Proceedings of Machine Learning Research vol 107:207–220, 2020

Calibrating Multivariate Lévy Processes with Neural Networks

Kailai Xu KAILAIX@STANFORD.EDU
Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, 94305

Eric Darve DARVE@STANFORD.EDU
Mechanical Engineering, Stanford University, Stanford, CA, 94305

Abstract

Calibrating a Lévy process usually requires characterizing its jump distribution. Traditionally this problem can be solved with nonparametric estimation using the empirical characteristic functions (ECF), assuming certain regularity, and results to date are mostly in 1D. For multivariate Lévy processes and less smooth Lévy densities, the problem becomes challenging as ECFs decay slowly and have large uncertainty because of limited observations. We solve this problem by approximating the Lévy density with a parametrized functional form; the characteristic function is then estimated using numerical integration. In our benchmarks, we used deep neural networks and found that they are robust and can capture sharp transitions in the Lévy density compared to piecewise linear functions and radial basis functions.

Keywords: Lévy Processes; Neural Networks; Automatic Differentiation

1. Introduction

Lévy processes generalize Gaussian processes by allowing jump-diffusion. Because of their ability to allow continuous evolution and abrupt jumps of random variables (Chen et al. (2010)), many models in finance, physics and biology have been built based on Lévy processes. For example, in the classical Black-Scholes model for risky assets, the price S_t of an asset at time t is governed by (Figueroa-Lopez and Houdré (2004))

$$S_t = S_0 e^{\sigma B_t + \mu t} \tag{1}$$

where B_t is the standard Brownian motion amd σ and μ are the standard deviation and the drift mean. To account for the excessive skewness and kurtosis in the log return distributions in empirical financial data, the Black-Scholes model has been generalized to the exponential Lévy process (Tankov (2011); Andersen and Lipton (2013); Benhamou (2000))

$$S_t = S_0 e^{X_t} \tag{2}$$

where X_t is a Lévy process. Yet, because of the lack of analytical closed-form density functions for general Lévy processes, an exact maximum likelihood estimator is not feasible. The lack of a closed-form density function leads to the difficulty of calibrating Lévy processes in the presence of jumps.

The multivariate Lévy process can be described by three parameters (Menn and Rachev (2006)): a positive semi-definite matrix $\mathbf{A} = \mathbf{\Sigma}\mathbf{\Sigma}^T \in \mathbb{R}^{d\times d}$, where $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$; a vector $\mathbf{b} \in \mathbb{R}^d$, and a

measure $\nu \in \mathbb{R}^d \backslash \{\mathbf{0}\}$. The Lévy process $\mathbf{X}_t$ is a superposition of a Wiener process $\mathbf{\Sigma B}_t + \mathbf{b}t$, where $\mathbf{B}_t$ is the standard i.i.d. Brownian motion, and a pure-jump Lévy process with the Lévy measure

$$\nu(A) = \frac{1}{t}\mathbb{E}\left(\sum_{s \le t} \mathbf{1}_A(\mathbf{X}_s - \mathbf{X}_{s-})\right) \tag{3}$$

where $\mathbf{1}_A$ is an indicator for A, i.e., $\mathbf{1}_A(\mathbf{x}) = 1$ for $\mathbf{x} \in A$ and 0 otherwise. The corresponding characteristic function is given by the Lévy-Khintchine representation (Papapantoleon (2008))

$$\phi(\boldsymbol{\xi}) = \mathbb{E}[e^{\mathrm{i}\langle \boldsymbol{\xi}, \mathbf{X}_t\rangle}] = \exp\left[t\left(\mathrm{i}\langle \mathbf{b}, \boldsymbol{\xi}\rangle - \frac{1}{2}\langle \boldsymbol{\xi}, \mathbf{A}\boldsymbol{\xi}\rangle + \int_{\mathbb{R}^d}\left(e^{\mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x}\rangle} - 1 - \mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x}\rangle \mathbf{1}_{\|\mathbf{x}\| \le 1}\right)\nu(d\mathbf{x})\right)\right] \tag{4}$$

The subject of this paper is to study the nonparametric calibration of pure jump processes $\nu(\mathbf{x})$ and thus we assume $\mathbf{A} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$ throughout the paper. In addition, we assume that ν is determined by a density function such that $\nu(d\mathbf{s}) = \nu(\mathbf{s})d\mathbf{s}$, and we call $\nu(\mathbf{s})$ the *Lévy density*.

The traditional nonparametric estimation for Lévy processes in 1D has two cases (Neumann et al. (2009)): (1) the Lévy process X_t is observed at high frequency at times t_i, i.e., $\max_i(t_i - t_{i-1})$ is small. In this case, a large increment $X_{t_i} - X_{t_{i-1}}$ indicates that a jump occurred. For example Comte and Genon-Catalot (2009) proposed nonparametric inference methods for Lévy process in this case. (2) in the low-frequency observation case, there are zero or several jumps present within the increment $X_{t_i} - X_{t_{i-1}}$. In this case, Neumann et al. (2009) applied a deconvolution algorithm to estimate $\nu(\mathbf{x})$ from the empirical characteristic function. Cont and Tankov (2004) discretized the Lévy density $\nu(\mathbf{x})$ on a grid and applied relative entropy minimization to find the optimal $\nu(\mathbf{x})$. We tackle the latter case, where we want to estimate $\nu(\mathbf{x})$ from Lévy process data given at equispaced time intervals, i.e., $t_1 = \Delta t$, $t_2 = 2\Delta t$, $t_3 = 3\Delta t$, ...

However, much of the attention in the literature has been restricted to 1D case and well-behaved $\nu(\mathbf{x})$. In the case where $\nu(\mathbf{x})$ is discontinuous, Neumann et al. (2009) proposed a deconvolution method but since the decay of the characteristic function is very slow, the method requires large computational domains. The method proposed in Cont and Tankov (2004) assigned one degree of freedom (DOF) to the discretized Lévy density $\nu(\mathbf{x})$ per grid point, which partially contributed to the ill-posedness of the nonlinear optimization problem. The ill-posedness problem becomes more severe in higher dimensions since DOFs grow exponentially.

In this paper, we tackle those challenges by proposing a novel approach for 2D nonparametric estimation of the Lévy density $\nu(\mathbf{x})$. This approach proceeds in four stages:

1. The Lévy density is approximated by a parametric functional form—such as piecewise linear functions—with parameters $\boldsymbol{\theta}$,
$$\nu(\mathbf{x}) \approx \nu_{\boldsymbol{\theta}}(\mathbf{x}) \tag{5}$$

2. The characteristic function is approximated by numerical integration
$$\phi(\boldsymbol{\xi}) \approx \phi_{\boldsymbol{\theta}}(\boldsymbol{\xi}) := \exp\left[\Delta t \sum_{i=1}^{n_q}\left(e^{\mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x}_i\rangle} - 1 - \mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x}_i\rangle \mathbf{1}_{\|\mathbf{x}_i\| \le 1}\right)\nu_{\boldsymbol{\theta}}(\mathbf{x}_i)w_i\right] \tag{6}$$
where $\{(\mathbf{x}_i, w_i)\}_{i=1}^{n_q}$ are quadrature nodes and weights and $\mathbf{1}$ denotes the indicator function.

3. The empirical characteristic functions are computed given observations $\{\mathbf{X}_{i\Delta t}\}_{i=0}^{n}$

$$\hat{\phi}_n(\boldsymbol{\xi}) := \frac{1}{n}\sum_{i=1}^{n} \exp(\mathrm{i}\langle \boldsymbol{\xi}, \mathbf{X}_{i\Delta t} - \mathbf{X}_{(i-1)\Delta t}\rangle),\ \boldsymbol{\xi} \in \mathbb{R}^d \tag{7}$$

4. Finally, we solve the following optimization problem with a gradient based method.

$$\min_{\boldsymbol{\theta}} \frac{1}{m}\sum_{i=1}^{m} \|\hat{\phi}_n(\boldsymbol{\xi}_i) - \phi_{\boldsymbol{\theta}}(\boldsymbol{\xi}_i)\|^2 \tag{8}$$

   Here $\{\boldsymbol{\xi}_i\}_{i=1}^{m}$ are collocation points and $\|\cdot\|$ denotes complex modulus.

Note that our method is not a direct fitting algorithm in the sense that we do not fit the function approximators with input-output pairs. The output of the function approximators are unobservable and must be calibrated by embedding the function approximator in a statistical model. The gradient descent method requires differentiating through both the statistical model and the function approximator.

One challenge for this approach is computing the error $\|\hat{\phi}_n(\boldsymbol{\xi}) - \phi(\boldsymbol{\xi})\|$ in the empirical characteristic function. In theory, the empirical characteristic function converges to the exact one given infinite observations under certain conditions (Marcus (1981); Feuerverger et al. (1977)). However, in practice the observations are limited and thus the empirical characteristic function is not exact. Another challenge is the discontinuity of Lévy densities. This occurs when the jump distribution experiences sudden changes for certain $\mathbf{x}$ (Eberlein (2009)).

To tackle these challenges, the choice of function approximators $\nu_{\boldsymbol{\theta}}(\mathbf{x})$ is essential. From the previous discussion, a potential form must have the following properties: (1) universal approximation, i.e., the capability of approximating any continuous functions given sufficient computing budget; (2) robustness to noise; (3) ability to handle discontinuity. In this paper, we apply and benchmark three popular parametric functional forms: neural networks (NN), piecewise linear functions (PL), and radial basis functions (RBF).

Neural networks enjoy many favorable properties such as exponential convergence for analytic functions (E and Wang (2018)) and rational function (Telgarsky (2017)) and we demonstrate empirically that they outperform the other two methods—PL and RBF—in several aspects. On the one hand, PL consists of local basis functions and therefore DOFs with no data points nearby around are not optimized. On the other hand, although the basis functions in RBF are global so it suffers less from the problem PL struggles with, it is well known that RBF is susceptible to noise and discontinuity (Boyd (2010); Xu and Belytschko (2005); Orr et al. (1996)). Besides, accuracy and stability of RBF depend on proper choice of shape parameters and determining optimal shape parameters is difficult in many situations. However, the problems are alleviated for NN, partially because it is adaptive to non-uniform data (Huang et al. (2019)), robust to noise (Rolnick et al. (2017)), and can overcome Gibbs phenomenon (Llanas et al. (2008); Xu and Darve (2019)). The capability of the neural network is demonstrated in Figure 1, where different function approximators are trained on 20 data points in a step function

$$f(x) = 0.5 + 1_{2<x<5}(x)$$

As we can see in the plots, when the number of training points is fixed, we need to limit the degrees of freedom (DOF) or impose regularization for local basis functions, such as piecewise linear functions, to avoid overfitting. In the presence of discontinuity, the radial basis functions suffer from oscillations near the discontinuous point. The neural network is capable of capturing the discontinuity and provides a large number of DOFs to approximate complex functions.

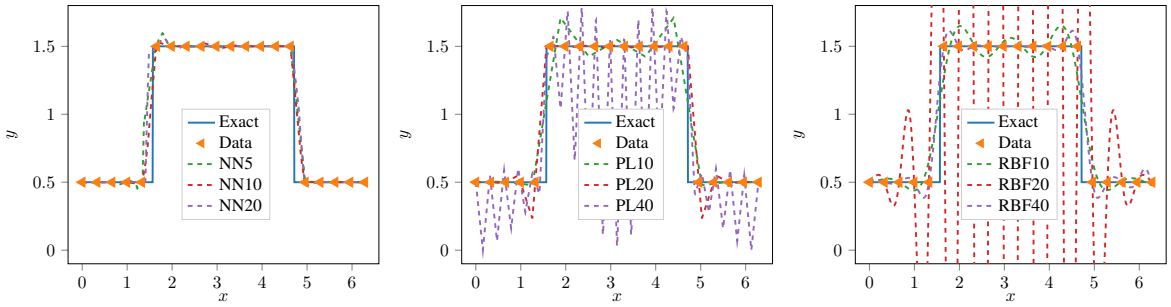

Figure 1: Fitting a step function $y = 0.5 + 1_{2<x<5}(x)$ using different function approximators. NNn stands for neural network model with n layers, 20 neurons per hidden layer and ReLU as activation function. PLn stands for piecewise linear function with n equispaced distributed nodes. RBFn stands for radial basis functions with n equispaced distributed centers.

With the re-parametrization technique, we show that the method can also be applied to multivariate symmetric α-stable processes (Gulian et al. (2018)), a subclass of Lévy processes. In this case, $\nu(\mathbf{x})$ is singular at $\mathbf{x} = \mathbf{0}$, but we can re-parametrize the characteristic function as

$$\phi(\boldsymbol{\xi}) = \mathbb{E}\left(\exp(\mathrm{i}t\langle\boldsymbol{\theta},\boldsymbol{\xi}\rangle)\right) = \exp\left[t\left(-\frac{1}{2}\langle\boldsymbol{\xi},\mathbf{A}\boldsymbol{\xi}\rangle + \mathrm{i}\langle\boldsymbol{\xi},\mathbf{b}\rangle - \int_{\mathbb{S}^d}|\langle\boldsymbol{\theta},\boldsymbol{\xi}\rangle|^{\alpha}\Gamma(\mathbf{s})d\mathbf{s}\right)\right] \tag{9}$$

where $\Gamma(\mathbf{s})$ is a function defined on $\mathbb{S}^d$. Here we can substitute $\Gamma(\mathbf{s})$ by a parametrized functional form $\Gamma_{\boldsymbol{\theta}}(\mathbf{s})$, apply the quadrature rule on the unit circle and minimize the discrepancy between $\hat{\phi}_n(\boldsymbol{\xi})$ and $\phi_{\boldsymbol{\theta}}(\boldsymbol{\xi})$.

Finally, we built a toolset `LevyNN` for calibrating Lévy processes based on the open source library `ADCME.jl`. The latter is an automatic differentiation library with `TensorFlow` backends and is specially designed for scientific computing. The library automates the gradient computation and integrates the optimization workflow.

2. Nonparametric Estimation of the Lévy processes

2.1. Characteristic Function Matching Method

The characteristic function matching method (Yu (2004)) minimizes the discrepancy between the empirical characteristic function $\hat{\phi}_n(\boldsymbol{\xi})$ (Equation (7)) and the characteristic function $\phi(\boldsymbol{\xi})$ (Equation (4)). The rationales are: (1) As $n \to \infty$, $\hat{\phi}_n(\boldsymbol{\xi}) \to \phi(\boldsymbol{\xi})$ because of the large number law; (2) there is a one-to-one correspondence between the characteristic function $\phi(\mathbf{x})$ and the density function for $\mathbf{X}_{i\Delta t} - \mathbf{X}_{(i-1)\Delta t}$, a.k.a., $\nu(\mathbf{x})$. Consequently, we can estimate $\nu(\mathbf{x})$ from $\hat{\phi}_n(\boldsymbol{\xi})$.

For Lévy processes, we have to make two additional approximations: (1) $\nu(\mathbf{x})$ is approximated with a parametrized functional form $\nu_{\boldsymbol{\theta}}(\mathbf{x})$ (Equation (5)); (2) Based on $\nu_{\boldsymbol{\theta}}(\mathbf{x})$, $\phi(\boldsymbol{\xi})$ is approximated by $\phi_{\boldsymbol{\theta}}(\boldsymbol{\xi})$ via numerical integration (Equation (6)). In the following, we will discuss details regarding those two approximations.

2.2. Approximation to the Lévy Density

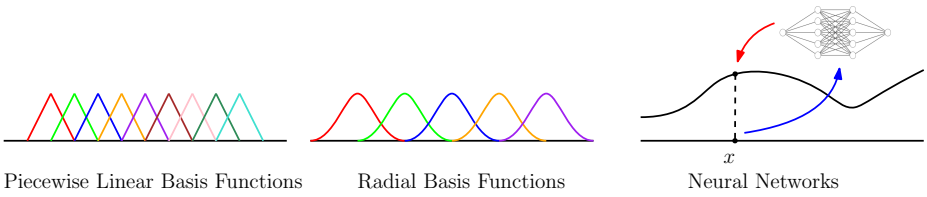

Figure 2: Different functional forms in 1D. In PL and RBF, the target function are approximated by linear combination of basis functions; in NN, it is approximated by composing linear transformations and nonlinear activation functions.

The Lévy density $\nu(\mathbf{x})$ is a mapping from the coordinates $\mathbf{x} \in \mathbb{R}^2$ to $\mathbb{R}$. We first truncate the infinite computational domain to $\mathbf{x} \in [-M, M]^2$ and then approximate $\nu(\mathbf{x})$ with $\nu_{\boldsymbol{\theta}}(\mathbf{x})$. In the following paragraphs, we discuss three functional forms for $\nu_{\boldsymbol{\theta}}(\mathbf{x})$ (Figure 2).

One type of neural network (NN) is a composition of linear operations followed by a nonlinear activation function. In this paper, we consider ReLU dense neural networks, where

$$\nu_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{W}_L \mathrm{ReLU}(\mathbf{W}_{L-1} \mathrm{ReLU}(\cdots \mathrm{ReLU}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) \cdots) + \mathbf{b}_{L-1}) + \mathbf{b}_L \tag{10}$$

here $\mathrm{ReLU}(x) = \max(x, 0)$ and it is applied elementwise, L is the number of layers and $\boldsymbol{\theta} = \{(\mathbf{W}_i, \mathbf{b}_i)\}_{i=1}^{L}$ are the weights and biases. For all the hidden layers, the output dimension is 20. NN is special because information at each data point is not represented by linear combination of predetermined basis functions by composing linear and nonlinear mappings.

Although the neural network does not require the input range or computational domain explicitly, it usually requires that the input data is properly scaled. For example, if M is very large, we need to scale $\mathbf{x}$ to $\frac{\mathbf{x}}{M}$ first.

For piecewise linear functions (PL), the computational domain is first triangulated and each vertex is associated with one DOF. The value $\nu_{\boldsymbol{\theta}}(\mathbf{x})$ is linearly interpolated from the nodal values of the triangle where $\mathbf{x}$ is located. $\boldsymbol{\theta}$ consists of all those DOFs. In this paper, we obtain the triangulation by splitting each square cell into two triangles on a uniform grid. One disadvantage of PL is the local DOF problem, where the DOFs with no data points nearby are not trained.

For radial basis functions (RBF), we have

$$\nu_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{i=1}^{M} a_i \frac{1}{\sqrt{(\mathbf{x} - \mathbf{x}_i)^2 + c^2}} \tag{11}$$

where $\{a_i\}_{i=1}^{M}$ are coefficients, $\{\mathbf{x}_i\}_{i=1}^{M}$ are centers, c is the shape parameter. In this paper, the centers are chosen as the grid points on a uniform grid. c is given by the grid step size, suggested by

Wu et al. (2012). Although the basis functions are global, the coefficients in RBF are more affected by data points that are closer to the corresponding centers. Hence, we expect RBF also suffers from the local DOF problem like PL.

2.3. Numerical Approximation to the Characteristic Function

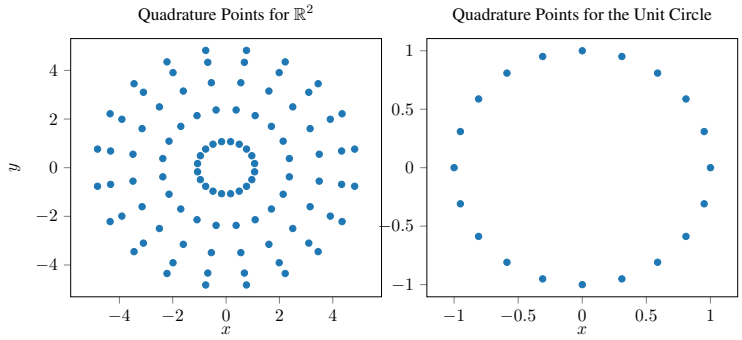

Figure 3: Quadrature points used for $\mathbb{R}^2$ and the unit circle. To avoid bloated plots, we show fewer quadrature points with order $n_q = 100$ and $n_q = 20$ respectively.

We assume that $\nu(\mathbf{x})$ decays as $\|\mathbf{x}\| \to \infty$; we use the quadrature rule on the truncated domain $\{(x, y) : x^2 + y^2 < M^2\}$ for approximating the integral in Equation (4)

$$\int_{\mathbb{R}^2} \left(e^{\mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x} \rangle} - 1 - \mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x} \rangle \mathbf{1}_{\|\mathbf{x}\| \le 1} \right) \nu(\mathbf{x}) d\mathbf{x} \approx \sum_{i=1}^{n_q} \left(e^{\mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x}_i \rangle} - 1 - \mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x}_i \rangle \mathbf{1}_{\|\mathbf{x}_i\| \le 1} \right) \nu_{\boldsymbol{\theta}}(\mathbf{x}_i) w_i$$

The quadrature points and weights $\{(\mathbf{x}_i, w_i)\}_{i=1}^{n_q}$ are obtained according to Cools and Kim (2000). For multivariate stable processes in the following, we use quadrature rules on the unit circle. Figure 3 shows examples of quadrature points with order $n_q = 100$ and $n_q = 20$.

Consequently, we obtain the expression for the approximation to the characteristic function

$$\phi(\boldsymbol{\xi}) \approx \phi_{\boldsymbol{\theta}}(\boldsymbol{\xi}) := \exp\left[\Delta t \sum_{i=1}^{n} \left(e^{\mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x}_i \rangle} - 1 - \mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x}_i \rangle \mathbf{1}_{\|\mathbf{x}_i\| \le 1} \right) \nu_{\boldsymbol{\theta}}(\mathbf{x}_i) w_i \right] \tag{12}$$

2.4. Optimization

The characteristic function matching method requires minimizing the discrepancy between $\hat{\phi}_n(\boldsymbol{\xi})$ and $\phi_{\boldsymbol{\theta}}(\boldsymbol{\xi})$. For computation, we consider a set of collocation points $\{\boldsymbol{\xi}_i\}_{i=1}^{m}$ uniformly drawn from $[-M', M']^2$ and solve the nonlinear least square problem

$$\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) := \frac{1}{m} \sum_{i=1}^{m} \|\hat{\phi}_n(\boldsymbol{\xi}_i) - \phi_{\boldsymbol{\theta}}(\boldsymbol{\xi}_i)\|^2 \tag{13}$$

The choice of M' is based on $\{\mathbf{X}_i\}$: we can choose M' such that $|\hat{\phi}_n(\boldsymbol{\xi})|$ is smaller than a certain value for $\boldsymbol{\xi} \in \mathbb{R}^2 \backslash [-M', M']^2$. In the objective function, we have used the characteristic function

instead of its logarithm (the characteristic exponent) because by taking the exponential of the characteristic exponent we can eliminate the 2π period in its imaginary part. Additionally, because of the exponential, the objective function will place more weights on the high frequency region (when the real part of the characteristic exponent is large), and this objective function is desired in certain scenarios because high frequency captures sharp changes in the Lévy density function.

The optimization problem Equation (13) is solved with `ADCME`. It computes the gradient $\nabla L(\boldsymbol{\theta})$ using automatic differentiation (Baydin et al. (2018); Abadi et al. (2016)) and applies a gradient-based optimizer such as `L-BFGS-B` (Dai (2013)) for minimization. The considerable flexibility makes it easy to test different approximation functional forms $\nu_{\boldsymbol{\theta}}(\mathbf{x})$ without deriving and implementing new gradients or optimization procedures.

2.5. Multivariate α-Stable Process: Re-parametrization

For the multivariate symmetric α-stable distribution, the characteristic function of the increment $X_{i\Delta t} - X_{(i-1)\Delta t}$ is given by the following theorem Samorodnitsky et al. (1994)

Theorem 1 *$\mathbf{X}$ is a symmetric α-stable vector in $\mathbb{R}^d$ with $0 < \alpha < 2$ if and only if there exists a unique symmetric finite measure Γ on the unit sphere $\mathbb{S}^d$ such that*

$$\phi(\boldsymbol{\xi}) = \mathbb{E}\left(\exp(\mathrm{i}\Delta t\langle \mathbf{X}, \boldsymbol{\xi}\rangle)\right) = \exp\left(-\Delta t \int_{\mathbb{S}^d} |\langle \mathbf{s}, \boldsymbol{\xi}\rangle|^\alpha \Gamma(d\mathbf{s})\right) \tag{14}$$

Γ is the spectral measure of the symmetric α-stable vector $\mathbf{X}$.

We assume that Γ is determined by a density function such that $\Gamma(d\mathbf{s}) = \Gamma(\mathbf{s})d\mathbf{s}$. The previous procedure will fail because $\nu(\mathbf{x})$ is singular at $\mathbf{x} = \mathbf{0}$ thus the given quadrature rule is unable to handle the singularities. For example, when $\Gamma(\mathbf{s}) = 1$, the corresponding Lévy density satisfies (Nolan (2008))

$$\nu(\mathbf{x}) \propto \frac{1}{\|\mathbf{x}\|^{\alpha+2}}, \quad \|\mathbf{x}\| \to 0 \tag{15}$$

Instead of working with $\nu(\mathbf{x})$, we approximate Equation (14) directly since $\nu(\mathbf{x})$ is singular in this case but $\Gamma(\mathbf{s})$ usually has better regularity. For calibrating the multivariate symmetric α-stable process, we apply the quadrature rule $\{(\mathbf{s}_i, w_i)\}_{i=1}^{n_q}$ on a unit circle instead of $\mathbb{R}^2$ and we obtain

$$\phi(\boldsymbol{\xi}) \approx \phi_{\boldsymbol{\theta}}(\boldsymbol{\xi}) := \exp(\mathrm{i}\Delta t\langle \boldsymbol{\theta}, \boldsymbol{\xi}\rangle) = \exp\left(-\Delta t \sum_{i=1}^{n_q} |\langle \boldsymbol{\xi}, \mathbf{s}_i\rangle|^\alpha \Gamma_{\boldsymbol{\theta}}(\mathbf{s}_i) w_i\right) \tag{16}$$

We have the additional constraint $\Gamma(\mathbf{s}) = \Gamma(-\mathbf{s})$ according to Theorem 1. This is enforced directly by the functional form $\Gamma_{\boldsymbol{\theta}}$. For example, we assume $\Gamma_{\boldsymbol{\theta}}(\mathbf{s}) = \Gamma'_{\boldsymbol{\theta}}(\mathbf{s}) + \Gamma'_{\boldsymbol{\theta}}(-\mathbf{s})$, where $\Gamma'_{\boldsymbol{\theta}}(\mathbf{s})$ is NN, PL or RBF in the 1D domain $[0, 2\pi)$ (since there exists a one-to-one correspondence between $\mathbb{S}^2$ and $[0, 2\pi)$).

3. Numerical Results

We now present the results of numerical experiments. We first compare the accuracy of three functional forms based on exact characteristic function, ignoring the uncertainty from observations.

Function	NN5	NN10	NN20	PL10	PL20	PL40	RBF10	RBF20	RBF40
Step	0.7500	0.7499	0.7498	0.7493	0.7494	0.7500	0.7482	0.7483	0.7504
Constant	0.7499	0.7500	0.7500	0.7500	0.7500	0.7499	0.7500	0.7500	0.7499

Table 1: Estimated α for different methods and test functions. The exact fractional index α is 0.75. We can see that our NN method is able to learn α quite accurately, regardless of the choices of basis functions.

Then we apply and compare the functional forms to symmetric α-stable processes and general Lévy processes in the presence of uncertainty from observations. We show that NN has very favorable properties in terms of being robust and capturing sharp transitions. We demonstrate the properties in the following three benchmarks and apply the algorithm to stock markets data.

3.1. Multivariate α-Stable Processes: Estimation from Exact Empirical Characteristic Functions

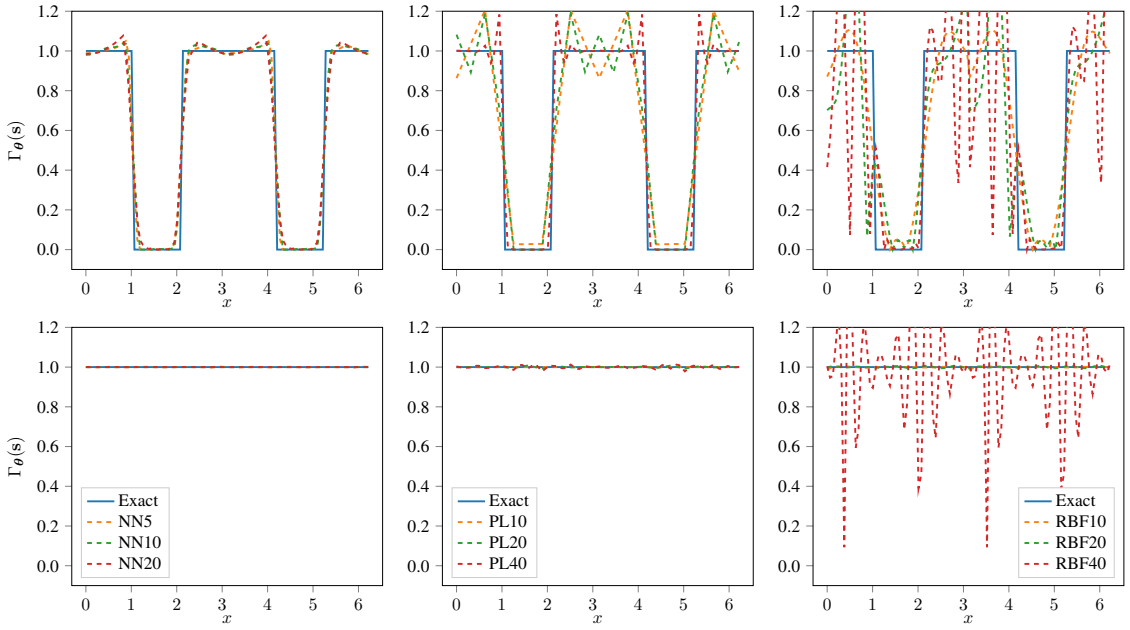

Figure 4: Estimated $\Gamma_{\boldsymbol{\theta}}(\mathbf{s})$ from exact characteristic functions using different methods. Here $\mathbf{s} = [\cos(x), \sin(x)]$, $x \in [0, 2\pi)$. For details about legend abbreviations, see Figure 4.

We assume that $\phi(\boldsymbol{\xi})$ is computed with accurate numerical quadrature rules $n_q = 10000$ for

$$\Gamma(\mathbf{s}) = \mathbf{1}_{|s_1|>0.5}(\mathbf{s}), \text{ and } \Gamma(\mathbf{s}) = 1, \ \mathbf{s} = (s_1, s_2), \ \mathbf{s} \in \mathbb{S}^2 \tag{17}$$

hence the error is negligible for estimating $\phi(\boldsymbol{\xi})$. We assume $\Delta t = 0.5$, $\alpha = 0.75$, and $n_q = 100$ for approximating $\phi_{\boldsymbol{\theta}}(\boldsymbol{\xi})$. The results in Figure 4 indicate that NN can capture the sharp transition better than others. For PL, if DOFs are too few, it is unable to capture the transition; however, some of

DOFs are not trained because there are no data points locating in the support of the associated basis functions. For RBF, results for RBF40 implies that too few data points compared to the number of centers make the optimization problem ill-posed. Additionally, RBF fails to capture the sharp transitions.

The fractional indices are estimated quite accurately (Table 1). This implies that compared to the "directional" information of the jump, the heavy tail information is easier to capture.

The effect of overfitting Despite having much more degrees of freedom than the PL and the RBF, NN does not overfit when trained from sample paths data and statistical constraint Equation (13). Piecewise linear functions are local basis functions, and thus to calibrate all the coefficients in the linear combination, the number of points in the training data must be at least the number of coefficients. That's why we see a lot of oscillations in the middle panel of Figure 1 when the DOF is large. For radial basis functions, it is known that when the number of the basis functions is larger than the number of training data, numerical instability occurs. As a numerical demonstration, we compare the results in Table 2 for $\alpha = 0.75$ using NN5, PL2000, and RBF2000, which have approximately the same number of parameters (2000 for the PL and RBF, and 1762 for the NN). We report the error norm $\|\Gamma(\hat{\mathbf{s}}) - \Gamma_{\boldsymbol{\theta}}(\hat{\mathbf{s}})\|_2$ using equispaced $\hat{\mathbf{s}}$.

Function	NN5	PL2000	RBF2000
Error	3.51	24.97	14.97

Table 2: Comparison of error $\|\Gamma(\hat{\mathbf{s}}) - \Gamma_{\boldsymbol{\theta}}(\hat{\mathbf{s}})\|_2$ for different methods with approximately the same DOF.

3.2. Multivariate α-Stable Processes: Estimation from Observations

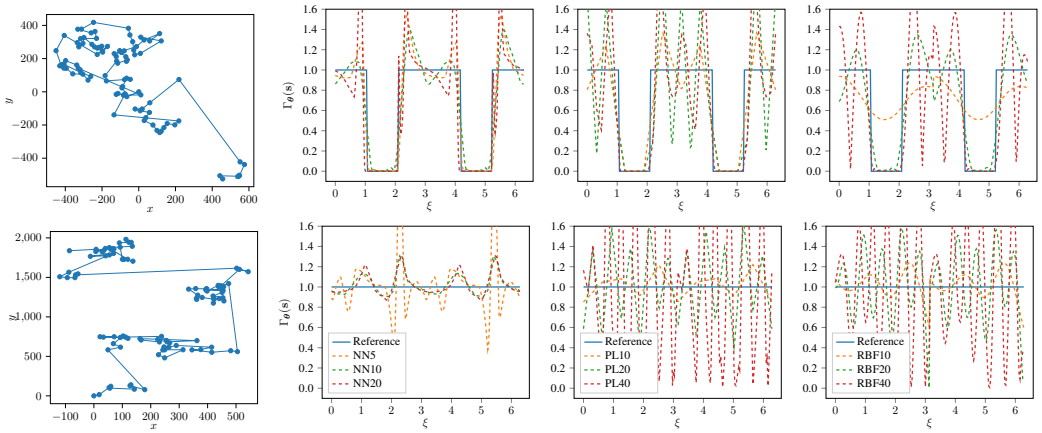

Figure 5: The first column shows sample paths of generated alpha-stable processes. The other three columns show estimated $\Gamma_{\boldsymbol{\theta}}(\mathbf{s})$ from observations with different methods. The $\xi \in [0, 2\pi)$ axis corresponds to the angle of $\mathbf{s}$, i.e., $\mathbf{s} = [\cos(\xi), \sin(\xi)]$. For details about legend abbreviations, see Figure 4.

Function	NN5	NN10	NN20	PL10	PL20	PL40	RBF10	RBF20	RBF40
Step	1.5164	1.5156	1.5162	1.5151	1.5166	1.5169	**3.2155**	1.5154	1.5171
Constant	1.5331	1.5329	1.5329	1.5330	1.5329	1.5330	1.5329	1.5329	1.5330

Table 3: Estimated α for different methods and test functions. The reference fractional index α is 1.5. For step functions, if we use too few centers for radial basis functions, the estimation is not accurate (RBF10). This is also demonstrated in Figure 5.

Now we consider estimating the multivariate α-stable process from $m = 1000$ observations. Different from the last section, $\phi(\boldsymbol{\xi})$ is unknown and is estimated with $\hat{\phi}_n(\boldsymbol{\xi})$. The difference $|\hat{\phi}_n(\boldsymbol{\xi}) - \phi(\boldsymbol{\xi})|$ introduces additional uncertainty, which can also be interpreted as "noise" in the nonlinear optimization problem.

The results in Figure 5 imply that NN is most robust in either case. The α indices are properly estimated in Table 3, except for the step function and RBF10 case because of the noise.

3.3. Multivariate Lévy Processes.

In this example, we consider the Lévy process where the jump distributions are truncated normal distributions. The Lévy density has the expression

$$\nu(\mathbf{x}) = \frac{2}{\pi} \exp\left(-\frac{\|\mathbf{x}\|^2}{2}\right) \mathbf{1}_{\mathbf{x}\in\mathbb{R}^2_+} \tag{18}$$

The density $\nu(\mathbf{x})$ is only nonzero for $\mathbf{x} \in \mathbb{R}^2_+$ and has sharp transition at axes $x = 0$ and $y = 0$ in the first quadrant. In our experiment, we assume $\Delta t = 0.5$, $m = 10000$, $n_q = 4096$. The data $\{\mathbf{X}_{i\Delta t}\}_{i=1}^n$ are simulated according to Nolan (2008). Notably, NN captures the sharp transition (Figure 6). However, the results from PL do not even capture the shape of the density function well. Additionally, RBF creates a smooth profile of the density distribution with a blurring edge and thus fails to capture the sharp transition.

3.4. Application to Stock Markets.

Finally, we apply our method to a stock market example. We investigate 12 stocks from 01/01/2016 to 08/01/2019, which are from the technology sector (MSFT, AAPL, AMZN, GOOG), the financial sector (JPM, C, WFC, CME) and the energy sector (EOG, XOM, COG, MPC). The α index is computed for each pair of stocks. The stock prices are turned into the log return and then shifted so data for each stock are unbiased.

We model the pairwise shifted log return of the stocks by a 2D symmetric α-stable process with unknown α and $\Gamma(\mathbf{s})$. Figure 7 shows the estimated pairwise α indices. Most of the indices are between 1.1 and 1.5. This implies that jumps do exist in the pairwise log return changes. We also show $\Gamma(\mathbf{s})$ for EOG vs. MSFT. We identify 4 peaks in the plot, which indicate that there is a larger tendency for prices to jump in those 4 directions compared to nearby directions.

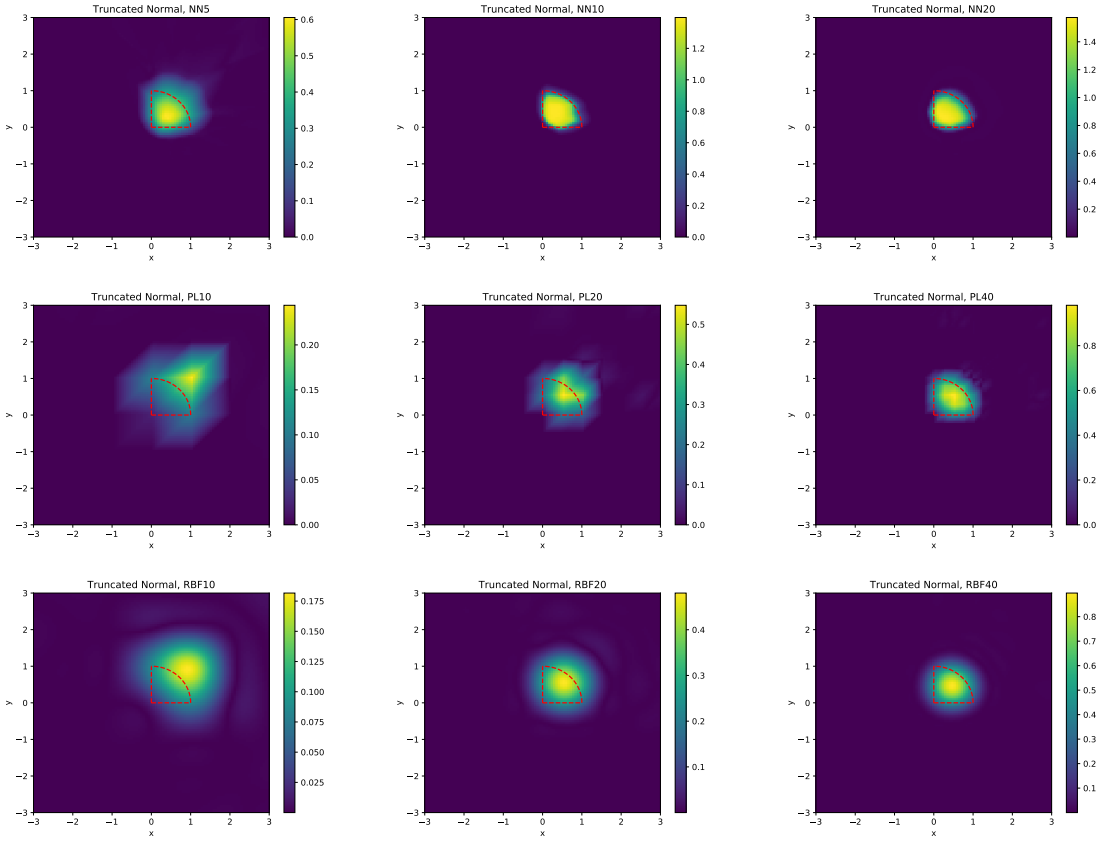

Figure 6: Estimated $\nu_{\boldsymbol{\theta}}(\mathbf{x})$ from observations with different methods. The intercepted arc of the dashed red sector is $\{(x, y) : x^2 + y^2 = 1, x \geq 0, y \geq 0\}$. The sector indicates the area with most of the density for the reference $\nu(\mathbf{x})$. For details about subtitle abbreviations, see Figure 4.

4. Conclusion

We have proposed a novel nonparametric estimation approach for Lévy processes and compared three approximation functional forms: (1) neural network (2) piecewise linear functions (3) and radial basis functions. We found that for the tested cases neural network performed best for being robust to noise and capturing sharp transitions. However, one should not expect that neural networks are always superior to others. Most likely, a certain functional form may be more suitable to a class of problems, since the performance highly depends on the characteristics of the training data.

Besides Lévy processes, the same idea—approximating an unknown function in a system model with neural network, and training by matching the model outputs with observations—can be applied to many other fields as well. In the future, a deeper understanding of the neural network approximation properties and improvement of the training algorithm will broaden the applications of our nonparametric estimation approach.

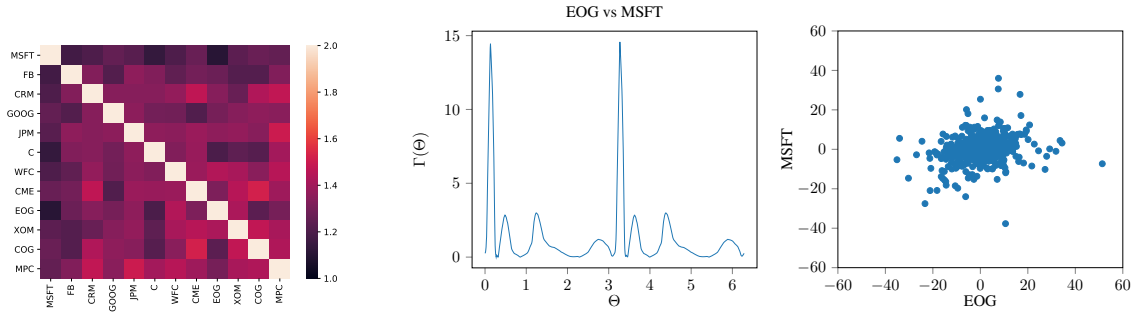

Figure 7: The first plot shows the pairwise α indices. For each stock, the α index against itself is not computed. The right two plots show the calibrated $\Gamma_{\boldsymbol{\theta}}(\mathbf{s})$ and shifted log return for EOG and MSFT. Here $\mathbf{s} = (\cos(\Theta), \sin(\Theta))$, here $\Theta \in (0, 2\pi]$.

Acknowledgments

This work is supported by the Applied Mathematics Program within the Department of Energy (DOE) Office of Advanced Scientific Computing Research (ASCR), through the Collaboratory on Mathematics and Physics-Informed Learning Machines for Multiscale and Multiphysics Problems Research Center (DE-SC0019453).

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Leif Andersen and Alexander Lipton. Asymptotics for exponential lévy processes and their volatility smile: survey and new results. *International Journal of Theoretical and Applied Finance*, 16 (01):1350001, 2013.

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic Differentiation in Machine Learning: a Survey. *Journal of Marchine Learning Research*, 18:1–43, 2018.

Eric Benhamou. Option pricing with lévy process. 2000.

John P Boyd. Six strategies for defeating the runge phenomenon in gaussian radial basis functions on a finite interval. *Computers & Mathematics with Applications*, 60(12):3108–3122, 2010.

Song X Chen, Aurore Delaigle, and Peter Hall. Nonparametric Estimation for a Class of Lévy Processes. *Journal of Econometrics*, 157(2):257–271, 2010.

Fabienne Comte and Valentine Genon-Catalot. Nonparametric Estimation for Pure Jump Lévy Processes Based on High Frequency Data. *Stochastic Processes and their Applications*, 119(12): 4088–4123, 2009.

Rama Cont and Peter Tankov. Nonparametric calibration of jump-diffusion option pricing models. *Journal of computational finance*, 7:1–49, 2004.

Ronald Cools and Kyung Joong Kim. A Survey of Known and New Cubature Formulas for the Unit Disk. *Korean Journal of Computational & Applied Mathematics*, 7(3):477–485, 2000.

Yu-Hong Dai. A Perfect Example For the BFGS Method. *Mathematical Programming*, 138(1-2): 501–530, 2013.

Weinan E and Qingcan Wang. Exponential convergence of the deep neural network approximation for analytic functions. *arXiv preprint arXiv:1807.00297*, 2018.

Ernst Eberlein. Jump–type lévy processes. In *Handbook of financial time series*, pages 439–455. Springer, 2009.

Andrey Feuerverger, Roman A Mureika, et al. The empirical characteristic function and its applications. *The annals of Statistics*, 5(1):88–97, 1977.

Enrique Figueroa-Lopez and Christian Houdré. Nonparametric Estimation for Lévy Processes with a view Towards Mathematical Finance. *arXiv preprint math/0412351*, 2004.

Mamikon Gulian, Maziar Raissi, Paris Perdikaris, and George Karniadakis. Machine Learning of Space-fractional Differential Equations. *arXiv preprint arXiv:1808.00931*, 2018.

Daniel Z Huang, Kailai Xu, Charbel Farhat, and Eric Darve. Predictive Modeling with Learned Constitutive Laws from Indirect Observations. *arXiv preprint arXiv:1905.12530*, 2019.

Bernardo Llanas, Sagrario Lantarón, and Francisco J Sáinz. Constructive approximation of discontinuous functions by neural networks. *Neural Processing Letters*, 27(3):209–226, 2008.

Michael B Marcus. Weak convergence of the empirical characteristic function. *The Annals of Probability*, pages 194–201, 1981.

Christian Menn and Svetlozar T Rachev. Calibrated FFT-based Density Approximations for α-stable Distributions. *Computational statistics & data analysis*, 50(8):1891–1904, 2006.

Michael H Neumann, Markus Reiß, et al. Nonparametric Estimation for Lévy Processes from Low-frequency Observations. *Bernoulli*, 15(1):223–248, 2009.

John P Nolan. An Overview of Multivariate Stable Distributions. *Online: http://academic2.american.edu/~jpnolan/stable/overview.pdf*, 2008.

Mark JL Orr et al. Introduction to radial basis function networks, 1996.

Antonis Papapantoleon. An Introduction to Lévy Processes with Applications in Finance. *arXiv preprint arXiv:0804.0482*, 2008.

David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

Gennady Samorodnitsky, Murad S Taqqu, et al. Lévy Measures of Infinitely Divisible Random Vectors and Slepian Inequalities. *The Annals of Probability*, 22(4):1930–1956, 1994.

Peter Tankov. Pricing and hedging in exponential lévy models: review of recent results. In *Paris-Princeton Lectures on Mathematical Finance 2010*, pages 319–359. Springer, 2011.

Matus Telgarsky. Neural networks and rational functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3387–3393. JMLR. org, 2017.

Yue Wu, Hui Wang, Biaobiao Zhang, and K-L Du. Using Radial Basis Function Networks for Function Approximation and Classification. *ISRN Applied Mathematics*, 2012, 2012.

Jingxiao Xu and Ted Belytschko. Discontinuous radial basis function approximations for meshfree methods. In *Meshfree methods for partial differential equations II*, pages 231–253. Springer, 2005.

Kailai Xu and Eric Darve. The Neural Network Approach to Inverse Problems in Differential Equations. *arXiv preprint arXiv:1901.07758*, 2019.

Jun Yu. Empirical Characteristic Function Estimation and its Applications. *Econometric reviews*, 23(2):93–123, 2004.