Representation Learning for Extremes

Ali Hasan* Duke University
Yuting Ng Duke University
Jose Blanchet Stanford University
Vahid Tarokh Duke University

Abstract

Extreme events are potentially catastrophic events that occur infrequently within an observation time frame, and it is necessary to understand the distribution of these events to properly plan for them. Extreme value theory provides a theoretical framework for extrapolating to the tails of a distribution using limited observations. However, for high-dimensional data such as images, covariates are generally not extreme but perhaps the features are extreme. In this work, we propose a framework for learning representations according to properties of extreme value theory. Specifically, we use the max-stability property of extreme value distributions to inform the representations of the model such that they extrapolate to the rare data observations. We theoretically characterize the properties of the model and provide an identifiability result for the parameters of the latent distribution. Our preliminary results suggest the promise of the method for extrapolating to regions of the distribution with little density.

1 Introduction

In this paper, we are primarily interested in describing regions of a data distribution with low density. This has particular importance in the setting of estimating risks or the behavior of data within the tail of a distribution. At first, this task may appear to be misguided — it seems impossible to describe regions of a distribution where there is little data by virtue of the lack of data. However, if we consider a factorization of the observations in terms of a *latent variable* that has a known extrapolation behavior, then we may be able to study the behavior in the tails based on the data we have observed. This structure is given by extreme value theory (EVT) which dictates that, for certain distributions, the maximum of n realizations converges to distributions with known form referred to as extreme value distributions. The EVT framework has historically been useful for representing a distribution in the tails. Thus, EVT tries to answer the question: “How will the data look in cases that the data are very large (or small)?"

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 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 factor is large and results 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 a latent variable distributed according to an extreme value distribution.

Concretely, consider the example of characterizing the distribution of abnormal pathologies in medical imaging. One can suppose that these abnormalities are generated through some underlying physiological process where components of the process representation are in the tail. Then, the question becomes *"can we write the observations as a function of a latent variable whose extremes*

*This work was supported in part by the Air Force Office of Scientific Research under award number FA9550-20-1-0397.

HeavyTails Workshop at 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

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.

Building on these thoughts, we will study representations of extreme data using two properties that we want our model to have:

> *(P1: Consistency) Extrapolation in the latent space leads to extrapolation in the observed space.*
> *(P2: Identifiability) Dependence structure of the latent variable can be identified from the observations.*

The first property (P1) requests we specify a model that has latent factors corresponding to tail data while maintaining consistency with the observation data. Intuitively, this means rare or extreme observations should correspond to representations deeper within the tails of the feature space. The second idea (P2) involves characterizing under which conditions of the mapping the parameters can correspond to a unique parameterization of an extreme value distribution. By studying the properties of this parameter, we are able to make more informed decisions about the behavior of the risky data.

To explicitly define how the model extrapolates, we introduce a type of *stability*, where certain operations on iid 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 1.1 (Max-Stable Distribution)**.** Let $X_1, X_2, \ldots, X_n \sim P$ be iid samples from P. If $a_n M_n + b_n = a_n \max\{X_1, X_2, \ldots, X_n\} + b_n \sim P$, where the maximum is 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.

Max-stability is a useful property since it says 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 Haan and Ferreira [7] for additional background on EVT. Analogously, consider n iid samples from a distribution $\{X_i\}_{i=1}^n$ and an operation φ that acts on $\{X_i\}_{i=1}^n$ such that $\varphi(\{X_i\}_{i=1}^n)$ has the *same* distribution up to a shift and scale of parameters dependent on n. We then define this distribution to be φ-stable.

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. For example, hidden regular variation considers how different components of a vector become extreme simultaneously and the underlying dependencies between components [11]. Clustered Archimax copulas [3] 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 with neural networks, e.g. in [8]. Some other examples include [5] where the authors consider techniques for sampling from tail events but do not consider the latent structure of the distribution. In [1] 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. In the present work, our goal is to extend the factorization of extremes in the ambient space to representations of extremes in the latent space.

2 Latent factorization

Having motivated the problem, we are now in the position to describe the latent factorization of the generative model. As previously noted, we want to consider the joint distribution of the observations X and latent variables z such that:

(a) The latent variable z is max-stable so that the shape of the distribution is known even in regions that we have no data;

(b) The ambient variable is consistent with the $\max$-stability of z in terms of φ, which is a generalization of the usual $\max$ operator from the latent space to the ambient space.

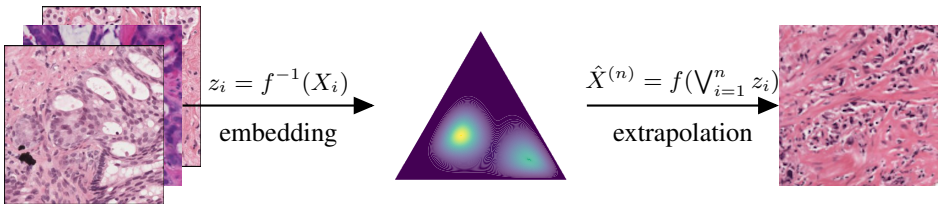

Figure 1: An overview of the method presented; prostate cell pathology slides are transformed to underlying max-stable variables which allows extrapolation according to φ. The embedded images are low Gleason grade images whereas the extrapolated image is a high Gleason grade. Pathology images from Bulten et al. [2].

To do this, we propose a latent variable model where the latent space is governed by a multivariate extreme value (MEV) distribution. This provides the max-stability property where taking the maximum of a series of realizations results only to changes in the scale and shift parameters of the distribution and not a change in the shape of the distribution. An overview of this concept is illustrated in Figure 1 with an example using pathology slides of prostate cancer cells for Gleason grading from Bulten et al. [2]. Images of all Gleason grades, where low Gleason grades correspond to images with healthy, well-differentiated cells, are embedded to the max-stable latent space. There, extrapolation to a higher grade is performed by taking the maximum over the latent variables. The maximum over the latent variables is then decoded according to f to obtain the high Gleason grade image corresponding to unhealthy, poorly-differentiated cells. We seek a correspondence between the decoded maximum and the observed maximum where the grading of the observed pathology slides is assumed to follow φ.

We consider the case where labels that correspond to the level of extremeness in the ambient space are given. In this case, we may train the ambient variable to be consistent with the max-stability of the latent variable. If labels are not given, we need to examine and constrain the relationship between the ambient and latent variables. For example, in the case of max-stable observations, we can constrain the decoder to be monotonic.

With these in mind, we proceed to describe the main components of the modeling technique.

2.1 Joint distribution

Denote $\mathrm{MEV}(\ell, \xi)$ as an MEV distribution with stable tail dependence function (stdf) ℓ and extremal index ξ. When ξ is not explicitly denoted, the extremal index is assumed to be 1. Suppose we are given iid data observations $\{X\}_n := \{X_i\}_{i=1}^n$ with each observation in $\mathbb{R}^d$ and assume that each X_i is generated as some function $f : \mathbb{R}^k \to \mathbb{R}^d$ of a latent variable $z \in \mathbb{R}^k$, i.e. $\hat{X} = f(z)$ where $\hat{X}$ are the noiseless decoded values, plus observation noise ε with distribution p_ε. Additionally, we suppose there exists a convex function $\varphi : \mathbb{R}^n \to \mathbb{R}$ that operates on the set $\{X\}_n$. Let $m_n = \max_{i=1\ldots n} z_i$ where the $\max$ is taken component-wise over all k components of z. Then, we suppose that $\hat{X}^{(n)} = f(m_n) = \varphi(\{X\}_n) = X^{(n)}$. Putting this together, we can write a generative model that is given by

$$\begin{aligned} P\left(X^{(n)}\right) &= \int p_\varepsilon \left(f(m_n) - X^{(n)}\right) p(m_n) \mathrm{d}m_n \\ &= \int p_\varepsilon \left(f(a_n z + b_n) - X^{(n)}\right) p_{a_n, b_n}(z) \mathrm{d}z \end{aligned} \tag{1}$$

where p_{a_n, b_n} describes the density of z scaled by $a_n \in \mathbb{R}^k$ and shifted by $b_n \in \mathbb{R}^k$. As noted above, since $p(z)$ is MEV, its parameters are given by an stdf ℓ when margins are appropriately normalized. This leaves us with the parameters we must estimate: (f, ℓ).

A point process viewpoint of $p(z)$ The correspondence between EVT and point processes is well known (see, e.g., in Coles et al. [4, Chapter 7]). Samples of an MEV are known to follow an

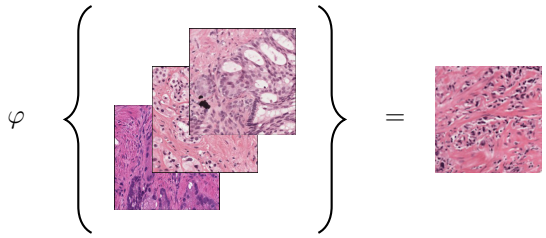

Figure 2: Example of φ operating on a batch of prostate cell images. Samples from the low Gleason grade images are combined according to φ to form the high Gleason grade images. Cell images from Bulten et al. [2].

inhomogenous Poisson point process N with intensity:

$$\mathbb{E}[N(\mathcal{C})] = \mu(\mathcal{C}) := \int_{\mathcal{C}} r^{-2} \mathrm{d}r H(\mathrm{d}w)$$

for a Borel subset $\mathcal{C}$ of $[0, \infty) \times \mathbb{R}_+^d$ where r corresponds to the radial component, w corresponds to the spectral component, and H is known as the *spectral measure*. It can be shown that this expectation corresponds to ℓ, with the details described in Coles et al. [4, Theorem 9.2].

In the case of the model (1), we can instead consider the latent distribution corresponding to the Poisson process related to the MEV distribution conditioned on the lack of observations in a particular region. This is then given by

$$P\left(X^{(n)}\right) = \int p_\varepsilon \left(f\left(a_n N_z + b_n\right) - X^{(n)}\right) p\left(N_z\left(\mathcal{C}^{a_n, b_n}\right) = 0\right) \mathrm{d}N_z. \tag{2}$$

Working with the point process perspective, we can write $p\left(N_z\left(\mathcal{C}^{a_n,b_n}\right) = 0\right) = \exp\{-\ell(a_n z - b_n)\}$ in terms of the stdf, which will become useful when proving the identifiability of ℓ.

2.2 Inference procedure

Recall that a major goal is to enforce a consistency between the max-stable latent variable and the observed variable. We can construct a new dataset by considering $\{\varphi(\{X\}_n)\}$, $n \in \mathcal{P}(\{1, \ldots, N\})$, where the dataset is first partitioned into blocks of size n, then the max-stable latent variable correspond to the (semantic) max observation $\hat{X}^{(n)} = f(m_n) = \varphi(\{X\}_n) = X^{(n)}$. Note that φ can remain unknown and only labels corresponding to n are needed (i.e. we only need $(n, X^{(n)})$) assuming that $X^{(n)}$ was generated according to the model above. The goal is to infer the tuple of parameters (ℓ_ϕ, f_θ) corresponding to the dependence of z and ambient mapping f. This results in a minimization problem given by

$$\min_{\phi,\theta} \mathbb{E}_{n \sim \mathcal{P}(\{1 \ldots N\})} \mathbb{E}_{z \sim \mathrm{MEV}_\phi(f^{-1}(\{X\}_n))} \left[\mathcal{L}\left(\varphi(\{X\}_n), f_\theta(M_n)\right)\right]. \tag{3}$$

In the case of unknown φ and known labels, the minimization problem is simply reduced to

$$\min_{\phi,\theta} \mathbb{E}_{(n, X^{(n)}) \sim P} \mathbb{E}_{z \sim \mathrm{MEV}_\phi(f^{-1}(\{X\}_n))} \left[\mathcal{L}\left(X^{(n)}, f_\theta(M_n)\right)\right].$$

2.3 Interpretation of φ

φ can be interpreted as a generalization of the $\max$ operator that extrapolates in the ambient space. As noted in the motivation, $\max$ in the ambient space may not be appropriate in many domains such as imaging, so other functions may be necessary. To provide intuition, we consider a few examples. The first is the case where $\varphi(\{X\}_n) = \max_{i=1}^n X_i$, where the max operator is taken component-wise. This implies that the distribution of x is also max-stable. The second is the case where $\varphi(\{X\}_n) = \sum_{i=1}^n w_i f^{-1}(X_i) + b_i$ with $\sum w_i = 1$ which can be thought of as the output of a single layer neural network acting upon the empirical measure induced by $\{X\}_n$. More abstractly, 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. Then, the consistency between the max-stable latent variable z and the ambient observation X is measured by the disease severity. This can be done by including labels ξ that correspond to disease severity (X, ξ), such that $\varphi(\{X\}_n) = X_{\text{argmax}(\xi_n)}$ and optimizing (3) accordingly. Thus φ generalizes $\max$ in the ambient space to $\max$ in the latent space. Figure 2 shows how φ could possibly act on a batch of data such that the rare sample is generated.

2.4 Choosing $\mathcal{L}$

The distance metric $\mathcal{L}$ used in (3) plays a particular role in the optimization procedure. We will discuss two cases: one where φ leads to extreme realizations (i.e. when $\varphi(\{X\}_n) = \max_{i=1}^n \{X_i\}$) and the other where φ leads to normal observations.

Extreme observations Assuming that $\varphi(\,\cdot\,) = \max\{\,\cdot\,\}$, X is distributed according to an MEV. We then propose using the interpretation of the MEV in terms of a Poisson point process. Specifically, consider an event denoted by A given as a Borel subset of $\mathbb{R}_+^d$ and let U denote the mapping to polar coordinates $U(X) = (R, W) = (\|X\|_1, X/\|X\|_1)$. We use the notion that the arrivals of the extreme events are distributed according to a Poisson point process with rate $\mu(A) = \int_{U(A)} r^{-2} \mathrm{d}r H(\mathrm{d}w)$ where H is the *spectral measure* denoting the dependence between the dimensions in the observations [6]. The cost function $\mathcal{L}$ can correspond to the agreement between the intensities given by the data and generated by the model over a set of events, i.e. $\mathcal{L} = \mathbb{E}_A[\mu(A) - \mu_\theta(A)]$.

Normal observations In the case of normal observations, we can use p_ε to be the Gaussian likelihood. The resulting likelihood is an objective that minimizes the mean squared error between the reconstructions and the original data. The extrapolation property is still enforced since the consistency between $f(m_n)$ and $\varphi(\{X\}_n)$ is applied.

3 Identifiability

We will now discuss identifiability of the extremal index and of the dependence function. In general, identifiability of the extremal index is difficult to obtain since f can map Z to multiple random variables with varying extremal indices. Instead, we focus on a class of f that preserves the extremal index to circumvent the issues with identifiability. For an MEV with unit Frechét margins given by Z, we can transform the distribution to one with extremal index ξ by Z^ξ/ξ. Since ξ controls the tail decay of the distribution, it is natural to understand under which conditions this parameter is identifiable. We assume that the decoder is a function that is regularly varying with tail index 1. In the following proposition, we describe a particular type of ReLU decoder architecture retains the latent extremal index.

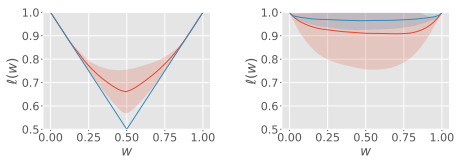

Figure 3: Estimated dependence functions for symmetric logistic distribution, complete dependence (left) and complete independence (right). Red is estimated and blue is ground truth. Confidence bands from 50 trials.

Proposition 3.1 (Regularly Varying Decoder)**.** *Let f be parameterized with a* ReLU *neural network with positive weights and let z be given by* $\mathrm{MEV}(\ell, \xi)$*. Then $X = f(z)$ is regularly varying with index ξ.*

Now we consider identifiability of the dependence function. We can show that, based on the result in Khemakhem et al. [10], the stdf is identifiable up to a translation.

Proposition 3.2. *Consider the model of X in* (2) *motivated by the point process viewpoint:*

$$\mathcal{L}(\cdot; f, \ell) := \int p_\varepsilon \left(f\left(a_n N_z + b_n\right) - X^{(n)} \right) \exp\{-\ell(a_n N_z - b_n)\} \mathrm{d}N_z$$

Then, for any pairs of solutions $(f, \ell), (\hat{f}, \hat{\ell})$ where $\mathcal{L}(\cdot; f, \ell) = \mathcal{L}(\cdot; \hat{f}, \hat{\ell})$, we have $\ell = \hat{\ell} + a$ where a is a constant under the following conditions: i. $f, \hat{f}$ are injective; ii. $p(z; \ell) \ll p(\hat{z}; \hat{\ell})$. Specifically, ℓ is identifiable up to a translation.

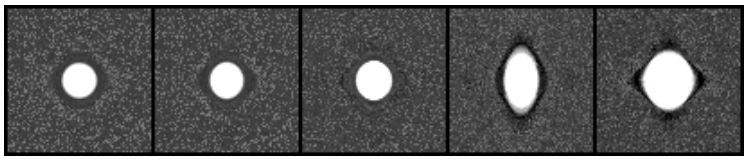

Figure 4: Example of extrapolation. Each image represents successive component-wise maxima in the latent domain. In this case, the latents are independent and the fourth image shows independence between the learned height and width.

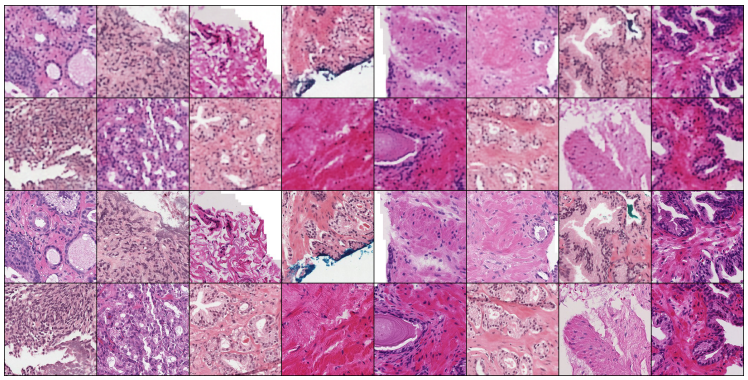

Figure 5: Held out reconstruction for prostate cancer data, each image is 256 x 256 pixels. Top two rows: model reconstruction. Bottom two rows: original data.

The proposition allows us to probe the behavior of the latent distribution according to the traditional tools of EVT. For example, we can characterize the stdf, which describes the clustering behavior of the extremes. This allows for proper planning and mitigation of extreme events.

4 Experiments

We consider a series of experiments to learn the support of the latent distribution and illustrate the extrapolation capabilities of the method. Descriptions of the different datasets are available in Appendix C. The first result is illustrated in Figure 3 where we find the dependence function associated with the data generating distribution of the symmetric logistic distribution. In this experiment, we simulate observations of ellipses where the semi-major and semi-minor axes are generated according to a symmetric logistic distribution with varying dependence parameter. In Figure 4, we consider how well the method is able to extrapolate to new images in the case of independent latents. We sample within the latent space to generate new images and we see that the generated images follow the predicted distribution according to the extrapolation.

In the second experiment, we consider the distribution of prostate cancer images where the extreme data corresponds to the high grade cancer. In Figure 5, we illustrate the reconstruction capabilities on held out data. The results suggest that the method is able to faithfully reconstruct held out data samples.

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 artchitecture that exploits this property to guarantee extrapolation in the observation space.

Acknowledgements

This work was supported in part by the Air Force Office of Scientific Research under award number FA9550-20-1-0397.

References

[1] 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.

[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.

[3] Simon Chatelain, Samuel Perreault, Johanna G. Nelehová, and Anne-Laure Fougères. Clustered archimax copulas, 2022.

[4] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

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

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

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

[8] 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.

[9] Hedegaard Anders Jessen and Thomas Mikosch. Regularly varying functions. *Publications de L'institut Mathematique*, 80(94):171–192, 2006.

[10] 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.

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

A Proofs

Proposition A.1. *Consider the model of X given by the following factorization*

$$\mathcal{L}(\cdot; f, \ell) := \int p_\varepsilon \left(f\left(a_n N_z + b_n\right) - X^{(n)} \right) \exp\{-\ell(a_n N_z - b_n)\} \mathrm{d}N_z$$

Then, for any pairs of solutions $(f, \ell), (\hat{f}, \hat{\ell})$ where $\mathcal{L}(\cdot; f, \ell) = \mathcal{L}(\cdot; \hat{f}, \hat{\ell})$, then $\ell = \hat{\ell} + a$ where a is a constant under the following conditions:

1. *$f, \hat{f}$ are injective;*
2. *$p(z; \ell) \ll p(\hat{z}; \hat{\ell})$.*

Specifically, ℓ is identifiable up to a translation.

Proof. The proof first follows Step I from Khemakhem et al. [10, Appendix B.2.2] where the observation noise distribution is removed and all that is left is the distribution of the latent space. We follow the steps to arrive at the following equivalence:

$$\exp(-\ell(\xi f^{-1}(X))) \mathrm{vol} J_{f^{-1}} \mathbb{1}_{X \in \mathcal{X}} = \exp(-\hat{\ell}(\xi \hat{f}^{-1}(X))) \mathrm{vol} J_{\hat{f}^{-1}} \mathbb{1}_{X \in \mathcal{X}}.$$

where we decompose the max-stable latent variable in terms of a radial component indicated by ξ, re-using the notation for extremal index, and a spectral component indicated by $f^{-1}(X)$. We take logarithms and compute the difference evaluated at ξ_0 to obtain

$$-\ell\left(\xi f^{-1}(X)\right) + \ell\left(\xi_0 f^{-1}(X)\right) = -\hat{\ell}\left(\xi \hat{f}^{-1}(X)\right) + \hat{\ell}\left(\xi_0 \hat{f}^{-1}(X)\right)$$

on the set of $X \in \mathcal{X}$. By Condition 2, we assumed that $p(f^{-1}(X)) \ll p(\hat{f}^{-1}(X))$, and thus we can write the Radon-Nikodym derivative $\frac{\mathrm{d}\hat{p}}{\mathrm{d}p}$. Our strategy to deriving an equivalence between ℓ and $\hat{\ell}$ uses the spectral decomposition of ℓ as an expectation with respect to the spectral component

$$\ell(z) = \int \bigvee_{i=1}^{k} z_i s_i \Lambda(\mathrm{d}s)$$

and the key idea that $p(f^{-1}(X)) \to \Lambda$ to apply the change of measure to Λ and write the equations on the same latent basis:

$$\begin{aligned}
-\int \bigvee_{i=1}^{k} s_i \xi_i f_i^{-1}(X) \Lambda(\mathrm{d}s) + \int \bigvee_{i=1}^{k} s_i \xi_{i,0} f_i^{-1}(X) \Lambda(\mathrm{d}s) & \\
&= -\int \bigvee_{i=1}^{k} s_i \xi_i \hat{f}_i^{-1}(X) \hat{\Lambda}(\mathrm{d}s) + \int \bigvee_{i=1}^{k} s_i \xi_{i,0} \hat{f}_i^{-1}(X) \hat{\Lambda}(\mathrm{d}s) \\
&= -\int \bigvee_{i=1}^{k} s_i \xi_i f_i^{-1}(X) \frac{\mathrm{d}\hat{p}}{\mathrm{d}p} \hat{\Lambda}(\mathrm{d}s) + \int \bigvee_{i=1}^{k} s_i \xi_{i,0} f_i^{-1}(X) \frac{\mathrm{d}\hat{p}}{\mathrm{d}p} \hat{\Lambda}(\mathrm{d}s) \\
&= -\int \bigvee_{i=1}^{k} s_i \xi_i f_i^{-1}(X) \tilde{\Lambda}(\mathrm{d}s) + \int \bigvee_{i=1}^{k} s_i \xi_{i,0} f_i^{-1}(X) \tilde{\Lambda}(\mathrm{d}s)
\end{aligned}$$

with $\tilde{\Lambda} = \frac{\mathrm{d}\hat{p}}{\mathrm{d}p} \hat{\Lambda}$ and letting $\tilde{\ell}(\,\cdot\,) = \int \bigvee_{i=1}^{k} s_i \,\cdot\, \tilde{\Lambda}(\mathrm{d}s)$, we get

$$-\ell\left(\xi f^{-1}(X)\right) + \ell\left(\xi_0 f^{-1}(X)\right) = -\tilde{\ell}\left(\xi f^{-1}(X)\right) + \tilde{\ell}\left(\xi_0 f^{-1}(X)\right).$$

Now we have both $\ell, \tilde{\ell}$ in the same domain, we can establish a linear relationship by conditioning on extrapolations (n) and on different points z_j. Using the max-stability of z where $z^{(n)} = a_n z^{(0)} + b_n$

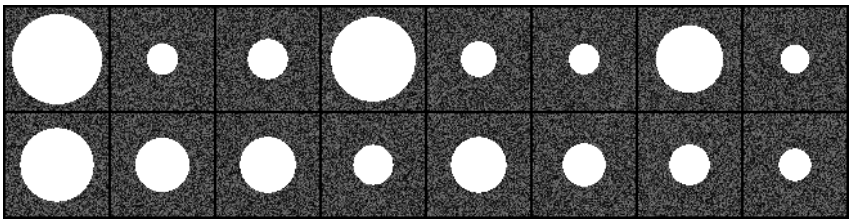

Figure 6: Circle data using the symmetric logistic distribution with $\alpha = 0$.

and the homogeneity of ℓ where $\ell(cx) = c\ell(x)$ for $c > 0$:

$$\begin{pmatrix} -a_1\ell(z_1+b_1)+d_1 & \cdots & -a_1\ell(z_m+b_1)+d_1 \\ \vdots & \ddots & \vdots \\ -a_n\ell(z_1+b_n)+d_n & \cdots & -a_n\ell(z_m+b_n)+d_n \end{pmatrix}$$
$$= \begin{pmatrix} -a_1\tilde{\ell}(z_1+b_1)+\tilde{d}_1 & \cdots & -a_1\tilde{\ell}(z_m+b_1)+\tilde{d}_1 \\ \vdots & \ddots & \vdots \\ -a_n\tilde{\ell}(z_1+b_n)+\tilde{d}_n & \cdots & -a_n\tilde{\ell}(z_m+b_n)+\tilde{d}_n \end{pmatrix}$$

which means $\ell(z) = \tilde{\ell}(z) + \tilde{d} - d$ by continuity and taking $m = n \to \infty$. □

Proposition A.2 (Regularly Varying Decoder)**.** *Let f be parameterized with a* ReLU *neural network with positive weights and let z be given by* $\mathrm{MEV}(\ell, \xi)$*. Then $X = f(z)$ is regularly varying with index ξ.*

Proof. We first follow [9, Lemma 4.6] with the setup that the weights of the network are sampled from Dirac measures centered at the weight value. This provides finite expectation and concludes that the output is regularly varying with index ξ. Nesting this structure and only considering the positive component with the ReLU activation, we obtain the desired result. □

B Implementation Details

All experiments are conducted with f, f^{-1} represented according to a convolutional autoencoder. The latent representation is assumed to be a Gumbel distribution, where the reparameterization trick is used to sample the rare event according to

$$z^{(n)} = \max_{i \geq 1} z_i^{(n)} = \max_{i \geq 1} \left\{ f^{-1}\left(X^{(n)}\right) - \log A_i^{(n)} \right\}$$

where $A_i^{(n)} = \sum_{j=1}^{n} \xi_{i,j},\ \xi_{i,j} \sim \mathrm{Exp}(1)$, and the infinite max is truncated to a finite number, 100. Both the encoder and decoder have 4 layers with width 8, 16, 32, 64 and use the ReLU activation function. We use the AdamW optimizer with learning rate of 0.0005 for 200 epochs.

C Additional Details on Data

C.1 Synthetic Data

For the synthetic experiments, we first sample from a 2d symmetric logistic distribution with parameter α. α is the dependence parameter with $\alpha \to 0$ implying complete dependence and $\alpha \to 1$ implying complete independence. We then scale the values to be within $(0, 1)$ by dividing by the maximum. We then take the semi-major and semi-minor axes to have lengths given by this vector and construct an image with the corresponding ellipse. Examples of samples are given in Figures 6 and 7. Image sizes are 80×80 pixels.

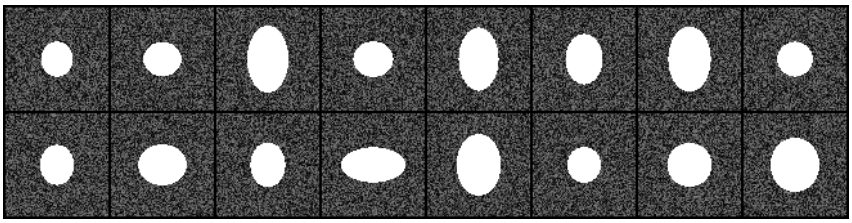

Figure 7: Circle data using the symmetric logistic distribution with $\alpha = 1$.

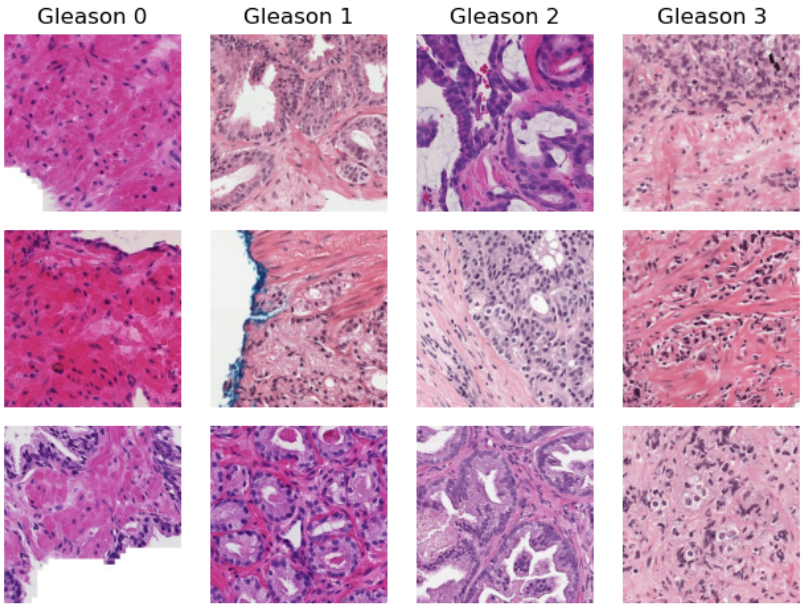

Figure 8: Examples of different types of Gleason scores for prostate cancer data.

C.2 Real Data

For the real dataset, we consider the prostate cancer Gleason score dataset from Bulten et al. [2]. We aim to characterize the extremeness of the data as it relates to the severity of the disease cancer. The examples of the different types of scores are illustrated in Figure 8.