CLUSTERING FINANCIAL RETURN DISTRIBUTIONS USING THE FISHER INFORMATION METRIC

STEPHEN TAYLOR

New Jersey Institute of Technology
184-198 Central Ave., Newark, NJ 07103
Email: smt-at-njit-dot-edu

Abstract. Information geometry provides a correspondence between differential geometry and statistics through the Fisher information matrix. In particular, given two models from the same parametric family of distributions, one can define the distance between these models as the length of the geodesic connecting them in a Riemannian manifold whose metric is given by the model's Fisher information matrix. One limitation that has hindered the adoption of this similarity measure in practical applications is that the Fisher distance is typically difficult to compute in a robust manner. We review such complications and provide a general form for the distance function for one parameter models. We next focus on higher dimensional extreme value models including the generalized Pareto and generalized extreme value distributions that will be used in financial risk applications. Specifically, we first develop a technique to identify the nearest neighbors of a target security in the sense that their best fit model distributions have minimal Fisher distance to the target. Second, we develop a hierarchical clustering technique that utilizes the Fisher distance. Specifically, we compare generalized extreme value distributions fit to block maxima of a set of equity loss distributions and group together securities whose worst single day yearly loss distributions exhibit similarities.

1. Introduction

Quantifying the similarity between two probability distributions is a central component of a variety of applications in the statistics, quantitative finance, and engineering literature. Such measures are one of the main inputs into classification, clustering, and optimization algorithms. In [16], the author compiles and categorizes a collection of sixty-five similarity measures and stresses that the identification of a proper notion of similarity for the application at hand is crucial to its success. Financial securities are often compared based on the values of statistics associated with their return distributions. For example, the historical volatility of a stock is defined to be the standard deviation of its daily return distribution and the 95% Value-at-Risk is the $q = 0.05$ quantile of this distribution. Our main aim will be to extend beyond return distribution summary statistic based comparisons of financial securities by utilizing their full return distributions when assessing their similarity. Specifically, we will consider the Fisher information distance on a model space of distributions which in turn will be used in both nearest neighbor and clustering applications that focus on the entire risk profiles of financial securities.

Grouping related probability distributions according to their similarity has been considered in a number of clustering applications. In [30], the authors develop a Kullback-Leibler divergence based clustering technique. More general Bergman divergences are used in [7] for clustering applications. The authors of [55] develop several distribution based clustering techniques using the L^1 norm between distributions as a similarity measure. In [50, 24, 25], the authors explore information geometry based clustering methods using the Fisher-Rao distance. The authors also mention that the development of clustering algorithms based upon probability distributions is a relatively unexplored area. The promise of these techniques is that they take into account the entirety of the distribution of the objects being clustered which may provide a more precise notion of similarity than traditional measures that use distribution summary statistics for comparison. For example, two distributions with equal means and variances can have significantly different probability density functions. A distribution based similarity measure should be able to distinguish such differences whereas a measure based on just the mean and variance statistics would conclude that these two models are identical.

Clustering techniques applied to financial return data predominately rely upon single statistics associated with these time series rather than a model for a full return distribution [14]. Specifically, similarity measures are defined in terms of functions of the correlation between these series. Each of [10, 11, 54], utilize correlation distance measures to develop hierarchical clustering algorithms in order to explore the network structures in

equity markets. These techniques take only the return time series of a collection of stocks as input and are able to identify hierarchical sector structure solely from this information. The authors of [20], utilize related clustering techniques to improve traditional index tracking strategies. Commodity network properties are explored with analogous methods in [53]. Correlation based clustering methods are useful for extracting network and sector structures from financial return data; however, have limited utility when comparing the risk profiles of these securities which is a strength of distributional based similarity measures.

Our aim is to develop a similarity measure to compare the return distributions of financial time series. Specifically, given end of day return data for a collection of equities, we will use maximum likelihood estimation or related model fitting techniques to identify a single model from a parametric family of distributions with data. Given the parameters of two such models, we consider the question of defining a distance between the models that is a function of these parameters. We use the Fisher information distance which is constructed from the Fisher information metric of the distribution family for this purpose. The main practical focus will be to group loss distributions by their risk profiles; specifically, how similar are the left tails of the two distributions? With this in mind, we focus on the generalized Pareto and generalized extreme value distributions from extreme value theory. This distance function may be used as an input into a nearest neighbor or hierarchical clustering method both of which provide a risk based clustering of financial securities analogous to the previously mentioned correlation methods.

The Fisher information metric and its associated distance are central concepts in the subject of information geometry [2, 3, 4, 5] which draws upon ideas from statistics, differential geometry, and information theory to study the geometric structure of statistical models. The main connection between a family of statistical models and differential geometry is the identification of the positive definite symmetric Fisher information matrix with a Riemannian metric on the space of all such distributions. The distance function associated with this metric is then taken to be a measure of similarity between models. In addition, a characterization theorem due to Chentsov [17] states that this is the unique metric one can place on a space of probability distributions that has certain desirable statistical properties explained further below.

One of the main limitations of the Fisher distance which has lead to its somewhat slow adoption in practical applications is the difficulty encountered in its computation. It is difficult to develop robust methods to directly solve the geodesic equations since the form of these equations, and in particular their singularity structure, is model dependent; we provide an example of this issue below. An alternative technique is provided in [12, 13], where the author developed a second order formula using Riemann normal coordinates that locally approximate geodesic distances. This technique may be merged with graph theoretic methods to develop a global geodesic distance approximation algorithm. However, coordinate transformations to and from normal coordinates make this technique quite complicated. We will utilize simple robust methods for computing geodesic distances in model spaces of dimension two to four developed in [39]. Here, the authors construct a Hamiltonian Fast Marching technique that solves an Eikonal equation for the distance function of a Riemannian manifold. This technique is robust and works well in many singular geometries and complex manifolds that the authors feature. The method may be roughly viewed as a generalization of Dijkstra's algorithm applied to a discretized version of the Riemannian manifold. We have found that it is well suited for the computation of geodesic distances for Fisher metric geometries, and we use it in the distance computations described in the below higher dimensional applications.

We provide several novel contributions to the information geometry, clustering, and financial risk literature. First, we develop a closed form expression for the Fisher information distance between one dimensional models. This may be used to measure the distance between members of a distribution family that depends on a single parameter or higher dimensional models where all but a single parameter is fixed. Next, we compute the components of the Fisher information matrix for the generalized Pareto and generalized extreme value distributions. Both are well suited for modeling the left tail of return distributions of financial securities as well as the distribution of the worst single day loss over a yearly period [11, 26, 29, 36, 37]. In particular, modeling the left tail of the distribution is important for identifying securities with similar risk metrics such as their Value-at-Risk and expected shortfall. We finally provide nearest neighbor and hierarchical clustering applications that group together equity return distributions based on the Fisher distance between them. This results in a new risk focused clustering technique of financial securities.

This article is organized as follows. In Section 2, we review relevant background in information geometry as well as complications that arise in developing robust techniques to directly solve the geodesic equations. In Section 3, we develop a closed form expression, up to quadrature, for the Fisher information distance of one dimensional models and provide several examples. In Section 4, we examine difficulties involved in computing geodesic distances in higher dimensional models, and compute the Fisher information matrix for univariate normal distributions, the generalized Pareto distribution, and the generalized extreme value

distribution. We also compare the Fisher distance to the Kullback-Leibler divergence in the case of univariate Gaussian models. In Section 5, we utilize these expressions in risk focused nearest neighbor and worst daily loss clustering applications for equity securities. Finally, in Section 6, we conclude and summarize future directions to explore.

2. Information Geometry Background

Information geometry is a relatively new discipline which combines statistics and differential geometry together in the study of statistical model Riemannian manifolds, c.f. [2, 3, 4, 5, 46]. Our main aim is to utilize the Fisher information distance on such a manifold as a similarity measure between probability distributions in subsequent applications. With this end in mind, we define the Fisher distance by first considering a random variable with probability density function $\phi(x;\theta)$ for $x \in D \subset \mathbb{R}^m$ with parameters $\theta = (\theta_1, \ldots, \theta_n)$. Denote the set of all probability distributions for this model by

$$S_\phi = \{\phi(x;\theta) | \theta \in \Theta\}, \tag{1}$$

where here $\Theta \subset \mathbb{R}^n$ is the set of admissible values of the parameters. We refer to an element of S_ϕ by its model parameters. Given $\theta_1, \theta_2 \in \Theta$, the Fisher distance is a function $d(\theta_1, \theta_2) : \Theta \times \Theta \to \mathbb{R}^+$ constructed from the log-likelihood of this model

$$l_\theta(x) \equiv \ln \phi(x;\theta). \tag{2}$$

The components of the Fisher information matrix $I_{ij}(\theta)$ are defined as the expected value of the negative Hessian matrix of the log likelihood of ϕ.

$$I_{ij}(\theta) = \mathbb{E}_\theta \left[\frac{\partial l_\theta}{\partial \theta_i} \frac{\partial l_\theta}{\partial \theta_j} \right] = \int_D \frac{\partial l_\theta}{\partial \theta_i} \frac{\partial l_\theta}{\partial \theta_j} \phi(x;\theta) dx = -\int_D \frac{\partial^2 l_\theta}{\partial \theta_i \partial \theta_j} \phi(x;\theta) dx. \tag{3}$$

The Fisher information matrix is positive definite symmetric and also has many applications in estimation and information theory, most notably the Cramér-Rao bounds [19, 49] that enable one to establish lower bounds on the variance of unbiased estimators of deterministic model parameters.

In order to construct a distance function from the Fisher information matrix, we identify I_{ij} with a Riemannian metric g_{ij} referred to as the Fisher information metric on the model space S_ϕ. This identification was first made in [49] and is the cornerstone of the subject of information geometry. Given the Riemannian manifold (g, S_ϕ), we may define a distance function in terms of the Fisher information metric in the usual manner in Riemannian geometry. In particular, the distance between two models $\theta_1, \theta_2 \in \Theta$ is given by the arc-length of the geodesic that connects these two points. More specifically, the components of such a geodesic $\gamma(t) = (\gamma^1(t), \ldots, \gamma^m(t))$, with $\gamma(0) = \theta_1$ and $\gamma(1) = \theta_2$ are determined by minimizing the length functional

$$L[\gamma] = \int_0^1 \sqrt{g(\dot\gamma(t), \dot\gamma(t))} dt, \tag{4}$$

over the space of all continuously differentiable curves $\gamma : [0,1] \to S_\phi$ that satisfy $\gamma(0) = \theta_1$ and $\gamma(0) = \theta_2$. One can show that this is equivalent to solving the second order quasi-linear system of geodesic equations boundary value problem

$$0 = \nabla_{\dot\gamma}(\dot\gamma) = \ddot\gamma^i(t) + \sum_{i,j=1}^n \Gamma^i_{jk} \dot\gamma^j(t) \dot\gamma^k(t), \quad \gamma(0) = \theta_1, \quad \gamma(1) = \theta_2, \quad i = 1, \ldots, n, \tag{5}$$

where here the Christoffel symbols are defined by

$$\Gamma^i_{jk} = \frac{1}{2} \sum_{m=1}^n g^{im} \left(\frac{\partial g_{mj}}{\partial x^k} + \frac{\partial g_{mk}}{\partial x^j} - \frac{\partial g_{jk}}{\partial x^m} \right), \tag{6}$$

assuming that γ is parameterized on the unit interval. Given a solution to these equations $\hat\gamma$, we can compute the distance between the two models by evaluating $L[\hat\gamma]$.

The main motivation for using the Fisher information distance as a similarity measure between different models in the same family of distributions is due to a characterization theorem of Chentsov [17]. As described in [3], Chentsov's theorem roughly states that the Fisher information metric equipped with any one of a special set of α-connections, which includes the Levi-Civita connection, are the unique metric/connection combinations that are invariant under sufficient statistic mappings of the model parameters; this includes all injective parameter mappings as well. This is a statistically desirable property as model re-parameterization, for example, should not influence any notion of distance between distinct models. There are several related characterization theorems for the Fisher metric [5, 15, 43], which extend Chentsov's initial work and establish

this result in the setting of continuous distributions. As a result, the Fisher information metric and its associated distance function provide the most natural means of defining a distance between two distributions of the same parametric family.

There are two major difficulties in using the Fisher information distance in applications. First, it is often not possible to obtain simple closed form expressions for the values of $I_{ij}(\theta)$. This results in the geodesic equations taking a complex form which requires a numerical quadrature method to be used inside a differential equation solver to calculate distances. Such methods are often intractable and unstable. Second, solving the geodesic equation (5) numerically can be quite challenging due to singularities that arise in its coefficient functions. These issues are illustrated in the case of the Pareto Power Law distribution [11, p.61]:

$$\phi(x;\alpha) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}} \quad \text{where} \quad x \geq x_m, \quad \alpha > 0. \tag{7}$$

The Fisher information metric and inverse take a simple closed form:

$$I = g = \begin{pmatrix} \alpha/x_m^2 & -1/x_m \\ -1/x_m & 1/\alpha^2, \end{pmatrix}, \quad g^{-1} = \begin{pmatrix} \frac{x_m^2}{\alpha(1-\alpha)} & \frac{\alpha x_m}{1-\alpha} \\ \frac{\alpha x_m}{1-\alpha} & \frac{\alpha^2}{1-\alpha}, \end{pmatrix}. \tag{8}$$

and geodesic equations for this model are given by

$$\ddot{\alpha} + \frac{1}{\alpha-1}\left[\frac{\alpha-2}{2x_m}\dot{\alpha}^2 - \frac{\dot{\alpha}\dot{x}_m}{\alpha} + \frac{x_m}{\alpha^2}\dot{x}_m^2\right] = 0, \quad \ddot{x}_m + \frac{1}{\alpha-1}\left[\frac{\alpha^2}{2x_m^2}\dot{\alpha}^2 - \frac{\alpha\dot{\alpha}\dot{x}_m}{x_m} + \frac{\dot{x}_m^2}{\alpha}\right] = 0, \tag{9}$$

which need to be solved numerically. One such method to solve these equations is the shooting point method. Here given two models (α_1, x_{m_1}) and (α_2, x_{m_2}), the boundary value problem defied by equations (9) with initial conditions $(\alpha(0), x_m(0)) = (\alpha_1, x_{m_1})$ and terminal conditions $(\alpha(1), x_m(1)) = (\alpha_2, x_{m_2})$ can be transformed into an initial value problem where the initial conditions are supplemented with a choice of the derivative $(\dot{a}(0), \dot{x}_m(0)) = v$ for a vector v in the tangent space of the initial point. For a fixed v, one then integrates outward and computes the distance between the geodesic and the target boundary point. The initial choice vector v is refined through a search over the tangent space until this distance vanishes.

We have had some success using the shooting point method to solve these equations in specific instances of initial and terminal points; however, producing a robust solution which works for any boundary conditions can be quite challenging. In particular, note that there are two singularities in this equation when $x_m(t) = 0$ or $\alpha(t) = 1$; these can cause the shooting point method to fail if the geodesic being solved for passes nearby these points. We discuss alternative methods to approximate geodesic distances in two or higher dimensions below, and now focus on the one dimensional setting.

3. Distance Functions for One Dimensional Models

In the case of one dimensional models consisting of a single free parameter, we can determine the Fisher distance exactly up to numerical quadrature. These models may be specified in terms of a single parameter or result from a higher dimensional model that has had all but one parameter fixed. We first consider the case of the exponential distribution before treating the generic one dimensional model.

The probability density of the exponential distribution is defined by

$$\phi(x;\lambda) = \lambda e^{-\lambda x}, \quad \text{where} \quad \lambda > 0, \tag{10}$$

for $x \in [0, \infty)$. The single component of the Fisher metric $g_{\lambda\lambda}$ is given by

$$g_{\lambda\lambda} = -\mathbb{E}\left(\frac{\partial^2 \ln \phi}{\partial \lambda^2}\right) = \frac{1}{\lambda^2}. \tag{11}$$

One can compute the distance function for this model directly by solving the geodesic equation. The inverse metric is given by $g^{\lambda\lambda} = \lambda^2$ and the Christoffel symbol is $\Gamma^\lambda_{\lambda\lambda} = g^{\lambda\lambda}\partial_\lambda g_{\lambda\lambda}/2 = -1/\lambda$. The geodesic equation for the model parameter $\lambda(t)$ is

$$\ddot{\lambda} - \frac{1}{\lambda}(\dot{\lambda})^2 = 0, \tag{12}$$

which has solution $\lambda(t) = c_2 e^{c_1 t}$. If we take $\lambda(0) = \lambda_1$ and $\lambda(1) = \lambda_2$, then solving for c_1, c_2, we find that

$$c_1 = \ln(\lambda_2/\lambda_1), \quad c_2 = \lambda_1, \tag{13}$$

and the distance between these two models is given by

$$d(\lambda_1, \lambda_2) = \left|\int_0^1 \sqrt{g_{\lambda\lambda}\dot{\lambda}^2}dt\right| = \left|c_1 c_2 \int_0^1 \frac{e^{c_1 t}}{\lambda}dt\right| = |c_1| = \left|\ln\left(\frac{\lambda_2}{\lambda_1}\right)\right|. \tag{14}$$

We follow a similar process to develop a generic formula for the Fisher distance on one-dimensional model spaces. Consider a one dimensional model $\phi(x;\theta)$ whose Fisher information matrix has a single component that takes the form $I(\theta) = u(\theta)^2$ so that the information metric is given by

$$g = u(\theta)^2 d\theta^2, \tag{15}$$

where here $u : \mathbb{R} \to \mathbb{R}$ is assumed to be at least C^1. The Christoffel symbol for this metric is given by

$$\Gamma^{\theta}_{\theta\theta} = \frac{1}{2} g^{\theta\theta} \partial_\theta g_{\theta\theta} = \frac{1}{2u^2}(2uu') = \partial_\theta \ln(u(\theta)). \tag{16}$$

For any fixed $z \in \mathbb{R}$ define

$$s(\theta) = \int_z^\theta u(y)dy. \tag{17}$$

Then in the coordinate s, the metric takes the form

$$g_{\theta\theta} = \left(\frac{\partial s}{\partial \theta}\right)^2 g_{ss} = u(\theta)^2 g_{ss} \to g_{ss} = 1, \tag{18}$$

which is the Euclidean metric. Here s is an arc length coordinate, and the distance between two points s_1 and s_2 is given by

$$d(s_1, s_2) = |s_2 - s_1| = \left| \int_{\theta_2}^z u(y)dy - \int_{\theta_1}^z u(y)dy \right| = \left| \int_{\theta_2}^{\theta_1} u(y)dy \right| = d(\theta_1, \theta_2). \tag{19}$$

Thus one can compute the distance function for any one dimensional model by integrating the Fisher metric and evaluating the difference of the result at the model parameters. We now turn to examining additional complexities that are involved in the two or higher dimensional setting.

4. Higher Dimensional Examples

There are only a few models with two or more parameters for which there exist known closed form expressions for their Fisher information distance. The most widely studied such example is the multivariate Gaussian distribution and its exponential family extension [21, 51]. We consider three models below that will be used in subsequent applications, the two dimensional normal distribution, generalized Pareto distribution with fixed location parameter and the three dimensional extreme value distribution. We compute the Fisher information matrix for each of these models, and provide discussions of how they have been utilized in the quantitative finance literature. Finally, we give a description of the techniques that will be used to compute geodesic distances for these models.

4.1. Gaussian Distribution.

We first consider one dimensional Gaussian models previously examined from the information geometry perspective in [18, 34, 51] with density function given by:

$$\phi(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad \mu \in \mathbb{R}, \quad \sigma \in \mathbb{R}^+, \tag{20}$$

defined for $x \in \mathbb{R}^+$. The Fisher information metric components for this model are given by

$$I_{\mu\mu} = \frac{1}{\sigma^2}, \quad I_{\mu\sigma} = 0, \quad I_{\sigma\sigma} = \frac{2}{\sigma^2}. \tag{21}$$

The model has the advantage that the geometry can be identified with a rescaled hyperbolic metric which has a closed form solution for its distance function. Specifically, in the Poincare upper half plane model of hyperbolic geometry, the geodesics are semicircles and vertical lines. The distance function is given by [18],

$$d((\mu_1,\sigma_1),(\mu_2,\sigma_2)) = \sqrt{2}\ln\left(\frac{\mathcal{F} + (\mu_1-\mu_2)^2 + 2\sigma_1^2\sigma_2^2}{4\sigma_1\sigma_2}\right), \tag{22}$$

$$\mathcal{F} = \sqrt{((\mu_1-\mu_2)^2 + 2(\sigma_1-\sigma_2)^2)\,((\mu_1-\mu_2)^2 + 2\sigma_1^2\sigma_2^2)}. \tag{23}$$

This model provides a benchmark for testing techniques to compute geodesic distances which can be compared against the exact solution. In addition, we will compare differences in the Fisher distance and the widely used Kullback-Leibler divergence extending the comparison given in [18] and also for purposes of building intuition about the Fisher metric.

4.2. **Generalized Pareto Distribution.** The generalized Pareto distribution has been used to model power law phenomena in economics [23] as well as in the development of tail-risk projection trading strategies in [45]. This distribution has also been used to model loss distributions of stock returns in [29] which was one of the first applications of extreme value theory to the field. This model distribution is defined by:

$$\phi(x;\mu,\sigma,\xi) = \frac{1}{\sigma}\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1-\xi^{-1}}, \quad x>\mu, \quad \sigma,\xi>0. \tag{24}$$

We compute the components of the Fisher information matrix to be

$$I_{\mu\mu} = \frac{(\xi+1)^2}{\sigma^2(2\xi+1)}, \quad I_{\sigma\sigma} = \frac{1}{\sigma^2(2\xi+1)}, \quad I_{\xi\xi} = \frac{2}{2\xi^2+3\xi+1}, \tag{25}$$

$$I_{\mu\sigma} = -\frac{\xi}{\sigma^2(2\xi+1)}, \quad I_{\mu\xi} = -\frac{\xi}{\sigma(2\xi^2+3\xi+1)}, \quad I_{\sigma\xi} = \frac{1}{\sigma(2\xi^2+3\xi+1)}. \tag{26}$$

The significance of the generalized Pareto distribution stems from the Pickands-Balkema-de Haan theorem [6, 47], which roughly states that the distribution of a random variable beyond a suitably high threshold is well approximated by a generalized Pareto distribution, c.f. [42] for further details.

4.3. **Generalized Extreme Value Distribution.** The second prominent distribution that arises in extreme value theory is the generalized extreme value (GEV) distribution. The functional form of this model is given by

$$\phi(x;\mu,\sigma,\xi) = \frac{1}{\sigma}f(x)^{\xi+1}e^{-f(x)}, \tag{27}$$

$$f(x) = \left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}, \quad \xi\neq 0, \quad f(x) = \exp\left(-\frac{x-\mu}{\sigma}\right), \quad \xi = 0, \tag{28}$$

with domain $x\in[\mu-\sigma/\xi,\infty)$ for $\xi>0$, $x\in\mathbb{R}$ for $\xi=0$, and $x\in(-\infty,\mu-\sigma/\xi]$ for $\xi<0$.

The GEV family of distributions was designed to be a generalization of the Gumbel, Weibull, and Fréchet limiting distributions of the Fisher-Tippet Theorem [22] for the block maxima of a sample from a random variable. We will apply it to the loss distribution of financial return series. Specifically, given a sequence of returns r_i for $i=1,\ldots,n$, consider the negative returns $r_i<0$, and then negate them to form a sample from the loss distribution denoted by $\tilde{r}_i$. Let $M_n=\max(\tilde{r}_1,\ldots,\tilde{r}_n)$. The application that we consider is based on a result of [22], which states that there exist location μ_n and scale sequences σ_n, such that in the limit $n\to\infty$, we have $x_n=(M_n-\mu_n)/\sigma_n\to\phi(x;\mu,\sigma,\xi)$.

The Fisher information matrix for the GEV distribution was derived in [48]:

$$I_{\mu\mu} = \frac{p}{\sigma^2}, \quad I_{\sigma\sigma} = \frac{1}{\sigma^2\xi^2}(1-2\Gamma(2+\xi)+p), \tag{29}$$

$$I_{\xi\xi} = \frac{1}{\xi^2}\left(\frac{\pi^2}{6}+\left(1-\gamma+\frac{1}{\xi}\right)^2-\frac{2q}{\xi}+\frac{p}{\xi^2}\right), \quad I_{\mu\sigma} = -\frac{1}{\sigma^2\xi}\left(p-\Gamma(2+\xi)\right), \tag{30}$$

$$I_{\mu\xi} = -\frac{1}{\sigma\xi}\left(q-\frac{p}{\xi}\right), \quad I_{\sigma\xi} = -\frac{1}{\sigma\xi^2}\left(1-\gamma+\frac{1-\Gamma(2+\xi)}{\xi}-q+\frac{p}{\xi}\right), \tag{31}$$

$$p = (1+\xi)^2\Gamma(1+2\xi), \quad q = \Gamma(2+\xi)\left(\psi(1+\xi)+\xi^{-1}+1\right). \tag{32}$$

Before turning to applications that use these three distributions, we comment on the methods that we use to determine the geodesic distance between two distributions in the same family.

4.4. **Computing Geodesic Distances.** Computing geodesic distances in two and higher dimensional models in a robust manner is challenging due to complex forms that the geodesics equations may take. Recent advances in [38, 39] have overcame some of these issues by focusing on solving for the distance function first from which geodesic paths are extracted. Specifically, the author develops numerical schemes to solve Eikonal equations for the distance function of a Riemannian manifold numerically on uniform grids. He focuses on a type of Hamiltonian formulation of the length functional minimization definition of geodesics. Then the fast marching algorithm, which is a generalization of Dijkstra's minimal path method, is used to solve these equations. These techniques work particularly well for complicated geometries which exhibit a high degree of anisotropy. We utilize them below to compute distances between GEV and generalized Pareto models.

5. Applications

We now proceed to consider three applications of the Fisher distance. First, we compare the Kullback Leibler divergence to the Fisher distance in the case of the normal distribution model to develop intuition for differences in these similarity measures. Second, we consider a nearest neighbor clustering application where we initially fit generalized Pareto distributions to equity return time series of members of the NASDAQ 100 index and consider their associated loss distributions. We then consider a single stock, AAPL, and identify which securities are nearest AAPL in a Fisher distance sense; thus finding the stocks whose risk profiles most closely resemble that of AAPL. Finally, we extend this idea to a hierarchical clustering application. Instead of focusing on a single stock, we iteratively cluster groups of stocks with a bottom-up hierarchical clustering algorithm that utilizes the Fisher distance as a similarity measure between their pairwise loss distributions along with Ward's linkage criteria for measuring the distance between clusters of stocks. This results in a risk focused hierarchical grouping of stocks.

5.1. **Comparison between KL divergence and Fisher Distance.** We first examine differences between the Kullback Leibler (KL) divergence, which is widely used as a similarity measure between probability distributions, and the Fisher information distance in the case of a univariate normal distribution through an empirical example. A related comparison was developed in [18] which we extend by considering an empirical example focused on equity returns from NASDAQ 100 stocks. To construct the dataset for this example, we download daily historical index components and end of day closing price data for the NASDAQ 100 index from Bloomberg between 2010-01-01 and 2015-12-31 and select stocks that remained in the index over this entire timeframe. Data was downloaded with Bloomberg's Python API package blpapi. We forward fill missing data at the price level so that a stock whose price was filled for a single day will have a corresponding zero daily return. We note that the price data is split and dividend adjusted by Bloomberg and that, with a few exceptions, each time series is fully populated over the timeframe being considered.

Our aim is to study differences between the Fisher distance and widely used KL divergence similarity measure. It is known that these two distance measures agree to second order when measuring the disparity between infinitesimally close distributions [28, 33]; however, significant differences arise on larger distance scales. Part of this comparison will examine how these distances differ when applied to the maximum likelihood parameters of each stock in the NASDAQ100 dataset. In particular, we fit the distribution to each of the return series using the unbiased MLE estimators for a normal distribution.

The general KL divergence for two probability distributions f, g is defined by

$$D_{KL}(f||g) = \int_{\mathbb{R}} f(x) \log \frac{f(x)}{g(x)} dx, \tag{33}$$

which for a univariate Gaussian distribution can be calculated explicitly

$$D_{KL}((\mu_1, \sigma_1)||(\mu_2, \sigma_2)) = \ln \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}. \tag{34}$$

This is not a distance function as it is neither symmetric nor satisfies the triangle inequality. However, it has been widely used as a similarity measure between distributions, in particular, in information theory applications [56]. We first provide a graphical comparison in Figure 1 for pairwise distances between all models in this NASDAQ 100 dataset as well as a visualization of the distance functions to a standard normal distribution of both similarity measures. Here, we compute all pairwise distances between the maximum likelihood model parameters of the 107 stocks which remained in the index over the timeframe of consideration. In the left most plot of Figure 1, we display histograms of pairwise distances for both similarity measures. The right tail of the KL divergence histograms was truncated at an upper limit of 3.0; however, we note that the KL divergence has several outlier distances. In particular, 13 values greater than 10.0 with

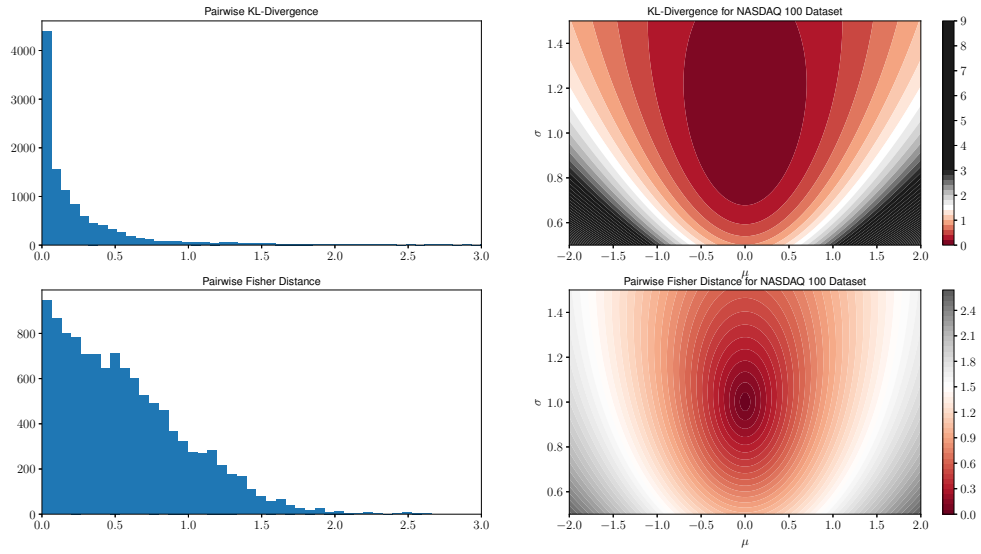

FIGURE 1. Comparison between pairwise KL-Divergence and Fisher information metric values for NASDAQ 100 parameters and distance functions to a $\mathcal{N}(0,1)$ distribution. Note that the KL divergence concentrates a number of distance values near zero and also has several large outliers whereas the Fisher distance distribution decreases roughly linearly for increasing distance.

a maximum value of 19.2 were observed. The pairwise Fisher distance histogram contains all data points, and by comparison, the maximum value is 2.7. Next, note that the KL divergence histogram has roughly an exponential decay whereas the Fisher distance histogram declines linearly. The KL divergence tends to concentrate pairwise distances near zero values when compared with the Fisher distance; this provides an indication that the Fisher distance may distinguish subtle differences in the return distributions between the equity series more strongly than the KL divergence. In addition, the greater dispersion of Fisher distances, in relation to those corresponding to the KL divergence, motivates its utilization in clustering algorithms. Specifically, one would expect a more informative cluster structure using the Fisher distance as opposed to having many data points tightly grouped together in the case of the KL divergence.

In the right two plots of Figure 1, we display contour plots that measure the similarity between all normal models in the domain and a standard $\mathcal{N}(0,1)$ distribution, i.e. we plot level sets of the function $f(\mu,\sigma) = d((0,1),(\mu,\sigma))$ for each (μ,σ) in the domain of consideration. Here we can see that the KL divergence initially concentrates around zero for models close to the standard Gaussian, but rapidly expands for further away models which have increasingly large values in comparison to the Fisher distance. In addition, the Fisher distance has slower growth and is more granular for models near the standard Gaussian.

5.2. **Generalized Pareto Nearest Neighbor Example.** We now consider an application of identifying the nearest neighbors of a given stock based on the Fisher distance between the loss distribution of the stock and the securities to which it is being compared. We use the generalized Pareto distribution as a model for the loss distributions of these stocks.

There are many ways to fit a generalized Pareto distribution to data, c.f. [32] for a comparison of a subset of such methods. Maximum likelihood estimation is notoriously difficult as the likelihood function is undefined for portions of the parameter domain. Multiple alternative fitting methods for the generalized Pareto distribution have been developed [8, 9, 31, 35, 41] which are generally more stable and robust when compared with typical estimation techniques such as the maximum likelihood estimation or the method of moments. Although any of these fitting methods may be used in this application, since our focus is model comparison and not estimation, we use a hybrid estimation procedure developed in [56] based on minimizing the Anderson-Darling statistic of this model which is both straightforward to implement and yields robust and intuitive results which we now briefly describe.

Given a sample x_i from an equity loss distribution, we note that the maximum likelihood estimator for the location parameter is $\hat{\mu} = \min(\{x_1, \ldots, x_n\})$. We first compute $\hat{\mu}$ for the loss distribution of each stock, which is typically zero, and subtract means from the samples of every security. As this parameter does not vary significantly over models, we will fix it for the remainder of this application and fit and compare models with the two free parameters (σ, ξ). As in [56], define $\theta = \xi/\sigma$, and fit this model by defining a set of functions

$$g_i(\theta) = 1 - (1 - \theta x_{(i)})^{-n/\sum_{j=1}^{n} \ln(1-\theta x_j)}, \tag{35}$$

for $i = 1, \ldots, n$ where here $x_{(i)}$ denotes the i-th order statistic of the sample. Then we minimize

$$G(\theta) = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\ln(g_i(\theta)) + (2n+1-2i)\ln(1-g_i(\theta)), \tag{36}$$

using a BFGS optimizer [58] setting the initializer at 10% of the maximum value of the sample. Denote the minimum value by $\hat{\theta}$, which defines the two model parameter estimators

$$\hat{\xi} = -\frac{1}{n}\sum_{i=1}^{n}\ln(1-\hat{\theta}x_i), \quad \hat{\sigma} = \frac{\hat{\xi}}{\hat{\theta}}. \tag{37}$$

Overall, we have found this optimization problem to be more stable than threshold based fitting methods and it also leads to intuitive fits. Specifically, in Figure 2, we display the best fit model parameters and distributions alongside return histograms for the loss distribution of six randomly selected stocks from our NASDAQ 100 sample.

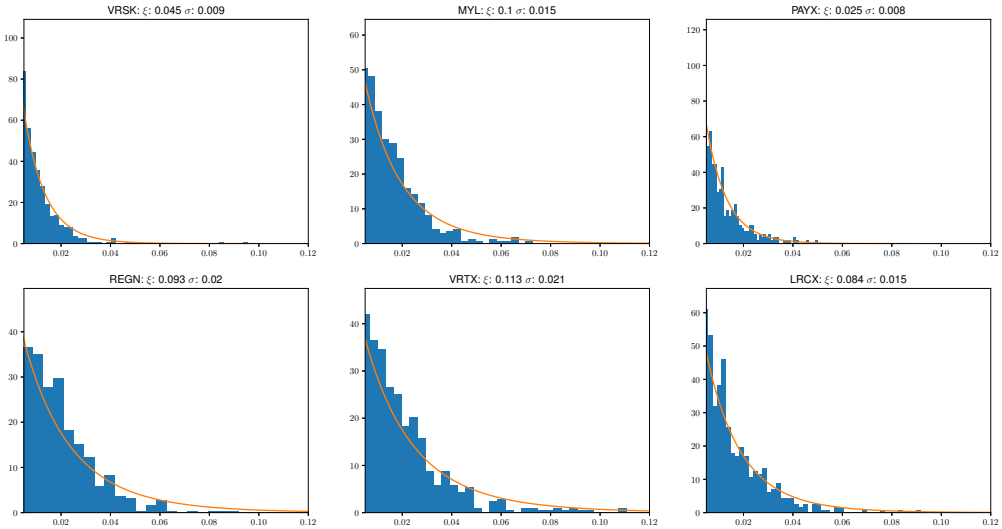

FIGURE 2. Generalized Pareto distributions fit using the hybrid method of [56] and empirical loss distribution histograms.

We note that this fitting technique works well for both low volatility stocks such as PAYX as well as those which consistently sustain larger losses like VRTX. Our next aim is to identify which stocks are the shortest distance away from a fixed stock. We fix APPL for this example, and calculate the geodesics that join the generalized Pareto models for each stock in the dataset to APPL and display the geodesics in Figure 3 for fifty stocks in the sample.

First note that the Fisher geodesic paths which connect model parameters deviate significantly from Euclidean geodesic lines. For models with ξ and σ values larger than those of AAPL, the geodesics have a strong arc which eventually reverts back and intersects the target model. If ξ is larger but σ is similar,

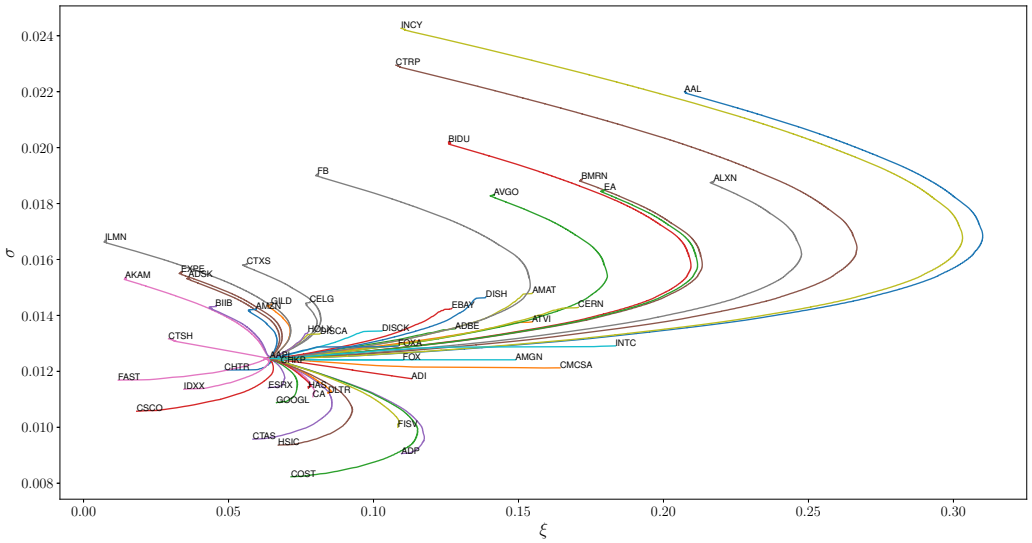

FIGURE 3. Geodesic paths between generalized Pareto distribution parameters of AAPL and fifty randomly selected NASDAQ 100 securities.

Rank	Stock	Distance	Rank	Stock	Distance	Rank	Stock	Distance	Rank	Stock	Distance
1	CHKP	0.012	11	KLAC	0.055	86	WYNN	0.397	96	MELI	0.533
2	XLNX	0.012	12	FOX	0.055	87	PAYX	0.425	97	KHC	0.540
3	MCHP	0.020	13	FOXA	0.056	88	AVGO	0.437	98	LILAK	0.544
4	QVCA	0.027	14	VIAB	0.060	89	FB	0.440	99	VRTX	0.579
5	ISRG	0.033	15	SBUX	0.061	90	NFLX	0.467	100	TSLA	0.612
6	LBTYA	0.037	16	TXN	0.063	91	EA	0.472	101	SWKS	0.612
7	CTSH	0.043	17	WBA	0.064	92	BMRN	0.487	102	AAL	0.672
8	ADI	0.048	18	JBHT	0.073	93	REGN	0.508	103	CTRP	0.678
9	MAT	0.048	19	ULTA	0.073	94	ALXN	0.514	104	LILA	0.727
10	CHTR	0.054	20	PCLN	0.077	95	BIDU	0.533	105	INCY	0.752

TABLE 1. Ranked Distances from AAPL to the top and bottom twenty stocks.

arclengths are much smaller and geodesics are approximately linear whereas lower values of σ exhibit the same bullet style geodesics of large values. Also, different share classes such as FOX and FOXA have similar models and geodesics as expected since the return series of different share classes are highly correlated. In addition, note that this technique links several comparable technology stocks such as GOOGL, AMZN, are more similar to AAPL than FB.

Given a geodesic path consisting of a sequence of parameter values (ξ_i, σ_i), for $i = 0, \ldots, n$, we approximate its arclength between models by first defining $\Delta\xi_i = \xi_i - \xi_{i-1}$, $\bar{\xi} = (\xi_i - \xi_{i-1})/2$ and $\Delta\sigma_i$, $\bar{\sigma}_i$ analogously. We compute explicit distances using a discrete analogue of equation (4)

$$d((\xi_0, \sigma_0), (\xi_n, \sigma_n)) = \sum_{i=1}^{n} \sqrt{g_{\xi\xi}(\bar{\xi}_i, \bar{\sigma}_i)\Delta\xi_i^2 + 2g_{\xi\sigma}(\bar{\xi}_i, \bar{\sigma}_i)\Delta\xi_i\Delta\sigma_i + g_{\xi\xi}(\bar{\xi}_i, \bar{\sigma}_i)\Delta\sigma_i^2}, \tag{38}$$

and gather results in Table 1 for the twenty nearest and furthest away stocks from AAPL. One example to note is that ADI and CHTR have similar Fisher distances; however, one can see from Figure 3 that CHTR is considerably closer to AAPL in the Euclidean distance than ADI. Also, the geodesic to ADI is nearly linear whereas that to CHTR is curved. This example demonstrates the potentially significant differences between the Fisher and Euclidean distance between model parameters.

Finally, in Figure 4, we plot the fitted generalized Pareto distributions to each of the NASDAQ 100 models being considered. In black, we plot the best fit distribution to AAPL and in red the ten nearest distributions

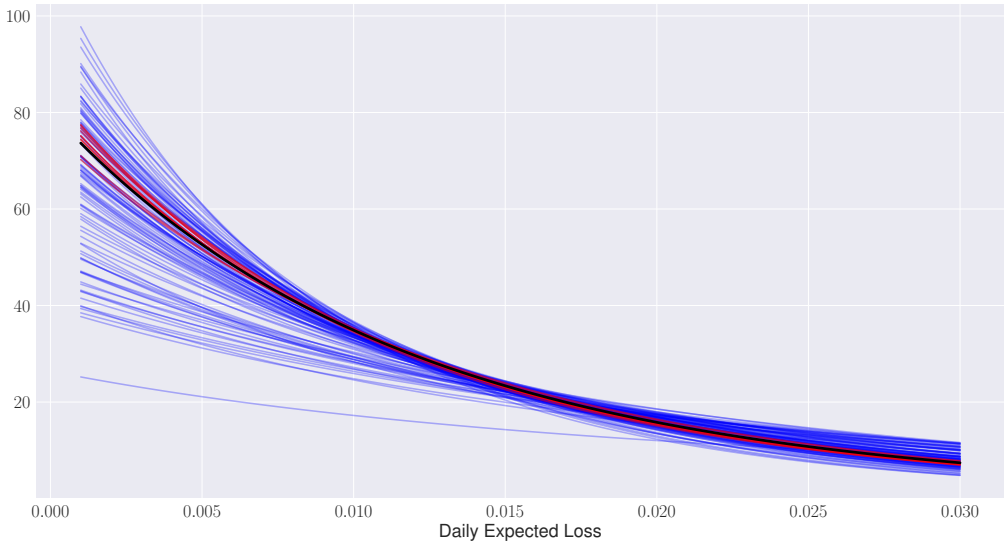

FIGURE 4. Generalized Pareto distribution fits for APPL (black), its ten nearest neighbors (red), and the remaining NASDAQ 100 stocks (blue).

to AAPL with respect to the Fisher distance. The remainder of model distributions are displayed in blue. This graph demonstrates that the Fisher distance groups stocks according to similarities in their distribution functional forms which was the original design intention of identifying stocks with common tail risk behavior.

5.3. **Application to Clustering Based on Worst Annual Loss for S&P 500 Stocks.** Next, we expand upon the nearest neighbor example by considering a clustering application using the Fisher distance based on an example of estimating the worst daily loss over an annual period provided in [29]. This application stems from the Fisher-Tippett-Gnedenko theorem [22, 27] which roughly states that for a set of independent identically distributed samples x_i from a probability distribution, the maximum $M_m = \max(\{x_1, \ldots, x_n\})$ converges to a generalized extreme value distribution modulo a shift and scale sequence. Our application uses the GEV distribution as a model for the distribution of the maximum return of the loss distribution (worse single daily loss) for a series of stock returns.

We expand the number of securities in this application by extracting the components of the S&P 500 as of 12-31-2016 and keep stocks that remained in the index over the prior thirty years since 1-1-1987. For each stock, we find its minimum daily return over the prior calendar year and fit a three dimensional GEV distribution with shift and scale parameters using maximum likelihood estimation to each set of thirty maximum loss values. The result is a set of 272 GEV model parameters which estimate the distribution of the maximum yearly loss for each stock.

Our aim will be to apply a hierarchical clustering method to these models in order to group stocks by their worst annual losses. Specifically, we use an agglomerative clustering method which requires two inputs. First, one must specify a notion of distance between two objects being clustered. In our example, these objects correspond to the max loss distributions of distinct stocks and the distance between them is given by the Fisher distance. Second, one must specify a notion of distance between clusters consisting of multiple points called the linkage function. We use Ward's linkage which is designed to minimize the intra-cluster variance in each fusion step of the method [44, 57].

This clustering technique consists of initially grouping together the two securities with the smallest Fisher distance and then iteratively clustering the closest two security/cluster pairs until all distributions have been grouped into a single cluster. One way of visualizing this clustering algorithm is through the dendrogram of the clustering procedure, portions of which we display in Figure 5.

Several dendrogram subclusters contain stocks belonging to the same sector. For example, in the upper left plot, the four stocks WMT, HRL, BMS, and ABT are part of the large right subcluster which is itself a subcluster of the entire group. In addition, we find that some clusters are composed of companies with similar businesses such as DE, SWK, and LEG which concentrate in the design and manufacturing of engineered

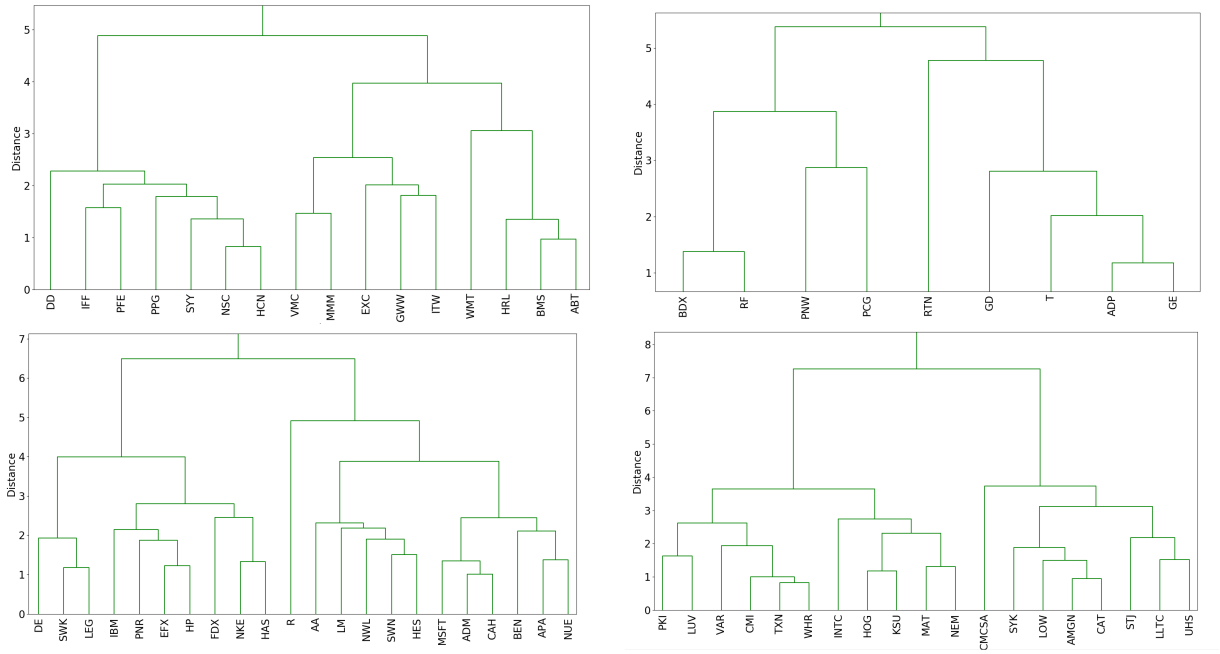

FIGURE 5. Portions of the dendrogram of a Fisher distance based hierarchical clustering of the best fit GEV distributions for the block maximum of S&P 500 stocks.

products and heavy machinery. Other clusters such as GE, ADP, T, and GD in the upper right subplot contain securities from a mix of the aerospace, technology, industrial, and telecommunications industries. This clustering procedure provides a new way to group similar stocks according to their worst annual return.

6. Conclusion

In summary, we have reviewed relevant ideas in information geometry and discussed issues that arise when computing geodesics for statistical models. We then derived a closed form expression up to quadrature for the Fisher distance for one dimensional models. Next, we compared the Fisher distance and KL divergence in an equity return distribution application for the univariate normal model. Finally, we developed two extreme value theory examples; a nearest neighbor comparison using the generalized Pareto distribution, and a maximum daily loss over and annual period distribution hierarchical clustering technique using the generalized extreme value distribution.

There are several further directions that may be pursued. First, it would be interesting to find and develop applications that involve the Fisher distance outside of quantitative finance, specifically in the areas of natural language and image processing. Second, the Fisher metric often cannot be represented in terms of a closed form algebraic expression. In such situations, one can develop numerical methods to compute geodesic distances given that the Fisher metric is only defined up to quadrature. Finally, one can consider Fisher distance analogous of clustering, classification, and regression techniques that take a distance function as input.

References

[1] Adler, R., Feldman, R., Taqqu, M., 1998. A Practical Guide to Heavy Tails: Statistical Techniques and Applications. *Birkhäuser*.
[2] Amari, S., 2016. Information Geometry and Its Applications. *Springer Applied Mathematical Sciences*.
[3] Amari, S., Nagaoka, H., 2000. Methods of Information Geometry. *Translations of Mathematical Monographs (American Mathematical Society)* **191**.
[4] Ay, N., Jost, J., Le H.V., Schwachhofer, L., 2017. Information Geometry. *Springer*.
[5] Ay, N., Jost, J., Le H.V., Schwachhofer, L., 2013. Information Geometry and Sufficient Statistics. *Probability Theory and Related Fields* **162** (1-2), 327-364.
[6] Balkema, A., and de Haan, L., 1974. Residual life at great age. *Annals of Probability*, **2** 792-804.
[7] Banerjee, A., et. al., 2005. Clustering with Bregman Divergences. *Journal of Machine Learning* **6**, 1705-49.
[8] Bermudez, Z., Kotz, S., 2010. Parameter estimation of the generalized Pareto distribution Part 1. *Journal of Statistical Planning and Inference*. **140** (6), 1353-1373.

[9] Bermudez, Z., Kotz, S., 2010. Parameter estimation of the generalized Pareto distribution Part 2. *Journal of Statistical Planning and Inference*. **140** (6), 1374-1388.

[10] Bonanno, G., et. al., 2004. Networks of equities in financial markets. *The European Physical Journal B*. **38** (2), 363-371.

[11] Bouchaud, J.P., Potters, M., 2009. Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management. *Cambridge University Press*.

[12] Brewin, L., 1996. Riemann Normal Coordinates. Lecture Notes.

[13] Brewin, L., 2009. Riemann normal coordinate expansions using Cadabra. *Class. Quantum Grav.* **26** (17), 175017.

[14] Cai, F., Le-Khac, N.A., and Kechadi, M.T., 2012. Clustering approaches for financial data analysis: a survey. *Proceedings of the 8th International Conference on Data Mining, Las Vegas, Nevada, USA*, 105-111.

[15] Campbell, L.L., 1986. An extended Cencov characterization of the information metric. *Proceedings of the American Mathematical Society*, **98**, 135-141.

[16] Cha, S.-H., 2007. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences*. **1** (4), 300-307.

[17] Chentsov, N.N., 2010. Statistical Decision Rules and Optimal Inference. *Translations of Mathematical Monographs (American Mathematical Society)* **53**.

[18] Costa S., Santos S., and Strapasson J., 2015. Fisher information distance: A geometrical reading. *Discrete Applied Mathematics* **197**, 59-69.

[19] Cramér, H., 1946. Mathematical Methods of Statistics. *Princeton University Press*.

[20] Dose, C., Cincotti, S., 2005. Clustering of financial time series with application to index and enhanced index tracking portfolio. *Physica A: Statistical Mechanics and its Applications*. **355** (1) 145-151.

[21] Efron, B., 1978. The Geometry of Exponential Families. *The Annals of Statistics*. **6**(2) 362-376.

[22] Fisher, R.A., Tippet H.C., 1928. Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. *Proceedings of the Cambridge Philosophical Society* **24**, 180-90.

[23] Gabaix, X., 2009. Power Laws in Economics and Finance. *Annu. Rev. Econ.* **1**, 255-293.

[24] Gattone, S.A., De Sanctis, A., Russo, T., Pulcini D., 2017. A shape distance based on the Fisher-Rao metric and its application for shapes clustering. *Physica A* **487**, 93-102.

[25] Gattone, S.A., De Sanctis, A., Puechmorel S., Nicol F., 2018. On the geodesic distance in shapes K-means clustering. *Entropy* **20**(9), 647.

[26] Gencay, R., Selcuk, F., 2004. Extreme value theory and Value-at-Risk: Relative performance in emerging markets. *International Journal of Forecasting* **20** (2) 287-303.

[27] Gnedenko, B.V., 1943. Sur la distribution limite du terme d'une série aléatoire. *Annals of Mathematics* **44**, 423-453.

[28] Guo, D., 2009. Relative Entropy and Score Function: New Information-Estimation Relationships through Arbitrary Additive Perturbation. *Proc. IEEE International Symposium on Information Theory*. 814-818.

[29] Gilli, M., Kellezi, E., 2006. An Application of Extreme Value Theory for Measuring Financial Risk. *Computational Economics* **27** (2-3), 207-228.

[30] Jiang, B., Pei, J., Tao, Y., Lin, X., 2013. Clustering Uncertain Data Based on Probability Distribution Similarity. *IEEE Transactions on Knowledge and Data Engineering* **25** (4), 751-763.

[31] Jockovic, J., 2012. Quantile Estimation for the Generalized Pareto Distribution with Application to Finance. *Yugoslav Journal of Operations Research* **22** (2), 297-311.

[32] Kang, S., Song, J., 2017. Parameter and quantile estimation for the generalized Pareto distribution in peaks over threshold framework. *Journal of the Korean Statistical Society* **46** (4), 487-501.

[33] Kullback, S., 1968. Information Theory and Statistics. *Dover*.

[34] Lebanon, G., 2005. Riemannian Geometry and Statistical Machine Learning. *CMU Dissertation: Lambert Academic Publishing*.

[35] Lenz, F., 2015. Generalized Pareto Distributions, Image Statistics and Autofocusing in Automated Microscopy. *Geometric Science of Information*, 96-103.

[36] Longin, F., 1996. The Asymptotic Distribution of Extreme Stock Market Returns. *The Journal of Business*. **69** (3) 383-408.

[37] Malevergne, Y., Pisarenko, V., Sornette D., 2006. On the power of generalized extreme value (GEV) and generalized Pareto (GPD) estimators for empirical distributions of stock returns. *Applied Financial Economics* **16** (3) 271-289.

[38] Mirebeau, J.M., 2014. Anisotropic Fast-Marching on cartesian grids Lattice Basis Reduction. *SIAM J. Numer. Anal.* **52** (4) 1573-1599.

[39] Mirebeau, J.M., Portegies, J., 2018. Hamiltonian Fast Marching: A numerical solver for anisotropic and non-holonomic eikonal PDEs. <hal-01778322>.

[40] McNeil A., Frey, R., Duffie D., Schaefer S., 2015. Quantitative Risk Management: Concepts, Techniques and Tools - Revised Edition (Princeton Series in Finance). *Princeton University Press*.

[41] McNeil A., Frey, R., 2000. Estimation of tail risk related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*. **7** (3-4) 271-300.

[42] McNeil A., 1997. Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory. *Astin Bulletin* **27** (1), 117-137.

[43] Montufar G., Rauh J., Ay, N., 2014. On the Fisher Metric of Conditional Probability Polytopes. *Entropy* **16**(6), 3207-3233.

[44] Murtagh, F., 2014. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification* **31**, 274-295.

[45] Packham, N., Papenbrock, J., Schwendner, P., Woebbeking, F., 2016. Tail-risk protection trading strategies. *Quantitative Finance* **17** (5), 729-744.

[46] Petersen, P., 2006. Riemannian Geometry. *Springer: Graduate Texts in Mathematics* **171**.

[47] Pickands, J., 1975. Statistical inference using extreme value order statistics. *Annals of Statistics* **3**, 119-131.

[48] Prescott, P., Walden, A. T., 1980, Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika* **67** (3), 723-724.
[49] Rao, C.R., 1945. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society.* **37** 81-89.
[50] De Sanctis, A., Gattone, S.A., 2016. Methods of Information Geometry to model complex shapes. *European Physics Journal-Special Topics.* **225**, 1271-1279
[51] Skovgaard, L.T., 1984. A Riemannian Geometry of the Multivariate Normal Model. *Scandinavian Journal of Statistics* **11** (4) 211-223.
[52] Strapasson, J., Porto, J., Costa, S., 2015. On bounds for the Fisher-Rao distance between multivariate normal distributions. *AIP Conference Proceedings* **1641**, 313.
[53] Tabak, B.M., Serra, T.R., Cajueiro, D.O., 2010, Topological properties of commodities networks. *Eur. Phys. J. B* **74**, 243-249.
[54] Tumminello, M., Lillo, F., Mantegna, R., 2010. Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior and Organization* **75** (1), 40-58.
[55] Van, T.V., Pham-Gia, T., 2010. Clustering Probability Distributions. *Journal of Applied Statistics* **37** (11), 1891-1910.
[56] Wang, C.P., Jo, B., 2013. Applications of a Kullback-Leibler Divergence for Comparing Non-nested Models. *Stat Modelling* **13** (5-6) 409-429.
[57] Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **58** (301), 236-244.
[58] Zhu, C., Byrd, R.H., Lu, P., Nocedal, J., 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transaction on Mathematical Software (TOMS)* **23** (4), 550-560.