Learning Max-Stable Representations that Extrapolate

Ali Hasan[1] **Patrick Kuiper**[2] **Yuting Ng**[2] **Jose Blanchet**[3] **Vahid Tarokh**[2]

[1]Machine Learning Research, Morgan Stanley, New York, NY
[2]Electrical and Computer Engineering, Duke University, Durham, NC
[3]Management Science and Engineering, Stanford University, Stanford, CA

Abstract

High-dimensional images and measurements may indicate extreme and rare behavior, such as advanced stages of cancer or material failure, and understanding these distributions is important for many disciplines. Classically, extreme value theory (EVT) provides a theoretical framework for extrapolating based on extremeness of *magnitude* of measurements, and it is used to model the tail of a distribution from limited observations. We propose a framework for learning representations from high dimensional observations that are amenable to analysis using classical EVT. Specifically, we propose extending the $\max$-stability property of EVT to φ-stability, which generalizes the $\max$ operator to a more general operator φ that has practical applications in high dimensional cases. We base φ-stability on representation learning techniques such that the resulting representations lend themselves to analysis by EVT and can model high-dimensional observations of extreme characteristics. This enables our method to extrapolate to observations of extreme behavior in the observation domain. We then extend our method to infinite dimensional observations such as time series. Empirical results indicate the utility of using max-stability for representations to extrapolate beyond the training data.

1 INTRODUCTION

Extrapolation beyond the support of training data is a long-standing goal in machine learning, with various techniques introduced to enable such properties. To motivate this idea, we will begin with the problem of characterizing the distribution of abnormal pathologies in medical imaging. In such settings, pathologists often have some examples of normal medical images with a decreasing number of diseased cases as a function of the severity of the disease, often with extremely small numbers of the most severe cases. To assist in image characterization, it may be important to have access to a method for the generation of examples of the most severe pathologies, particularly those not present in the observed data, via an effective extrapolation technique.

The above and many other applications, such as those in risk estimation and understanding the behavior of rare events, motivate us to study extrapolation to regions of a data distribution with low density. On the surface, it may seem impossible to describe regions of a distribution where there is little data. However, the mathematical tools of EVT provide a solution when extremity is defined by the *magnitude* of observations. In cases such as images that indicate extreme behavior, such as advanced stages of cancer or material failure, these are often supported within some known domain (e.g. the pixel depth of a sensor) and the data may not exhibit large magnitudes. This motivates our approach in this paper, where we consider the extraction of *latent variables* from the data where these variables are amenable to analysis using extreme value theory (EVT). EVT specifies that tail data inherit a property known as max-stability. This property enables extrapolation to low density areas based on the observed data, analogous to the familiar central limit theorem describing sums of random variables. We use max-stability to relate observed data to its extrapolated behavior.

We will now recall that EVT dictates that, for certain distributions, properly scaled and shifted maximum of n independent realizations of data converges to *extreme value distributions* as $n \to \infty$ [De Haan and Ferreira, 2006]. Historically, EVT provides an asymptotic answer to the question: "How will the data distribution look in cases that the data values are very large?" As mentioned, for high-dimensional data, all components growing very large simultaneously does not often happen nor is it a useful description of the corresponding "tail" of the data. In the case of image data, for example, it may not be the case that all pixel values are large for data in the tails – rather, the appearance of a particular structure

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:2075–2084.

corresponds to data in the tails. Instead, it can be indicative that some (possibly latent) variable is extreme that generates the high risk observation. In other words, some hidden factors are large and result in the rare or high risk observation rather than the values of the observation itself. We use this idea to guide our framework by describing a factorization of the observation in terms of latent variables/vector distributed according to an extreme value distribution.

To give some ideas about how these variables are related to different cases where data are in the tail of the distribution, we present a few examples.

Example 1.1 (Pathology data)**.** *Consider a data set which exhibits pathology slides of various grades of cancers. The different grades are established through features a pathologist designates important (e.g. break down of the cell). The latent variable may correspond to how diffusive the cells are within the background material and can be characterized by a max-stable latent variable.*

Example 1.2 (Material failures)**.** *Consider a dataset of crystallography data that corresponds to material failures with varying levels of severity. An engineer may be interested in simulating worst case scenarios to design optimal strategies to mitigate against severe effects due to material fatigue. The latent variable can correspond to the number of gaps between neighboring atoms with particular configurations corresponding to more severe failures.*

Example 1.3 (Environmental disasters)**.** *Consider satellite images given in the aftermath of different natural disasters of different severity. The latent variable can roughly correspond to the extent of damage accrued during the event.*

In all these examples, we suppose there is some underlying latent process where components of the process representation are in the distributional tail. Then, the question becomes *"can we write the observations as a function of a latent variable whose extremes correspond to the (semantic) extremes of the observations?"* For this case, we need not observe extreme values (i.e. values that are physically large), but we represent the data in terms of variables that grow very large. It is desirable that this representation of extreme data have the following properties:

1. *Extremes in the latent domain correspond to extremes in the observation domain.*
2. *Extrapolation in the latent domain enables extrapolation in the observation domain.*

These properties seek a model that has latent factors corresponding to tail data while maintaining consistency with the observation data. Intuitively, this means rare or extreme observations in the observation domain should correspond to latent domain representations that are deeper in the tails of the feature space. To explicitly define how the model extrapolates, we introduce a type of *stability*, where certain operations on independent and identically distributed samples preserve the shape of the distribution. While well known examples such as α-stability or $\max$-stability exist, we are interested in a characterization that is appropriate for high-dimensional data.

Definition 1.4 (Max-Stable Distribution)**.** *Let $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ be n independent and identically distributed (iid) samples drawn according to the distribution P. If there exists sequences $\{a_n\} \in \mathbb{R}^d$ and $\{b_n\} \in \mathbb{R}^d$ such that $a_n \max\{X_1, X_2, \ldots, X_n\} + b_n \sim P$, where the multiplication, addition and maximum are taken component-wise for $X \in \mathbb{R}^d$, then we say that P is* max-stable*. That is, if the maximum of iid samples from P is also distributed according to P up to a scale and shift, then P is max-stable. We may write $\vee_{i=1}^n X_i$ for $\max\{X_1, X_2, \ldots, X_n\}$ in the sequel.*

Max-stability is a useful property since it indicates that as long as we know the shape of a distribution, the tails of the distribution have the same shape, but differ in location and scale. We refer the interested reader to De Haan and Ferreira [2006] for additional background on EVT. As in the motivating example, stability with respect to the $\max$ operator may not be relevant in many high-dimensional data cases. In addressing this issue, we propose a new notion of distributional stability in the latent domain. Consider a well-defined mapping $\varphi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ such that for all $X_i, X_j, X_k \in \mathbb{R}^d$ in the image domain: $\varphi(X_i, X_i) = X_i$; $\varphi(X_i, X_j) = \varphi(X_j, X_i)$; and $\varphi(\varphi(X_i, X_j), X_k) = \varphi(X_i, \varphi(X_j, X_k))$. For $n \geq 3$, we can recursively define $\varphi(\{X_1, X_2, \ldots, X_n\})$ by $\varphi(X_1, X_2, \ldots, X_n) = \varphi(\varphi(X_1, X_2, \ldots, X_{n-1}), X_n)\})$. It can be seen that $\varphi(X_1, X_2, \ldots, X_n)$ is invariant to permutations of $X_1, X_2, \ldots, X_n$. In this light, we may write $\varphi(\{X_i\}_{i=1}^n)$ for $\varphi(X_1, X_2, \ldots, X_n)$. Given φ as above, we can now generalize the concept of $\max$-stability.

Definition 1.5 (φ-Stable Distribution)**.** *Let $X_1, X_2, \ldots, X_n \in \mathbb{R}^d \sim P(X)$ be iid samples drawn according to the distribution P in image domain. Consider an operator φ as above and let $X^{(n)} := \varphi(\{X_i\}_{i=1}^n)$. P is said to be φ-stable, if there exists an invertible map $g : \mathbb{R}^d \to \mathbb{R}^k$ such that $g(\varphi(\{X_i\}_{i=1}^n)) = \max(\{g(X_i)\}_{i=1}^n)$, and $g(X)$ is max-stable.*

For a φ-stable distribution as above, if g has inverse $f : \mathbb{R}^k \to \mathbb{R}^d$, we have $\varphi(\{X_i\}_{i=1}^n) = f(\max(\{g(X_i)\}_{i=1}^n)$. This means that there exists an *encoder* $g : \mathbb{R}^d \to \mathbb{R}^k$ and a *decoder* $f : \mathbb{R}^k \to \mathbb{R}^d$ from images in the image domain to latent variables in $\mathbb{R}^k$ such that the latent domain representation of images is $\max$-stable. Having provided this construct, we can approximate g and f by neural networks (or other expressive parametric models) such that latent vector $z = g(X)$ is $\max$-stable and the relationship between

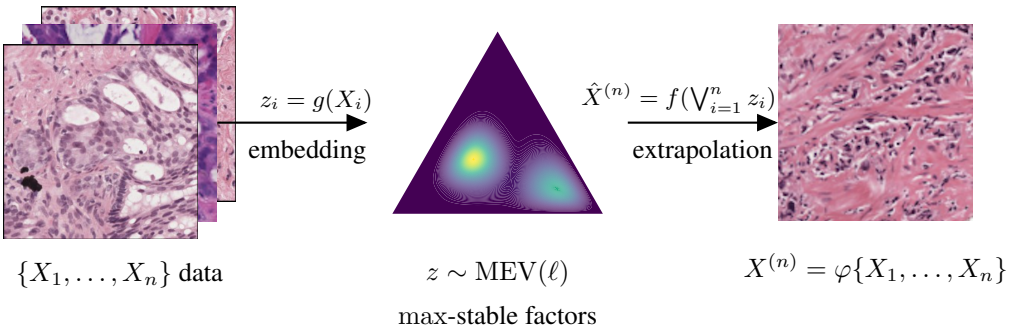


Figure 1: An overview of the method presented; prostate cell pathology slides are transformed to underlying max-stable variables which allows extrapolation according to φ. The embedded images are low Gleason grade images whereas the extrapolated image is a high Gleason grade. Pathology images from Bulten et al. [2022].

the max stability of the latent variables and φ-stability of the observed variables holds. This implies that the *shape* of the latent distribution is known even in regions that we have no data and which allows for extrapolation to unseen data. As we describe the computational method, we will make some approximations in our implementation to make training and generation computationally feasible.

Interpretation of φ φ may be interpreted as a generalization of the $\max$ operator that extrapolates in the observation space. As noted in the motivation, $\max$ in the observation space may not be appropriate in many domains such as imaging (e.g. the $\max$ of many images would result in a fully saturated image) so φ instead recombines a set of observations into one that is at least as extreme as all of its arguments. To provide intuition, we next consider a few examples. The first is when $\varphi(\{X\}_n) = \max_{i=1}^n X_i$, where the $\max$ operator is taken component-wise. If we suppose that the distribution of X is also $\max$-stable, by letting $g(X) = X$ we can include max-stability as a special case of φ-stability when $\varphi = \max$. As another example, suppose that X_i is an image sample of the manifestation of a disease and φ chooses the image with the greatest severity out of n samples. Although an analytical expression for φ is not known, it can be implemented by including labels ξ for image X that correspond to disease severity giving pairs (X, ξ) of data. In this case, $\varphi(\{X\}_n) = X_{\text{argmax}(\xi_n)}$ as depicted in Figure 2. In the case of crack propagation in materials, φ could be a segmentation algorithm that indicates where a failure occurred in a material and synthesizes a new, severely failed material based on the observations. Concretely, if $A \subset X$ is the region where cracks have occurred, $\varphi(X_i, X_j) = X_i \mathbb{1}_{x \in A | x \in X_i} + X_j \mathbb{1}_{x \in A | x \in X_j} + \frac{1}{2}(\phi(X_i) + \phi(X_j))$ with ϕ describing the non-extreme information. Finally, suppose φ is a feature extractor that operates on a set of observations. This extractor localizes sets of relevant features $A \subset \mathbb{R}^d$ such that $\varphi(X_i, X_j) = X_i \mathbb{1}_{x \in A | x \in X_i} + X_j \mathbb{1}_{x \in A | x \in X_j} + \sum_{k=i,j} w_k h(X_k) : \mathbb{R}^d \to \mathbb{R}^d$ for some function h describing non-extreme components of the data.

1.1 RELATED WORK

A variety of work exists that considers factorizations of extremes, but these generally consider the case where all data are in the tail of the distribution and well described by max-stable observations. For example, hidden regular variation considers how different components of a vector become extreme simultaneously and the underlying dependencies between components [Resnick, 2002]. Clustered Archimax copulas [Chatelain et al., 2022] attempts to achieve a similar goal in the sense that the framework models different components with different dependencies. A number of machine learning techniques have been developed that combine some of the properties of EVT. In Rudd et al. [2017], the authors considered using EVT for finding data points far away from the bulk of the data distribution. In terms of neural networks, [Hasan et al., 2022] presented an approach which represents max-stable distributions for multi-dimensional distributions. Ng et al. [2022] described a neural approach based on combining regions of the bulk and the tail of the distribution, but did not consider this for latent observations. Some other examples include [Cont et al., 2022] where the authors consider techniques for sampling from tail events, but do not consider the latent structure of the distribution. In [Améndola et al., 2022] the authors consider a max-linear framework where dependencies between different variables are given by a graphical model. In all cases, the methods are concerned with the cases of extreme observations and not the case of observations corresponding to an underlying extreme event. From the perspective of distribution shift, Dong and Ma [2023] provide idealized conditions under which a model can extrapolate. Other works such as Krueger et al. [2021], Kong et al. [2024] propose perturbing the data space to robustify classifiers distributionally over the extrapolation set, but do not consider the problem of estimating the extrapolated distribution. In the present work, our goal is to extend the factorization of extremes in the data domain to representations of extremes in the latent domain.

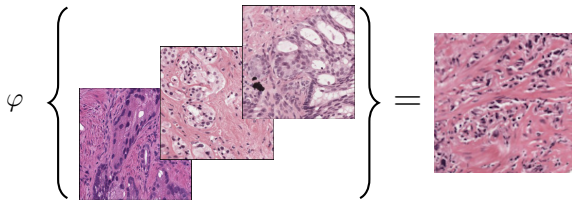


Figure 2: Illustration of the max-stable functional φ operating on a set of prostate cancer cell images. Given images with varying Gleason grades, φ selects the image with the highest severity. Cell images from Bulten et al. [2022].

2 LATENT FACTORIZATION

Suppose that φ and samples $\{X_i\}_{i=1}^N$ from a φ-stable distribution are given. Our approach is to learn the encoder g and decoder f defined above by constructing approximating Deep Neural Networks (DNNs), or other parametric models g_Θ and f_Ω using training data $\{X_i\}_{i=1}^{N'}$ (where N' is the upper-bound index of training data sequence). Our training method is based on forming subsets S_n of the training data of size n (for various values of n) and forcing $\varphi(\{X_i\}_{X_i\in S_n})$ and $f(\max(\{g(X_i)\}_{X_i\in S_n})$ be close to each other (in terms of some application specific distance). We will validate the approach by forming subsets V_n of the hold-out data (for various values of n) and checking if $\varphi(\{X_i\}_{X_i\in V_n})$ and $f(\max(\{g(X_i)\}_{X_i\in V_n})$ are close to each other (in terms of some application specific distance). Once validated, we will use the max-stability of $g(X)$ to generate extreme samples.

To illustrate our approach, we continue with the example of pathology images and consider the image domain to be prostate cancer histopathology slides used for Gleason grading, following Bulten et al. [2022]. Low Gleason grades correspond to healthier, well-differentiated cellular structures, while high Gleason grades correspond to images lacking clear cellular organization.

For any set of pathology images $\{X_i\}_{i=1}^n$, let $\varphi(\{X_i\}_{i=1}^n)$ denote the image in the set with the highest Gleason score. An example is shown in Figure 2, where applying φ to a set of images with increasing grades returns the image with the greatest severity.

All images, regardless of Gleason grade, are embedded into a max-stable latent domain. In this domain, extrapolation to higher grades is performed by taking the maximum over latent representations and decoding the result using f, yielding an image corresponding to unhealthy, poorly-differentiated cells. During training, we enforce agreement between the decoded latent maximum and the observed image with the maximum Gleason score.

Next, we describe the main components of the proposed modeling technique. To this end, we briefly review the required mathematical background and introduce relevant notation.

2.1 MATHEMATICAL BACKGROUND

Let Δ_{k-1} denote the k-dimensional simplex. As in the above, consider a k-dimensional multivariate extreme value (MEV) distribution $z = (z^1, z^2, \cdots, z^k)$. MEVs are often represented using their copula which separates modeling univariate marginals from multi-dimensional dependence, and that the copula $C(u) = C(u^1, \cdots, u^k)$ of z can be described by the stable tail dependence function (stdf) [De Haan and Ferreira, 2006, Ressel, 2013]:

Definition 2.1 (stable tail dependence function)**.** *A k-variate stdf $\ell : [0,\infty)^k \to [0,\infty)$ is defined as:*

$$\ell(u) = k \int_{\Delta_{k-1}} \max_{j\in\{1,\cdots,k\}} \{u^j w^j\}\, \mathrm{d}\Lambda(\mathbf{w}) \tag{1}$$

where $\Lambda(w) = \Lambda(w^1, \cdots, w^k)$ is a corresponding spectral measure for a spectral random variable $\mathbf{W} \in \Delta_{k-1}$ satisfying moment constraints $\mathbb{E}_{\mathbf{\Lambda}}[W_j] = 1/k$ for $j \in \{1, \cdots, k\}$ [Ressel, 2013].

We note that the marginals $z_1, z_2, \cdots, z_k$ of z are generalized extreme value (GEV) distributions. Once these marginals and the stable tail dependence $\ell(u)$ are described, the MEV distribution of z is entirely known. We refer to ℓ as the *stdf* of z and write $\mathrm{MEV}(\ell, z)$ for a MEV with stdf ℓ for a random vector z whose marginals of $z_1, z_2, \cdots, z_k$ are specified by normalized GEVs (often normalized Frechet margins).

The decomposition of ℓ in equation 1 illustrates that the Λ provides the dependence between variables, and samples of MEVs may be represented according to *radial* and *spectral* components [De Haan and Ferreira, 2006]. Specifically, if we let $Y_1, Y_2, \cdots$ be iid samples drawn according to a spectral distribution $\Lambda(\cdot)$ and A_n be the n^{th} arrival-time of a unit rate Poisson point process, then a max-stable random variable is represented as

$$z = \max_{n\geq 1} \frac{Y_n}{A_n}. \tag{2}$$

In light of the above, an alternative way to model z is to specify the marginals and the spectral measure $\Lambda(\cdot)$. Samples of A_i can be generated by adding i independent samples of a unit rate exponential random variable.

2.2 INFERENCE PROCEDURE

Given iid data observations $\{X\}_n := \{X_i\}_{i=1}^n$ with each observation in $\mathbb{R}^d$, we assume that each X_i is generated by applying the decoder $f : \mathbb{R}^k \to \mathbb{R}^d$ to a max-stable

latent vector $z_i \in \mathbb{R}^k$, where $z_i = g(X_i)$ i.e. $\hat{X}_i = f(z_i)$ where $\hat{X}_i$ are the decoded values. Let $m_n = \max_{i=1\dots n} z_i$ where max is taken component-wise over all k components of z. Let $\hat{X}^{(n)} = f(m_n)$, where we enforce φ-stability by requiring that $\hat{X}^{(n)}$ and $\varphi(\{X_i\}_{i=1}^n) = X^{(n)}$ be as close to each other as possible for a domain-specific distance $\mathcal{L}(\hat{X}^{(n)}, X^{(n)})$ for any choice of n. This minimization is accomplished by an appropriate choice of stdf ℓ (equivalently spectral measure $\Lambda(\cdot)$ and decoder f, i.e. we must estimate the pair (f, ℓ) or equivalently (f, Λ)). We must also account for the fact that z_i are required to be max-stable in the design of g.

Since max-stable distributions have a specific algebraic structure that describes the extrapolation behavior, we will exploit this to design an appropriate decoder structure that allows for extrapolation in the ambient space, which we formalize in this remark:

Remark 2.2 (Condition for Extrapolation)**.** *Let φ satisfy the above conditions and suppose that the true observation model is given by $X = f(z)$. Then, if $f(z)$ satisfies the equality in distribution*

$$f(a_n z + b_n) \overset{d}{=} \varphi\left(f(z_1), \dots, f(z_n)\right)$$

where z_i are iid samples of z, then we say that the model extrapolates according to φ.

This remark allows us to sample $a_n z + b_n$ in the latent space and enforce consistency in the observation space. We discuss further implications and architectures based on this in Appendix A. In terms of practical implementation, the function f should have dependence on the batch size to allow for extrapolation.

Inference Using Spectral Representation Building off of this connection, we can sample latent vectors using a Monte Carlo method and extrapolate through sampling. Rather than directly calculating the scale and shift parameters, a_n and b_n, of the MEV for each iteration, we consider a sampling approach using the spectral representation. As such, our main tool in this approach is given by equation 2, and the spectral measure $\Lambda(\cdot)$. From equation 2, the only component we must estimate is the distribution of Y_n given by Λ, since A_n already has a prescribed form. Then, sampling the max-stable vector z may be approximately achieved by sampling from Y_n and calculating the maximum over a finite number of samples, as is similarly done in Dombry et al. [2016]. When considering this finite approximation, we establish an unnormalized version Γ of the spectral measure Λ; that is, $\Lambda = \Gamma / Z_\Gamma$, where Z_Γ is the normalizing partition function. If instead of Λ, we use Γ in calculating z in equation 2, then this representation is equivalent to the scaling of g by Z_Γ. This may be compensated for during training by scaling the decoder f by a factor of $1/Z_\Gamma$. Subsequently, assuming that Λ is not normalized has no effect on the training algorithm. Furthermore, from the moment constraints $\mathbb{E}_{\mathbf{W}}[W_j] = 1/k$ in equation 2, it can easily be seen that $\mathbb{E}[Y_i] = (1/k, 1/k, \cdots, 1/k)$ for $i = 1, 2, \cdots, n$. If these constraints are relaxed, then components of Y_i and z will be scaled. Again, we observe that this scaling may be compensated for in the decoder f in the training phase. In this light, without loss of generality, we can relax these conditions. We note that this also follows from EVT as presented by [De Haan, 1984], which establishes that relaxing these conditions only results in the distribution of components of z to be unnormalized (for example to non-unit Frechét margins).

Algorithmic Development The proposed inference procedure using extrapolation conditions motivates the development of Algorithms 1- 6 for designing extrapolating models (where Algorithms 1- 6 are provided in Appendix D). Given $\Lambda(\cdot)$, there exists a neural network $K_\Theta(X)$ with parameters Θ such that $Y = K_\Theta(X)$ is approximately distributed according to $\Lambda(\cdot)$. This neural network must have normalization in the last layer (such as softmax) to ensure that $K_\Theta(X) \in \Delta_{k-1}$. The existence of such an approximating neural network is akin to the construction of normalizing flows where arbitrary distributions are constructed from given (typically Gaussian) distributions. We approximate $\max_{n \geq 1} \frac{Y_n}{A_n}$ in equation 2 with a finite number of samples (size of mini-batch). These observations lead to training Algorithm 1 in Appendix D, which learns $K_\Theta(\cdot)$ mapping images to samples from the spectral distribution. To generate samples from the trained encoder / decoder neural networks, we present Algorithm 2 in Appendix D, which considers a classical max-stable sampling algorithm as described in Dey and Yan [2016, Chapter 9] based on extremal Gaussian processes.

Extrapolation Implications As a consequence of this latent formulation, we can gauge how well a network trained on observations up to $\varphi(\{X\}_k)$ is expected to perform on observations $n > k$. We will impose a few assumptions:

1. The latent distribution z_i has standard Gumbel margins;
2. There exists a true decoder $f_\star(M_n) \overset{d}{=} X^{(n)}$;
3. The error $e(x) = \|f_\star(x) - f(x)\|_\infty$ is L_e-Lipschitz (i.e. $\|e(x) - e(y)\| \leq L_e \|x - y\|_\infty$.

Now, suppose that $\mathbb{E}[\|f(M_k) - X^{(k)}\|] < \varepsilon$ for all $k \leq n$ then for any $n > k$,

$$\mathbb{E}[\|f(M_n) - X^{(n)}\|] \leq 2\varepsilon + 2L_e^2 \log^2\left(\frac{n}{k}\right).$$

The proof is provided in Appendix B. As a consequence, the error grows roughly logarithmically as we extrapolate deeper into the tail of the distribution.

3 INFINITE DIMENSIONAL OBSERVATIONS

So far we discussed a model using a finite dimensional max-stable vector in the latent space and assumed a finite dimensional observation space. However, many applications require data that are sampled from an infinite dimensional space where the dimensionality may change according to some features. For example, we can consider a time series of sensor traces that result in component failure. In the individual traces, there may not be a signature that approaches infinity, but a latent factorization of this time series in terms of a max-stable process may be helpful. We will consider these objects as φ-stable processes. Consider a domain $\mathcal{D} \subset \mathbb{R}^k$ and elements $u \in \mathcal{D}$ that act as an index of the domain. We suppose that our observations are given by functions $X(u) : \mathcal{D} \to \mathbb{R}^d$. For the finite dimensional case, we specified a function φ that maps realizations of data to more extreme realizations. In the infinite dimensional case, we instead define an **operator** Φ acting upon a set of functions $\{X_i(u)\}_{i=1}^N$. Similarly to the finite dimensional case, the larger the N, the more extreme the observation transformed by Φ is.

The latent space is then governed by a *max-stable process* rather than a finite dimensional multivariate extreme distribution. Specifically, we have functions $a_n(u), b_n(u)$ analogous to a_n, b_n in Definition 1.4. We can also define the Φ-stable processes in an analogous manner to φ-stable distributions. Max-stable processes enjoy similar factorizations in the sense that a spectral distribution is specified. Specifically, a max-stable process $z(u)$ can be written as $z(u) = \max_{n \geq 1} \frac{Y_n(u)}{A_n}$ where $Y_n(u)$ are independent copies of a non-negative spectral process $Y(u)$ satisfying $\mathbb{E}[Y(u)] = 1$, and A_n are samples of n-the arrival time of a unit rate homogeneous Poisson process. We note that finite dimensional sampled versions of max-stable processes are max-stable distributions. Given this, the methods we discussed above readily extend to φ-stable processes with some minor modification.

4 EXPERIMENTS

We consider a series of experiments to learn the support of the latent distribution and illustrate the extrapolation capabilities of the φ-stable representation method. As a baseline for comparisons to the proposed φ-stable method, we employ alternate models formulated to achieve identical results including a conditional variational autoencoder (cVAE) [Sohn et al., 2015], extreme autoencoder (Ext-VAE) [Lafon et al., 2023], and Tail-GAN [Cont et al., 2022]. Detailed descriptions of each experiment and data set are available in Appendix F.

We demonstrate the effectiveness of the φ-stable representation method by modeling the extremes of noisy sinusoidal curves, showcasing the ability to extrapolate beyond the bounds of the observed data. We employ two approaches: a max-stable multi-layer perceptron (MLP) encoding a fixed length time domain, and a max-stable convolutional neural network (CNN) encoding a variable length domain.

Synthetic Experiments We evaluate the φ-stable model on extrapolation tasks beyond the bounds of observed (training) data by constructing a set of observations calculated from a sinusoidal curve with a multiplicative noise component. This data is given by the function $s_i(t) = \log(t) \times \sin(t) \times \epsilon_i + c$, where $i = 0, \ldots, n$ is the number of noisy observations, $t = t_0, \ldots, t_m$ is a range of values, and c remains fixed for all $i = 0, \ldots, n$. The noise ϵ_i is sampled from one of three distributions: log-normal, Weibull, and uniform. The anticipated behavior of the proposed φ-stable model is to minimize loss between the extrapolated curve and the maximum over all observations, $\max_{i=0}^n s_i(t)$ for all $t \in \{t_0, \ldots, t_m\}$ of the noisy curves. To demonstrate this performance, we provide the mean squared error (MSE), denoted by ℓ_m, between the maximum values of the training data, $m_{n,t} = \max_{i=0}^n s_i(t)$ for all $t = t_0, \ldots, t_m$, and the extrapolated maximum values $\hat{m}_{n,t} = \max_{i=0}^n \hat{s}_i(t)$ for all $t = t_0, \ldots, t_m$ with $\ell_m = \mathbb{E}_t[(\hat{m}_{n,t} - m_{n,t})^2]$, where $n = 1 \times 10^5$. To achieve the desired performance, ℓ_m should be small. Two methods are used for baseline comparison: a conditional variational autoencoder (cVAE), and an extreme variational autoencoder (Ext-VAE), as proposed for similar extrapolation tasks by Sohn et al. [2015] and Lafon et al. [2023] respectively. As demonstrated in Table 3 and Figure 4, we find that the φ-stable model is able to extrapolate to the maximum curve of the noisy data set consistently, with monotonically decreasing MSE as the size of the blocks increases, while neither the cVAE nor the Ext-VAE baselines achieve this goal.

We extend this experiment to the infinite dimensional case where we sample φ-stable processes. In this experiment there are multiple noisy trigonometric curves, encoded using a CNN, where the φ-stable process is used to extrapolate based on latent factors. We provide the results of the technique in Figure 5, where we see the model exhibits the anticipated behavior of extrapolating outside the training data, progressing to higher values as the block size is increased for all channels. We compare the errors in Figure 4d for the different methods. Further information concerning the function for the noisy sinusoidal curves used to generate each channel of data is available in Appendix F.2.

Extrapolating Fractal Images We next study how different methods extrapolate on fractal data. Fractals are an ideal test since the scale of the fractal can represent the degree of extremeness while not including meaningfully large magnitudes of values in the observation space. We test this on three different fractal families: Sierpinski carpet, Barnsley

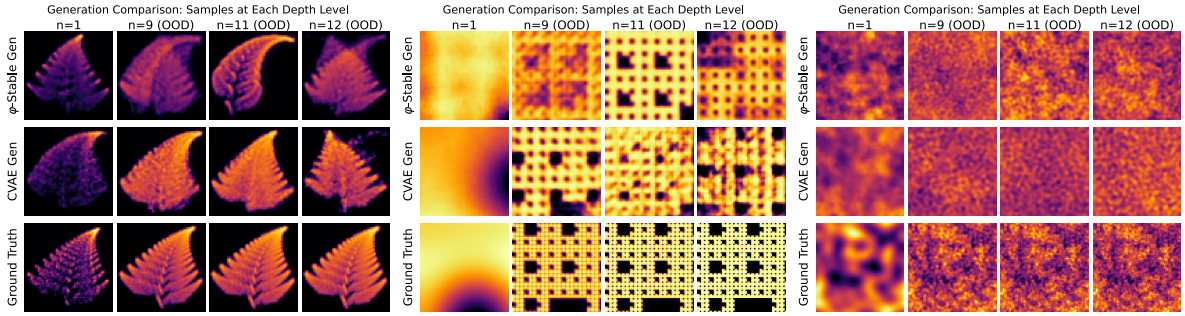


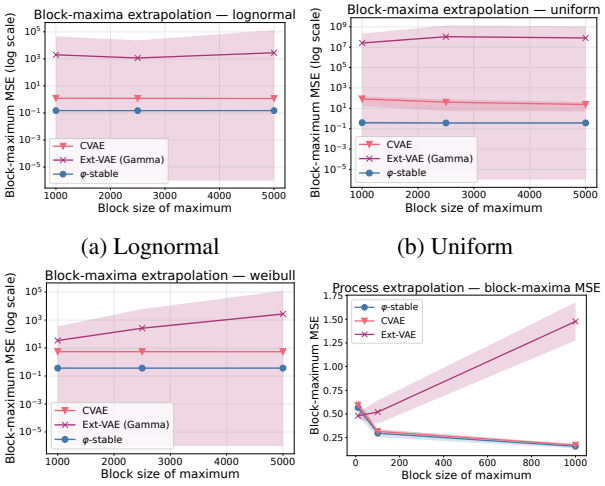
Figure 3: Comparison of generation quality for Barnsley fern, Sierpinski, and fractional Brownian motion fractal datasets.

Table 1: Out-of-distribution sampling (generation) MSE on the fractal experiments, averaged over the OOD depth levels $n = 9, \ldots, 12$. Values are mean $\pm$ std (pooled across OOD levels).

Fractal	cVAE	φ-stable	Tail-GAN
fBm	1.67×10^{-2}	5.80×10^{-3}	3.39×10^{-2}
IFS	1.12×10^{-2}	8.40×10^{-3}	3.88×10^{-2}
Sierpinski	9.76×10^{-2}	4.46×10^{-2}	1.56×10^{-1}

fern, and fractional Brownian motion. In our experimental setup, we observe a fixed level of fractal scales and try to extrapolate to unseen scales. Examples for the different fractals are given in Figure 3.

We compute the MSE on held out scales of generated data in Table 1. The results indicate the stability of the φ-stable method in recovering the distribution of the data within the tails of the distribution.



(a) Lognormal (b) Uniform

(c) Weibull (d) Multi-channel

Figure 4: Comparison of MSE values for the φ-stable model with Ext-VAE and cVAE. Ext-VAE's variance dominates most of the plot.

4.1 REAL DATA EXPERIMENTS

Financial Max-Drawdown using φ-stable Factors We estimate latent max-stable factors that contribute to the simultaneous worst-case drawdown of multiple blue-chip components of the S&P 500 index. We encode the daily percent change in stock values from six large and diversified corporations traded on the S&P 500. We observe in Figure 6a, that as the block size is increased, the performance of the φ-stable model continues to improve. We again compare this to a standard cVAE which is trained to generate max-drawdown values conditioned on observed minimum daily percent changes over five day trading periods, for varying block sizes. Details concerning the experimental setting and data used are available in Appendix F.3.

Extrapolating Climate Data We reference data collected by Gershunov et al. [2017], concerning atmospheric rivers in the western United States. Atmospheric rivers are narrow bands of moisture that transport heavy rainfall. Annual rainfall in California specifically varies according to the presence of major winter storms, which makes drought highly dependent on atmospheric river activity. We assess the magnitude of atmospheric rivers, and related drought conditions, based on vertically integrated water vapor measurements recorded by satellites based on data in Gershunov et al. [2017]. This provides analysts with conditions of the worst-case drought scenario in multiple locations simultaneously. A visual performance comparison is provided in Figure 6c, where we observe qualitatively (and quantitatively confirmed) that the φ-stable model outperforms the cVAE in predicting the worst-case drought conditions during the five measurement periods in all six locations. A quantitative comparison between the baseline cVAE and φ-stable model error is given in Figure 6b, where we observe the anticipated behavior as block size increases.

Extrapolating Handwriting Based on Age We reference a dataset of handwritten numerical digits conditioned on the age of the writer labeled for each example [Beaulac and Rosenthal, 2022]. With this dataset we demonstrate the φ-

Model	Classifier FID		InceptionV3 FID		
	Recon	Sample	Recon	Sample	Recon MSE
cVAE	1.56×10^3	1.61×10^3	4.46×10^2	4.47×10^2	2.63×10^{-2}
φ-stable	7.04×10^2	7.04×10^2	1.66×10^2	1.55×10^2	7.05×10^{-3}

Table 2: Comparison of FID and MSE for the proposed φ-stable method and a cVAE baseline on conditional and unconditional generation for out-of-distribution Gleason score 3 images. Unconditional MSE is the minimum MSE over pairings between real and generated examples.

Age	φ-stable	cVAE
7	6.00×10^3	1.13×10^4
8	2.55×10^3	3.59×10^3
9	1.82×10^3	1.96×10^3

Table 4: Comparison of validation error for out-of-sample ages. The φ-stable model performs increasingly better for younger ages.

stable model's ability to construct out of sample children's handwriting. We train the φ-stable model with image data as the features and the age of the writer as context for the max-stable distribution. For evaluation, we consider the model's ability to reconstruct and generate the handwritten digits of the youngest members of the dataset: 7, 8, and 9-year-olds, while training on ages 12 and older. The reconstruction error for all out of sample data (specifically for $9-$digit) of the φ-stable model, compared to a baseline cVAE, is provided in Table 4. We observe the φ- stable model's outperformance over all out-of-sample ages, and increasing improvement (relative to the cVAE baseline) as the writers' age is younger (more extreme). We provide further results, including validation digit examples, with age comparisons in Appendix F.4.

Extrapolating to Rare Prostate Cancer Pathology Slides We consider the problem described in the motivation where we extrapolate to unseen pathology slides for prostate cancer based on the data in Bulten et al. [2022]. For this experiment, each slide image is labeled as an integer of $\{0, 1, 2, 3\}$ representing the grade of cancer found in the slide. Our goal is to train on grades $\{0, 1, 2\}$ and determine how well we can a) conditionally reconstruct grade 3 image based on an unseen input image and b) estimate how well we can sample within the region of the grade 3 images without conditioning on an input image. For b) we sample a batch of images within this region and use the smallest error to the target set as our metric for validation. The results are illustrated in Table 2 for the quantitative comparison on the two experiments. We include two additional metrics based on the Fréchet Inception Distance (FID) using both a classifier trained to classify pathology slides and the traditional InceptionV3 architecture. The results indicate that the φ-stable model successfully extrapolates to the grade 3 images, especially in the case where we do not condition on an input image.

5 DISCUSSION

In this paper, we proposed an extension of the usual use cases of EVT to scenarios where the observation data can be factorized into a function of extreme data. Specifically, we represent the data in terms of max-stable latent variables. The $\max$-stability of the latent variables is then used to extrapolate outside the support of the training set. We illustrated the promise of the proposed method on both synthetic and real data. As future work, we will consider properties of f that allows extrapolation in the ambient space. Using the monotonic behavior of the latent distribution when extrapolating, we can develop a specific neural architecture that exploits this property to guarantee extrapolation in the observation space. Additionally, considering the max-stable processes in the context of other applications could be helpful. For example, in large language models, it may be that when embeddings reach a region of the state space that has low density, hallucinations occur. Informing the latent representations with the max-stable algebra could be a method of anticipating embeddings in this region and be used as a way to prevent hallucinations within models.

Future Directions The idea that we can use latent representations endowed with $\max$-stability to extrapolate outside of the training domain can be applied in multiple settings. Here, we simply use the encoder-decoder approach to demonstrate a proof of concept. There are multiple avenues for extension that may have more applicability than the initial domain we studied. First, we can think of how to modify sequence based architectures to encode $\max$-stability in their representations. Architectures such as the transformer can be modified in a myriad of ways to include $\max$-stability, and this is a strong direction for studying. Second, we can use this knowledge to design more optimal pre-training distributions. Recent work [Azizian and Hasan, 2026] sheds light on how the tail parameter of the pre-training distribution affects performance for in-context learning, and a unified analysis that connects the latent dis-

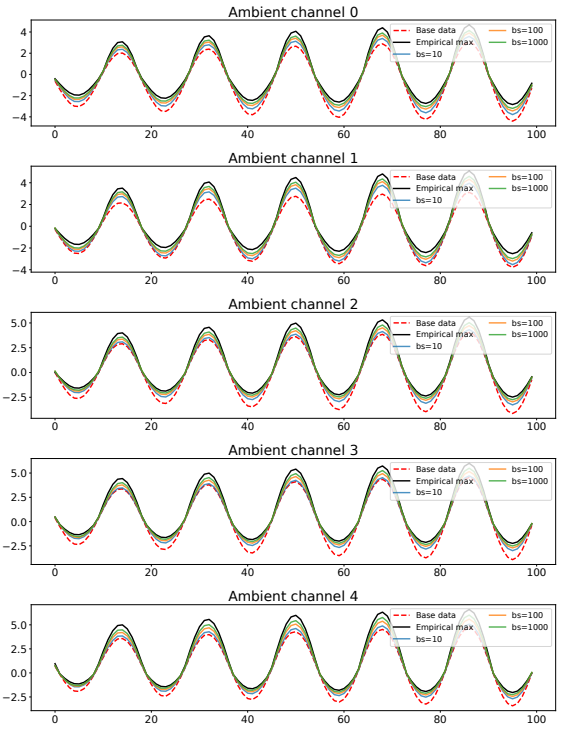


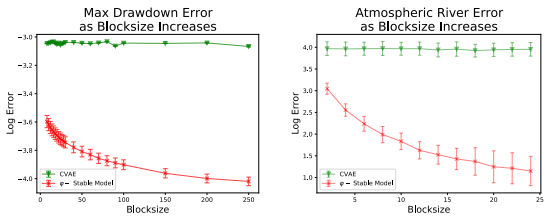
Figure 5: Extrapolation over 5 observation dimensions outside the training data set using the φ-stable model.

	φ-stable	Ext-VAE	cVAE
Lognorm.	2.26×10^{-1}	1.64×10^{0}	2.84×10^{-1}
Weibull	3.51×10^{-2}	7.12×10^{0}	1.43×10^{0}
Uniform	6.84×10^{0}	6.38×10^{5}	2.11×10^{4}

Table 3: Comparison of the φ-stable method with Ext-VAE and cVAE baselines ($n = 1 \times 10^5$). Values represent MSE.

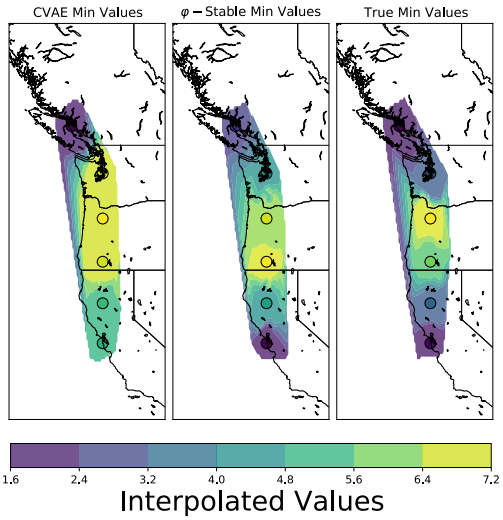
tribution with the pre-training distribution could be helpful to understand when new, out of distribution concepts can be learned from data.

Limitations Without explicitly constraining the function class of f, it is difficult to guarantee extrapolation in the ambient space. In this work, we considered validation sets with data that are more extreme than the observations that are trained on. However, this validation technique may fail for data that is extremely far from the validation set. In that sense, constraining f such that its extrapolation properties are known provides a constraint that guarantees extrapolation and circumvents this problem. Additionally, when φ is unknown, we require information from a semantic perspective on how extreme an example would be. An alternative would be understanding the whole data distribution and finding regions of low support automatically.



(a) S&P 500 drawdowns.

(b) Atmospheric events.



(c) Visual drought comparison

Figure 6: Comparison of the φ-stable model against a cVAE baseline two application settings - extreme S&P 500 daily drawdown and spatial drought conditions. (a) Extrapolation of extreme daily drawdowns for selected S&P 500 components error analysis. (b) Extrapolation of extreme atmospheric river events error analysis. (c) Spatial extrapolation to extreme drought conditions. Across all cases, the proposed model achieves lower error and scales appropriately with block size.

Broader Impacts The proposed methods are intended to be useful in risk estimation settings where a reasonable estimate of tail data is needed. The ideal use-case is in situations where out of distribution data should be estimated and the method can be used to ensure that algorithms are more robust against data occurring in the tails of the distribution and out of distribution examples.

Acknowledgements

Material in this paper is partially based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397.

REFERENCES

Carlos Améndola, Claudia Klüppelberg, Steffen Lauritzen, and Ngoc M Tran. Conditional independence in max-linear bayesian networks. *The Annals of Applied Probability*, 32(1):1–45, 2022.

Waïss Azizian and Ali Hasan. How does the pretraining distribution shape in-context learning? a fundamental trade-off. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=QiTBmWiH8G.

Cédric Beaulac and Jeffrey S. Rosenthal. Introducing a new high-resolution handwritten digits data set with writer characteristics. *SN Comput. Sci.*, 4(1), November 2022. doi: 10.1007/s42979-022-01494-2. URL https://doi.org/10.1007/s42979-022-01494-2.

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022.

Simon Chatelain, Samuel Perreault, Johanna G. Nešlehová, and Anne-Laure Fougères. Clustered archimax copulas, 2022.

Rama Cont, Mihai Cucuringu, Renyuan Xu, and Chao Zhang. Tail-gan: Nonparametric scenario generation for tail risk estimation. *arXiv preprint arXiv:2203.01664*, 2022.

L. De Haan. A Spectral Representation for Max-stable Processes. *The Annals of Probability*, 12(4):1194 – 1204, 1984.

Laurens De Haan and Ana Ferreira. *Extreme value theory: an introduction*, volume 3. Springer, 2006.

Dipak K Dey and Jun Yan. *Extreme value modeling and risk analysis: methods and applications*. CRC Press, 2016.

Clément Dombry, Sebastian Engelke, and Marco Oesting. Exact simulation of max-stable processes. *Biometrika*, 103(2):303–317, 2016.

Kefan Dong and Tengyu Ma. First steps toward understanding the extrapolation of nonlinear models to unseen domains. In *The Eleventh International Conference on Learning Representations*, 2023.

Alexander Gershunov, Tamara Shulgina, F. Martin Ralph, David A. Lavers, and Jonathan J. Rutz. Assessing the climate-scale variability of atmospheric rivers affecting western north america. *Geophysical Research Letters*, 44(15):7900–7908, August 2017. doi: 10.1002/2017GL074175. URL https://ui.adsabs.harvard.edu/abs/2017GeoRL..44.7900G. Provided by the SAO/NASA Astrophysics Data System.

Ali Hasan, Khalil Elkhalil, Yuting Ng, João M Pereira, Sina Farsiu, Jose Blanchet, and Vahid Tarokh. Modeling extremes with d-max-decreasing neural networks. In *Uncertainty in Artificial Intelligence*, pages 759–768. PMLR, 2022.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

Lingjing Kong, Guangyi Chen, Petar Stojanov, Haoxuan Li, Eric Xing, and Kun Zhang. Towards understanding extrapolation: a causal lens. *Advances in Neural Information Processing Systems*, 37:123534–123562, 2024.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021.

Nicolas Lafon, Philippe Naveau, and Ronan Fablet. A vae approach to sample multivariate extremes, 2023. URL https://arxiv.org/abs/2306.10987.

Yuting Ng, Ali Hasan, and Vahid Tarokh. Inference and sampling for archimax copulas. *arXiv preprint arXiv:2205.14025*, 2022.

Sidney Resnick. Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5:303–336, 2002.

Paul Ressel. Homogeneous distributions—and a spectral representation of classical mean values and stable tail dependence functions. *Journal of Multivariate Analysis*, 117:246–256, 2013.

Ethan M Rudd, Lalit P Jain, Walter J Scheirer, and Terrance E Boult. The extreme value machine. *IEEE transactions on pattern analysis and machine intelligence*, 40 (3):762–768, 2017.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.