{R}^n$

Setting $t = 1$ and $v = 2a$, we write

$$G_v(x) := (2\pi v^2)^{-n/2} \exp\left\{-\frac{1}{2}\frac{\|x\|^2}{v^2}\right\}$$

Let $\mathcal{M}$ be a symmetric space with Riemannian metric $d\mu$.

Definition: The Intrinsic Laplacian

The *intrinsic Laplacian operator* Δ^* on $\mathcal{M}$ is the operator obtained by transforming Δ according to the change of coordinates $\mathbb{R}^n \to \mathcal{M}$.

Definition: The Intrinsic Gaussian

The *intrinsic Gaussian* $G_v(x)$ for $(\mathcal{M}, d\mu)$ is the fundamental solution to the heat equation for the intrinsic Laplacian.

The Heat Equation on the Poincaré Upper Half-Plane

Definition: Poincaré Upper Half-plane

The *Poincaré* upper half-plane is the space

$$H = \{x + iy \mid x, y \in \mathbb{R}, y > 0\}, \quad \text{where} \quad i = \sqrt{-1}.$$

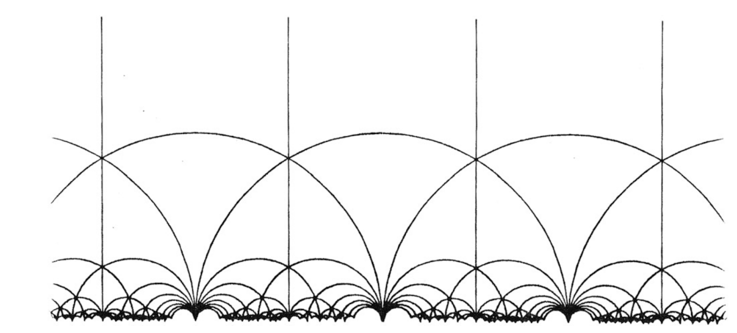


Figure 2: Tesselation of H for $SL(2, \mathbb{Z})$. [Terras 2013]

The Heat Equation on the Poincaré Upper Half-Plane

The Poincaré Disk model is also commonly used to represent $\mathbb{H}$.

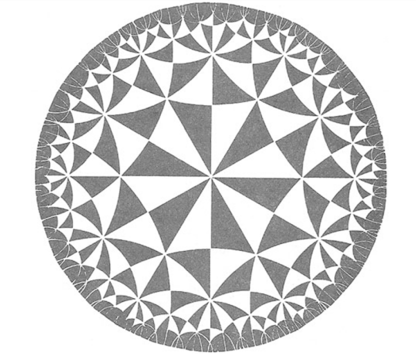


Figure 3: Illustration by Coxeter that inspired Escher. [Terras 2013]

Definition: Arc Length on H

The *arc length* on H is given by

$$ds^2 = y^{-2}(dx^2 + dy^2)$$

Definition: Geodesic polar coordinates on H

$$x = y \sinh r \sin(2\theta), \quad y = (\cosh r + \cos(2\theta) \sinh r)^{-1}$$

The Heat Equation on the Poincaré Upper Half-Plane

Definition: Heat Equation on H

$$\Delta^* u = \frac{\partial u}{\partial t} = y^2 \Delta u(x+iy,t), \quad u(z,0) = \delta(z)$$

We can solve the hyperbolic heat equation to find an expression of the hyperbolic Gaussian density function.

Density of the Hyperbolic Gaussian

$$G_t(e^{-r}i) = \frac{\sqrt{2}}{(4\pi t)^{3/2} e^{t/4}} \int_r^\infty \frac{b e^{-b^2/4t}\, db}{\sqrt{\cosh b - \cosh r}}$$

Hyperbolic Gaussian

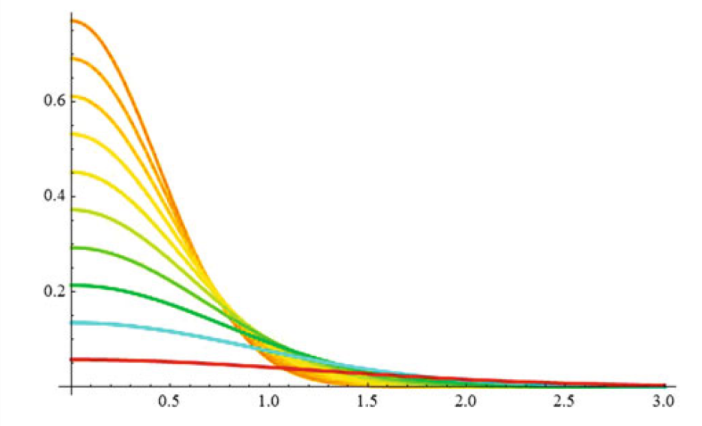


Figure 4: Hyperbolic Gaussian density along the y-axis for various values of t. [Terras 2013]

Hyperbolic Gaussian

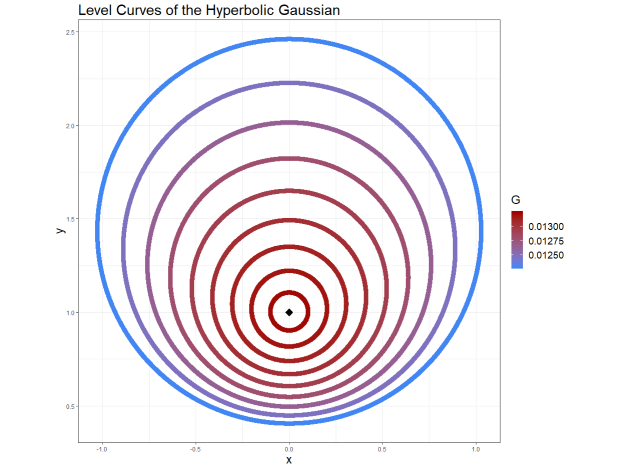


Figure 5: Level curves of the hyperbolic Gaussian, upper half-plane model.

Hyperbolic Gaussian

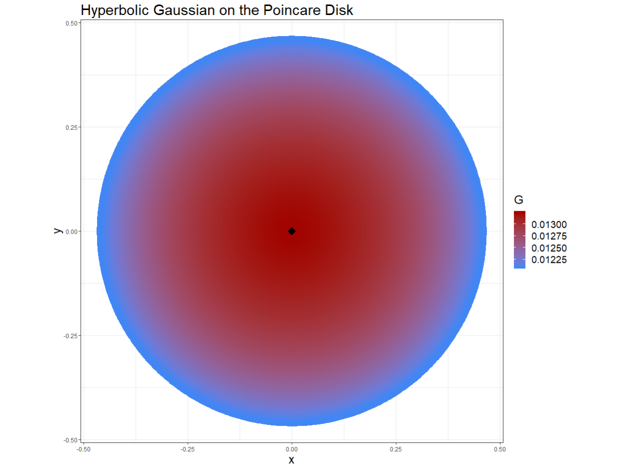


Figure 6: Density of the hyperbolic Gaussian, Poincaré disk model.

Central Limit Theorem on the Hyperbolic Upper Half-Plane

CLT on H

Suppose that $\{Z_n\}_{n\geq 1}$ is a sequence of iid $SO(2)$-invariant random variables in $\mathbf{H}$ with density $f(z)$. Let $S_n = Z_1 \circ \cdots \circ Z_n$ be normalized as $S_n^{\#}$. Then for measurable sets $A \subset H$ we have

$$\int_A f_n^{\#}(z)\, d\mu \to \exp(d/4) \int_A G_{d/4}(z)\, d\mu \quad \text{as } n \to \infty,$$

where G_c is the Gaussian on $\mathbf{H}$, and with $d = 2\pi \int_{r>0} f_Z(e^{-r}i) r^2 \sinh r \, dr.$

Application to Adaptive Mantel Test

Application to the Mantel Test and Metric Learning

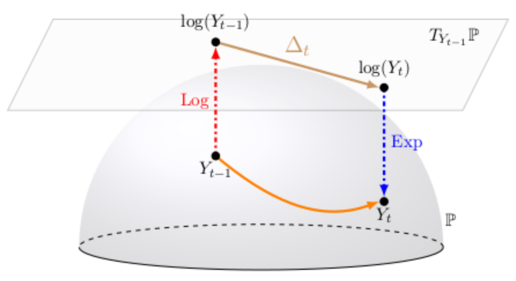
Using the intrinsic Gaussian G on H (or $SL(n,\mathbb{R})/SO(n)$ for $n > 2$), we can state a Bayesian model for the observed data

$$
\begin{aligned}
&Y \sim \mathcal{N}(0, \sigma^2 X^T W X + \sigma_\varepsilon^2 I_m) \\
&W \sim G_{\mathcal{SP}_n}(I, v) \\
&\sigma^2 \sim Gamma(a, b) \\
&\sigma_\varepsilon^2 \sim Gamma(c, d),
\end{aligned}
$$

where values are chosen for v (prior variance of W) and a, b, c, d (according to prior beliefs about variance of Y and $h^2 = \sigma^2/(\sigma^2 + \sigma_\varepsilon^2)$).

Application to Modeling Covariance Matrices

- The Gaussian on $\mathcal{P}_d$ can be derived using a similar strategy as for H.
- The Log-Euclidean and intrinsic Gaussians agree at the identity, and diverge as one moves towards the boundary.
- Since covariance matrices resulting from real data are often approximately low-rank, the differences between the two distributions could be substantial in practice.



Summary

1. For many symmetric Riemannian manifolds, we can define an intrinsic Gaussian distribution as the solution to the heat equation for the intrinsic Laplacian.
2. The intrinsic distributions may give better behavior near the boundary of the manifold than Gaussians defined on the tangent space.
3. The intrinsic Gaussian shares many properties of the usual normal distribution, including a Central Limit Theorem.
4. Consequently, if random errors "combine" according to the Riemannian geometry of the manifold, the combined errors will approximately follow an intrinsic Gaussian distribution.

Thanks!

Reference

A. Terras. *Harmonic Analysis on Symmetric Spaces and Applications.* 2013 Springer.

Slides available at:

`https://github.com/dspluta/HASS/`

Figure 7: Escher, *Circle Limit III*