

2014

Reduced perplexity: Uncertainty measures without entropy

Kenric P. Nelson. 2014. "Reduced Perplexity: Uncertainty measures without entropy." Recent Advances in Info-Metrics

<https://hdl.handle.net/2144/28821>

Downloaded from DSpace Repository, DSpace Institution's institutional repository

Reduced perplexity: A simplified perspective on assessing probabilistic forecasts

Kenric P. Nelson^{1,2}

A simple, intuitive approach to the assessment of probabilistic inferences is introduced. The Shannon information metrics are translated to the probability domain. The translation shows that the negative logarithmic score and the geometric mean are equivalent measures of the accuracy of a probabilistic inference. Thus there is both a quantitative reduction in perplexity as good inference algorithms reduce the uncertainty and a qualitative reduction due to the increased clarity between the original set of probabilistic forecasts and their central tendency, the geometric mean. Further insight is provided by showing that the Rényi and Tsallis entropy functions translated to the probability domain are both the weighted generalized mean of the distribution. The generalized mean of probabilistic forecasts forms a Risk Profile of the performance. The arithmetic mean is used to measure the decisiveness, while the $-2/3$ mean is used to measure the robustness.

1 Introduction

The objective of this chapter is to introduce a clear simple approach to assessing the performance of probabilistic forecasts. The goal is to ground the approach in information theory [1], [2] while framing the perspective based on the central tendency and fluctuation of the probabilities used in the evaluation. This approach is important because the role of ‘scoring rules’ [3], [4] which translate probabilities onto a separate utility scale has resulted in confusion regarding how to select an appropriate metric and how the results of that metric should be interpreted. To achieve this objective, Section 2 reviews the relationship between probabilities, perplexity and entropy. The geometric mean of probabilities is shown to be the central tendency of a set of probabilities. In Section 3, the relationship between probabilities and entropy is expanded to include generalized entropy functions [5], [6]. From this, the generalized mean of probabilities is shown to provide insight into the fluctuations of a forecast and thus the risk sensitivity. From this analysis, a *Risk Profile* [7] defined as the spectrum of generalized means of a set of forecasted probabilities, is used in Section 4 to evaluate a variety of models for a n -dimensional random variable.

The goal is to reduce both the perplexity regarding how to evaluate the performance of a forecast and to provide a method which establishes a better standard for designing algorithms which can be verified to reduce perplexity. The clarity provided by expressing average uncertainty as a probability rather than as entropy or perplexity is important in communicating the performance of information systems. Engineers and managers seek performance metrics which are simple and intuitive, so that system objectives can be communicated clearly. As an example, a classification system has two basic requirements;

¹ Raytheon Company, 235 Presidential Way, Woburn, MA 01801, kenric_p_nelson@raytheon.com

² Boston University, 8 St Marys St, Boston, MA 02215, kenricpn@bu.edu

correct classification and accurate probabilities. Reporting the percentage of correct classification is routine and standard. In contrast measuring accurate probabilities has produced a perplexing array of literature for several decades. Grounding the accuracy of probabilities with the geometric mean of reported probabilities will be consistent with information theoretic metrics, while utilizing the intuitive scale of probabilities. Further insight into the risk sensitivity of the predicted probabilities can be achieved using the generalized mean.

2 Probability, Perplexity and Entropy

The arithmetic mean and the standard deviation of a distribution are the elementary statistics used to describe the central tendency and uncertainty respectively of a random variable. Less widely understood, though studied as early as the 1870s by McCalister [8], is that a random variable which is formed by the ratio of two independent random variable has a central tendency determined by the geometric mean rather than the arithmetic mean. Probabilities, which are always formed from a ratio, thus require the geometric mean to measure the central tendency. We will see in the next section, that the central tendency of probabilities formed from non-exponential distributions, can be analyzed with the generalized mean, but this a refinement. The main point is recognizing the importance of the geometric mean with regard to uncertainty analysis.

Long-standing tradition within mathematical physics has utilized the entropy function to express the average uncertainty of a distribution. There are at least three important reasons for this, physically entropy defines the change in heat energy per temperature, mathematically entropy provides an additive scale for measuring uncertainty and computationally entropy has been shown to be a measure of information[9], [10]. Unfortunately, what is lost is the intuitive relationship between the underlying probabilities of a distribution and a summarizing average probability of the distribution. Perplexity, which determines the average number of uncertain states, provides a bridge between the average probability and the entropy of a distribution. For a random variable with a uniform distribution of n states, the perplexity is n . For other distributions, the perplexity is determined by taking the exponential of the entropy. The central tendency of the distribution or the average probability of the distribution is defined by the inverse of the perplexity. Understanding the relationship between average probability P_{avg} , perplexity PP and entropy H provides a valuable perspective on quantifying average uncertainty. These relationships can be summarized as

$$P_{avg} \equiv PP^{-1} = \exp(-H(\mathbf{p})) = \exp\left(-\sum_{i=1}^N p_i \ln p_i\right) = \prod_{i=1}^N p_i^{p_i} \quad (2.1)$$

where $\mathbf{p} = \left\{ p_i : \sum_{i=1}^N p_i = 1 \right\}$ is a probability distribution. The natural logarithm is used here for the entropy function, though base 2 is also common, particularly within information theory. For a continuous distribution $f(x)$ of a random variable X these expressions become

$$f_{avg} \equiv PP^{-1} = \exp(-H(f(x))) = \exp\left(-\int_{x \in X} f(x) \ln f(x) dx\right), \quad (2.2)$$

where f_{avg} is the average density of the distribution and PP still refers to perplexity.

Figure 1 illustrates these relationships for the standard normal distribution. The key point is that by expressing the central tendency of a distribution as a probability (or density for continuous distributions), the context with the original distribution is

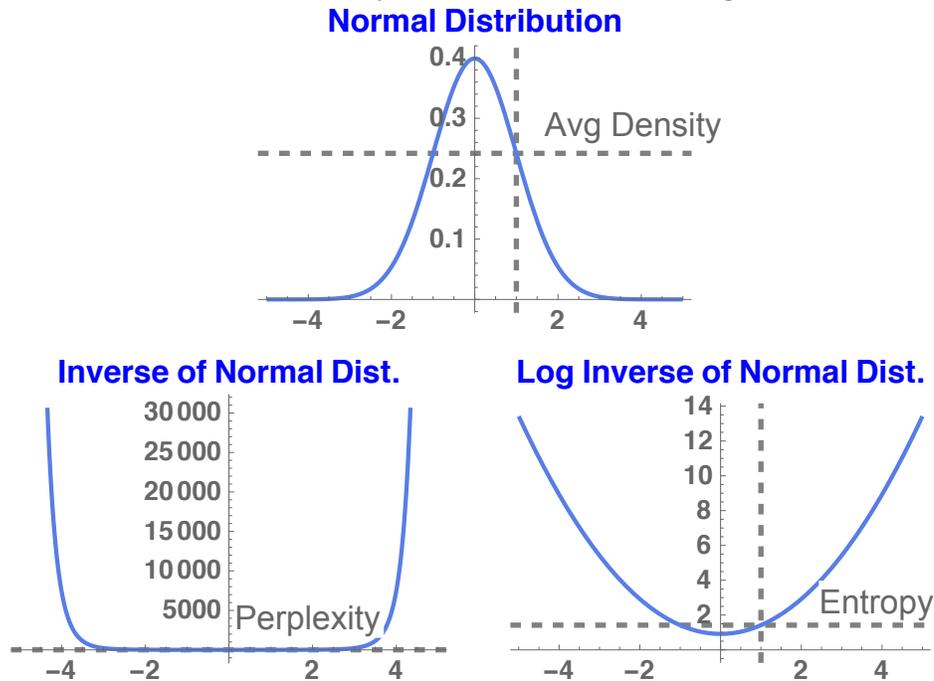


Figure 1 Comparison of the average density, perplexity and entropy for the standard normal distribution. Plots of the inverse distribution and the log of the inverse of the distribution provide visualization of the perplexity and entropy. The intersection for each of these quantities with the distribution is at the mean plus the standard deviation.

maintained. For the exponential and Gaussian distributions, translating entropy back to the density domain [11] results in the density of the distribution at the location μ plus the scale σ

$$\exp\left(\int_{\mu}^{\infty} \frac{1}{\sigma} \exp\left(-\frac{x-\mu}{\sigma}\right) \ln\left(\frac{1}{\sigma} \exp\left(-\frac{x-\mu}{\sigma}\right)\right) dx\right) = \frac{e^{-\left(\frac{\mu+\sigma-\mu}{\sigma}\right)}}{\sigma} = \frac{1}{\sigma e} \quad (2.3)$$

$$\exp\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ln\left(\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx\right) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(\mu+\sigma-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi e\sigma}}. \quad (2.4)$$

Thus, it should be more commonly understood that for these two important members of the exponential family, the average uncertainty is the density at the width of the distribution defined by $f(\mu+\sigma)$. While perplexity and entropy are valuable concepts, it is not common to plot distributions on the inverse scale (perplexity) or the log inverse scale (entropy), thus the intuitive meaning of these quantities is disconnected from the underlying distribution.

Table 1 shows the translation of three basic entropy functions to the perplexity and probability scales. In each case the translation is e^H and e^{-H} , respectively for the perplexity and probability. While the perplexity is useful in counting the number of combinations represented by the uncertainty, it still lacks the simplicity of representing uncertainty as a probability. The additive combination of logarithmic probabilities, translates into a multiplicative combination of the probabilities. The weight on the mean, also a probability, is now a power term. The weighted geometric mean is then the proper method for determining the central tendency of the probability or average probability of a distribution. While an optimal decision would have an uncertainty equal to the maximum probability of the distribution, the central tendency of the distribution can be thought of as representing the uncertainty associated with wages proportional to the probability of each state of the distribution.

Similarly, the additive relationship between cross-entropy, entropy, and divergence $H(p,q) = H(p) + D_{KL}(p||q)$, is multiplicative in the probability space

$$\begin{aligned} P_{cross-entropy} &= P_{entropy} P_{divergence} \\ &= \left(\prod_{i=1}^N p_i^{p_i}\right) \left(\prod_{i=1}^N \left(\frac{q_i}{p_i}\right)^{p_i}\right) \\ &= \prod_{i=1}^N q_i^{p_i}. \end{aligned} \quad (2.5)$$

Jaynes [12], [13] established the principal of maximum entropy as a method for selecting a probability distribution such that known constraints were satisfied, but no additional knowledge was represented in the distribution. Two basic examples are the exponential distribution, which satisfies the constraint that the range is 0 to ∞ and a known mean $\mathbb{E}(X) = \int_0^{\infty} xf(x)dx = \mu$, and the Gaussian distribution which satisfies a known mean and variance $\mathbb{E}(X^2) - \mathbb{E}(X)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \sigma^2$. The principal of maximum entropy

Table 1: Translation of entropy functions to perplexity and probability scales

Info-Metric	Entropy Scale	Perplexity Scale	Probability Scale
Entropy	$-\sum_i p_i \ln p_i$	$\prod_i (p_i)^{-p_i}$	$\prod_i (p_i)^{p_i}$
Divergence	$-\sum_i p_i \ln \left(\frac{q_i}{p_i} \right)$	$\prod_i \left(\frac{q_i}{p_i} \right)^{-p_i}$	$\prod_i \left(\frac{q_i}{p_i} \right)^{p_i}$
Cross-Entropy	$-\sum_i p_i \ln q_i$	$\prod_i (q_i)^{-p_i}$	$\prod_i (q_i)^{p_i}$

could thus be framed in the probability domain as a minimization of the weighted geometric mean of the distribution. In section 4 a related principal of minimizing the cross entropy between a discrimination model and the actual uncertainty of a forecasted random event will be translated to maximizing the geometric mean of the reported probability.

Just as the arithmetic mean of the logarithm of a probability distribution determines the central tendency of the uncertainty or the entropy, the standard deviation of the logarithm of the probabilities, $\sigma_{\ln p}$ is needed to quantify variations in the uncertainty,

$$\sigma_{\ln p} \equiv \left[\sum_{i=1}^N p_i (-\ln p_i)^2 - \left(-\sum_{i=1}^N p_i \ln p_i \right)^2 \right]^{1/2}. \quad (2.6)$$

Unfortunately, the translation to the probability domain, $e^{-\sigma_{\ln p}}$ does not result in a simple function with a clear interpretation. Furthermore, because the domain of entropy is one-sided, just determining the standard deviation does not capture the asymmetry in the distribution of the logarithm of the probabilities. Instead, the next section demonstrates that the generalized mean can be used to measure fluctuations about the central tendency measured by the geometric mean.

3 Relationship between the generalized entropy and the generalized mean

In this section the effect of sensitivity to risk will be used to generalize the assessment of probabilistic forecasts. The approach is based on generalizations of the entropy function as is shown to related to economists measure of relative risk aversion. As with the Boltzmann, Gibbs, Shannon entropy, the generalized entropy can be transformed back to the probability domain. The resulting function, for several different generalizations and particularly the Rényi and Tsallis entropies [14]–[16] is the weighted generalized mean or weighted p -norm of the probabilities. Using the symbol r , to avoid confusion with

p for probabilities and q for the traditional variable of Tsallis entropy the metric given a vector of probabilities \mathbf{p} is

$$P_r(\mathbf{w}, \mathbf{p}) \equiv \left(\sum_{i=1}^N w_i p_i^r \right)^{\frac{1}{r}}. \quad (2.7)$$

The symbol P_r is used here rather than the traditional symbols of M_r or $\|x\|_r$ for the generalized mean and p -norm respectively to emphasize that the result is a probability which represents a particular aggregation of the vector of probabilities. The weights must sum to one for the function to represent a mean. While several different generalizations of entropy can be shown to transform into the form of Equation (2.7) the definitions for weights and the variable r can differ. Appendix A provides details of the derivation from the generalized entropy functions. The weights are a modified set of probabilities formed by raising the probabilities of the distribution to a power and renormalizing and is referred to as the coupled probability

$$P_i^{(r)}(\mathbf{p}) \equiv \frac{p_i^{1-r}}{\sum_{j=1}^N p_j^{1-r}}. \quad (2.8)$$

This new distribution is the normalized probability of $1-r$ independent events rather than one event. Substituting Equation (2.8) for the weights in Equation (2.7) and simplifying gives the following expression for the weighted generalized mean of a distribution

$$P_r(\mathbf{P}^{(r)}, \mathbf{p}) = \left(\sum_{i=1}^N \left(\frac{p_i^{1-r}}{\sum_{j=1}^N p_j^{1-r}} \right) p_i^r \right)^{\frac{1}{r}} = \left(\sum_{i=1}^N p_i^{1-r} \right)^{\frac{-1}{r}} = P_{-r}(\mathbf{p}, \mathbf{p}). \quad (2.9)$$

The normalized probability of $1-r$ events as a weight has the effect of reversing the sign of power r with the original probabilities now the weights, as shown in the right most expression.

For the assessment of probabilistic forecasts, the coupled probabilities will not be necessary, as the probability of each test sample will be treated as equiprobable and the coupled probabilities are also equiprobable. Figure 2 which shows the weighted geometric mean for three different distributions is plotted in terms of $-r$ so the visual orientation of graphs is similar to those appearing later regarding assessment of probabilistic forecasts. The distributions examined are members of the coupled-Gaussians, which are equivalent to the Student's-t, but defined in terms of a nonlinear coupling term κ which is the inverse of the degree of freedom. For simplicity the coupled-Gaussians are expressed here in terms of the power of the generalized mean $r_D = \frac{-2\kappa}{1+\kappa}$. The subscript D distinguishes the parameter

of the distribution from the power of the mean. The numeral 2 relates to the squared term of a Gaussian and is 1 for the coupled exponential distribution. The coupled-Gaussian is

$$f(x) = \frac{1}{Z(r, \sigma)} \left(1 - \left(\frac{r_D}{2+r_D} \right) \frac{x^2}{\sigma^2} \right)_+^{\frac{1}{r_D}}, \quad r > -2 \quad (2.10)$$

where $(a)_+ \equiv \max(0, a)$, Z is the normalization of the distribution and σ is the scale parameter of the distribution. For $-2 < r_D < 0$ the distribution is heavy tail, $r_D = 0$ is the Gaussian, and $r_D > 0$ is a compact-support distribution. Applying the continuous form of equation (2.9) with the matching value of r gives the following result

$$\begin{aligned} f_{r_D}(f(x, r_D, \sigma)) &= \left(\int_{x \in X} f(x, r_D, \sigma)^{1-r_D} dx \right)^{\frac{-1}{r_D}} \\ &= Z(r, \sigma)^{\frac{1-r_D}{r_D}} \left(\int_{x \in X} \left(1 - \left(\frac{r_D}{2+r_D} \right) \frac{x^2}{\sigma^2} \right)_+^{\frac{1-r_D}{r_D}} dx \right)^{\frac{-1}{r_D}} \\ &= Z(r, \sigma)^{\frac{1-r_D}{r_D}} \left(Z(r, \sigma) \left(1 - \frac{r_D}{2+r_D} \right)^{-1} \right)^{\frac{-1}{r_D}} \\ &= \frac{1}{Z(r, \sigma)} \left(1 - \frac{r_D}{2+r_D} \right)^{\frac{1}{r_D}} = f(x = \sigma, r_D, \sigma) \end{aligned} \quad (2.11)$$

While not derived here, the equivalence between of the generalization maximum entropy principal using the Tsallis entropy and the minimization of the weighted generalized mean is such that the distribution $f(x, r_D, \sigma)$ is the minimization of f_{r_D} given the constraint that the scale is σ .

In Figure 2 the weighted generalized mean (wgm) is shown for the Gaussian distribution $r_D = 0$ and two examples of the Coupled Gaussian with $r_D = -2/3, 1$. For each of the distributions the scale is $\sigma = 1$. In order to illustrate the intersection between the distribution and its matching value of the wgm, the mean of each distribution is shifted by $\mu = r_D - \sigma$. The wgm is plotted as a function of $2r_D - r$ rather than r so that the increase in wgm is from left to right as it will be when evaluating probabilistic forecasts. The coupled exponential distribution and the coupled Gaussian distribution have the following relationship with respect to the

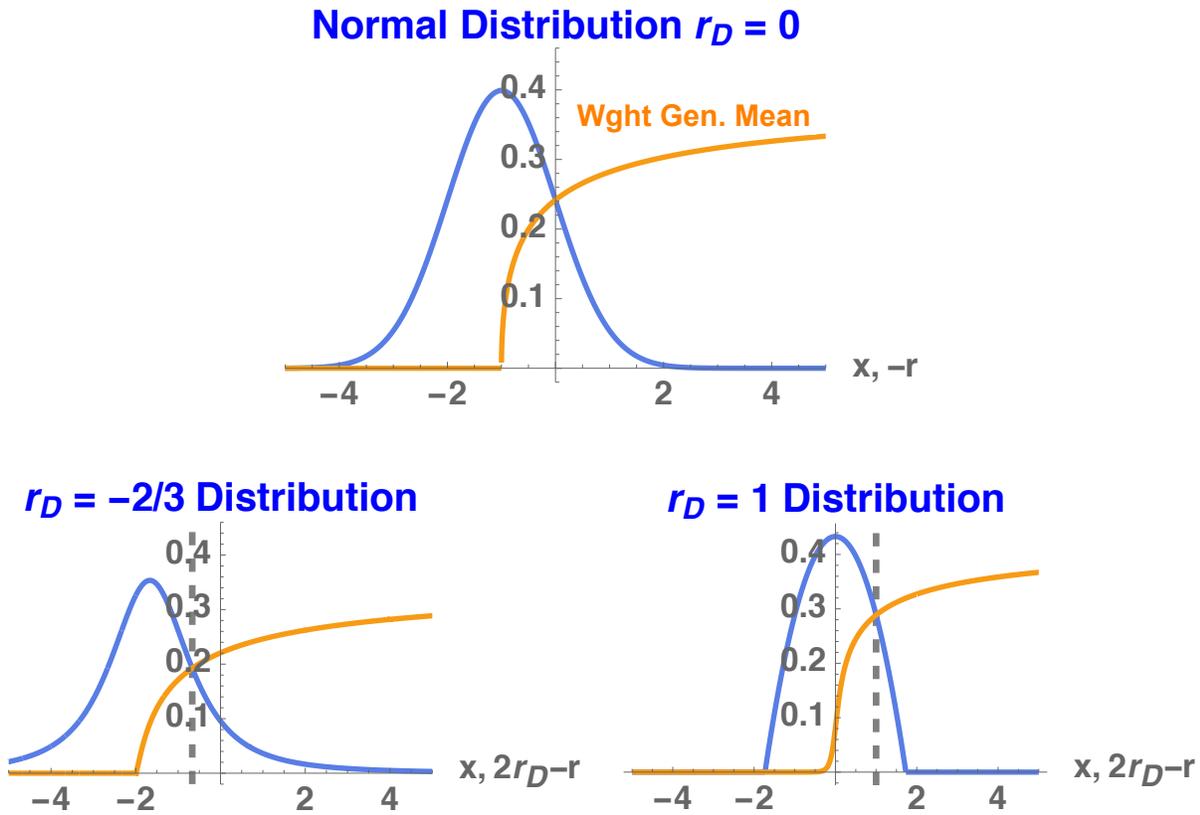


Figure 2 Plots of the weighted generalized mean (wgm) overlaid with the distribution which minimizes the wgm at the value $r = r_D$. The mean of each distribution is adjusted to show the wgm intersecting the density at the mean plus width parameter of the distribution. a) Normal distribution $N(-1,1)$ (blue) with its wgm (orange). The normal distribution is Coupled Gaussian with $r_D = 0$ and minimizes the wgm at $r = 0$ (weighted geometric mean). The wgm at $r = 0$ is equal to the density of the normal at the mean plus standard deviation. b) The Coupled Gaussian with $r_D = -2/3$, $\mu = -5/3$, $\sigma = 1$ minimizes the wgm at $r = -2/3$. The orientation of the wgm plot for b and c is inverted and shifted by $2r_D - r$. c) The Coupled Gaussian with $r_D = 1$, $\mu = 0$, $\sigma = 1$ minimizes the wgm at $r = 1$. For both b and c the wgm at r_D is equal to the density at the mean plus the generalized standard deviation.

generalized average uncertainty

$$\exp_r \left(\int_{\mu}^{\infty} \frac{1}{\sigma} \exp_r \left(-\frac{x-\mu}{\sigma} \right) \ln_r \left(\frac{1}{\sigma} \exp_r \left(-\frac{x-\mu}{\sigma} \right) \right) \right) = \frac{e_r^{-\left(\frac{\mu+\sigma-\mu}{\sigma}\right)}}{\sigma} = \frac{1}{\sigma e_r} \quad (2.12)$$

$$\exp_r \left(\int_{-\infty}^{\infty} \frac{1}{Z_r} e_r^{-\frac{(x-\mu)^2}{2\sigma^2}} \ln \left(\frac{1}{Z_r} e_r^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) \right) = \frac{1}{Z_r} e_r^{-\frac{(\mu+\sigma-\mu)^2}{2\sigma^2}} = \frac{e_r^{-1/2}}{Z_r}. \quad (2.13)$$

These relationships provide evidence of the importance of the generalized mean as expressions of the average uncertainty for non-exponential distributions. The approach also strengthens the connection with established principals of statistics, since the coupled exponential distribution is equivalent to the generalized Pareto distribution and the

coupled Gaussian is equivalent to the Student's t-distribution. In the next section use of the generalized mean as a metric to evaluate probabilistic inference is demonstrated.

4 Assessing probabilistic forecasts using a Risk Profile

The goal of an effective probabilistic forecast is to “reduce perplexity”; i.e. to enhance decision making by providing accurate information about the underlying uncertainties. Just as the maximum entropy approach is important in selecting a model which properly expresses the uncertainty, minimization of the cross-entropy between a model and a source of data is essential to accurate forecasting. In Section 3 the relationship between the weighted generalized mean of a distribution and the generalized entropy functions was established; likewise, the generalized cross-entropy can be translated into the weighted generalized mean in probability space. The result is a spectrum of metrics which modifies the sensitivity to surprising or low-probability events; as such it is referred to as a *Risk Profile*.

The most basic definition of risk R is the expected cost of a loss L times the probability of the loss

$$R = E(L) = \sum_{i=1}^N L_i p(L_i). \quad (2.14)$$

It can also be defined as the degree of variance or standard-deviation for a process, such as an asset price, which has a monetary or more general value. An individual or agent can have different perceptions of risk, expressed as the utility of a loss (or gain). Thus a risk-averse person would seek to lower exposure to high variances given the same expected loss. With regard to a probabilistic forecast, the cost is being surprised by an event which was forecasted to be low probability. While a particular application may also assign a valuation to events, with regard to evaluating the quality of the forecast the ‘surprisal – S ’ will be the only cost. A neutral perspective on the risk of being surprised is the information theoretic measure, the logarithm of the probabilities. The expected surprisal cost is the arithmetic average of the logarithmic distance between the forecasted probabilities and perfect forecasts

$$S = E[S_i] = -\frac{1}{N} \sum_{i=1}^N (\ln p_i + \ln 1) = -\sum_{i=1}^N \ln p_i. \quad (2.15)$$

This is known as the logarithmic scoring rule and has the property of being the only scoring rule which is both proper and local. A proper scoring rule is one in which optimization of the rule leads to unbiased forecasts relative to what is known by the forecaster. A local scoring rule is one in which only the probabilities of events which occurred are used in the evaluation.

The influence of risk-seeking and risk-aversion in forecasting can be evaluated using a generalized surprisal function

$$S_r = E[S_{r,i}] = -\frac{1}{N} \sum_{i=1}^N (\ln_r p_i + \ln_r 1) = -\sum_{i=1}^N \ln_r p_i \quad (2.16)$$

$$\ln_r x \equiv \frac{1+r}{r} (x^r - 1).$$

The generalized logarithmic function is fundamental to the generalization of thermodynamics introduced by Tsallis and its role for a generalized information theory is explained further in Appendix A. The generalized surprisal function is still a local scoring rule, but is no longer proper. Its properties as a biased function of risk have been studied in economics due to its preservation of a constant coefficient of relative risk. In economics the variable x of Equation (2.16) is the valuation and the relative risk aversion [17], [18] is defined in terms of $1-r$ since $r=1$ is a linear function and thus is considered to be neutral risk. Here the bias is with respect to the neutral measure of information, namely $\ln p$ when $r=0$. Thus for purposes of this discussion the relative risk sensitivity is defined as

$$r \equiv 1 + p \frac{d^2(\ln_r p) / dp^2}{d(\ln_r p) / dp}. \quad (2.17)$$

For negative values of r , the generalized surprisal is risk-averse, since the cost of being surprised by an event forecasted not to exist, i.e. $p=0$, goes to infinity faster. This is referred to as the domain of robust metrics, since it encourages algorithms to be conservative or robust in probabilistic estimation. For positive values of r , the measure is risk-seeking and is referred to as a decisive metric since it is more like the cost of making a decision over a finite set of choices, as opposed to the cost of properly forecasting the uncertainty of the decision.

For evaluating a probabilistic forecast use of the logarithmic or generalized logarithmic scale is needed to assure that the analysis properly measures the cost of a surprising forecast; nevertheless, it leaves obscure what is truly desired in an evaluation: knowledge of the central tendency and fluctuation of the forecasts. Following the procedures introduced in Sections 2 and 3, the generalized scoring rule can be translated to a probability by taking the inverse of the generalized logarithm, which is the generalized exponential

$$\exp_r(x) \equiv \left(1 + \frac{r}{1+r} x \right)^{\frac{1}{r}}. \quad (2.18)$$

Applying (2.18) to (2.16) shows that the generalized mean of the probabilities is the translation of the generalized logarithmic scoring rule to the probability domain

$$P_{r-avg}(\mathbf{p}) \equiv \exp_r(-S_r(\mathbf{p})) = \left(1 + \frac{r}{1+r} \left(\frac{1}{N} \sum_{i=1}^N \frac{1+r}{r} (p_i^r - 1) \right) \right)^{\frac{1}{r}} = \left(\frac{1}{N} \sum_{i=1}^N p_i^r \right)^{\frac{1}{r}}. \quad (2.19)$$

The generalized mean of the forecasted probabilities, thus forms a spectrum of metrics which profile the performance of the forecast relative to the degree of risk aversion. The spectrum is thus referred to as a *Risk Profile* of the probabilistic forecast.

To illustrate the utility of the Risk Profile in evaluating statistical models, the contrast between robust and decisive models of a multivariate Gaussian random variable is evaluated. The origin of the Student's t-distribution was the insight by William Gosset [19] that limited sample size from a source known to have a Gaussian distribution requires a model which modifies the Gaussian distribution to have a slower than exponential decay. Again using the equivalent coupled-Gaussian distribution defined in Equation (2.10) but now accounting for the dimensions of the distribution, $r_D = \frac{-2\kappa}{1+d\kappa}$ where D refers to the distribution and d is the dimensions. The full expression for the multivariate coupled-Gaussian is

$$G_r(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{Z_r(\boldsymbol{\Sigma})} \left(1 - \frac{r_D}{2+r_D} (\mathbf{x} - \boldsymbol{\mu})^\top \cdot \boldsymbol{\Sigma}^{-1} \cdot (\mathbf{x} - \boldsymbol{\mu}) \right)_+^{\frac{1}{r_D}} \quad (2.20)$$

where the vectors \mathbf{x} and $\boldsymbol{\mu}$ are the random variable and mean, $\boldsymbol{\Sigma}$ is the correlation matrix³, Z_r is the normalization, and the symbol $()_+$ used for the compact-support domain indicates that negative values are truncated to zero.

The problems with trying to model a Gaussian random variable using a multivariate Gaussian as the model is shown in Figure 3. In this example 10 independent features which are generated from Gaussian distributions are modelled as a multivariate Gaussian with a varying number of dimensions. Although reasonable classification performance is achieved (84%), the accuracy of the modelled uncertainty is reduced beyond 6 dimensions. Furthermore, the robustness as measured by the -2/3 generalized mean drops to zero when all 10 dimensions are modelled.

³ While $\boldsymbol{\Sigma}$ is the covariance matrix for a Gaussian distribution, for the coupled-Gaussian this parameter is a generalization of the covariance and in association with the Student's t distribution is known as the correlation matrix.

Even without seeking to optimize the coupling value, improvement in the accuracy and robustness of the multivariate model can be achieved using heavy-tail decay. Figure 4

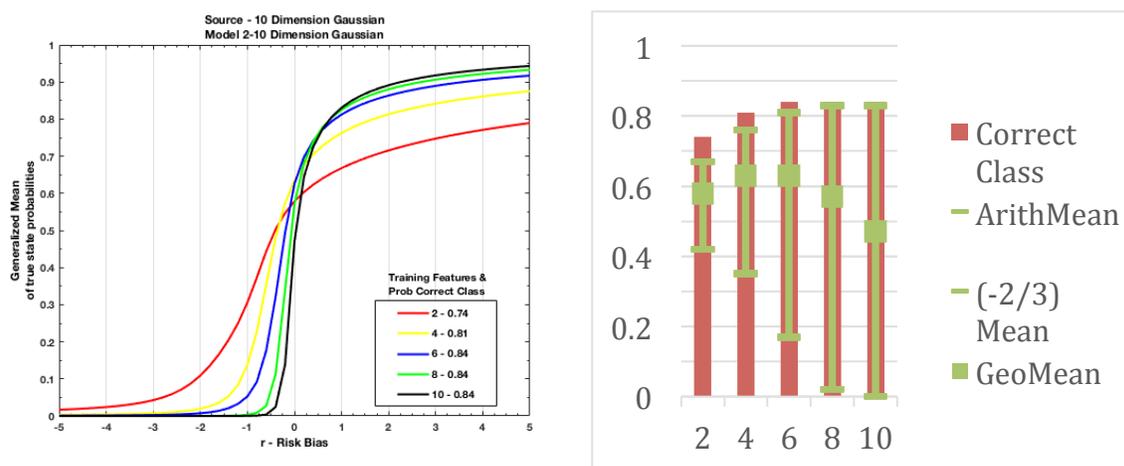


Figure 3 A source of 10 independent Gaussian random variables is overfit given 25 samples to learn the mean and variance of each dimension and a model which is also a multivariate Gaussian. a) The risk profile shows that as the number of dimensions increases the model becomes more decisive. b) At 6 dimensions, the classification performance saturates to 84% at and the accuracy of the probabilities reaches is maximum of 63%.

shows an example with $r_D = -0.15$ in which the accuracy is improved to 0.69 and is stable for dimensions 6-10. The robustness continues to decrease as the number of dimensions increases, but is improved significantly over the multivariate Gaussian model. The classification improves modestly to 86%, but is not the principal reason for using the heavy-tail model.

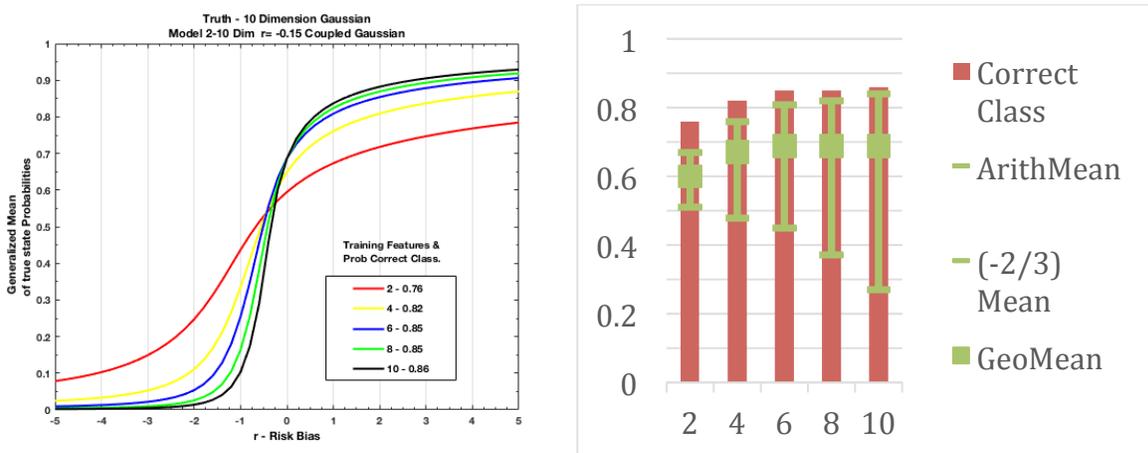


Figure 4 The risk of overfitting is reduced by using a heavy-tail coupled-Gaussian. Shown is an example with $r = -0.15$. a) The risk profile shows that the accuracy of 0.69 continues to hold as the dimensions modeled is increased from 6 to 8. b) The percent correct classification (red bar) improves to 86% with 10 dimensions modeled. The robustness does go down as the number of dimensions is increased, but could be improved by optimizing the coupling value used.

The problems with overconfidence in the tails of a model are very visible when a compact-support distribution is used to model a source of data which is Gaussian. In this

case, the reporting of $P=0$ for states which do occur results in the accuracy being zero. An example of this situation is shown in Figure 5 in which the distribution power is $r_d = 0.6$. Although the model is neither accurate or robust in the reporting of uncertainty, it is still capable of modest classification performance (75% for 4-dimensions and reduced to 67% for 10-dimensions); nevertheless, characterization of only the classification performance would not show the severity of the problem with inappropriately using a compact-support model.

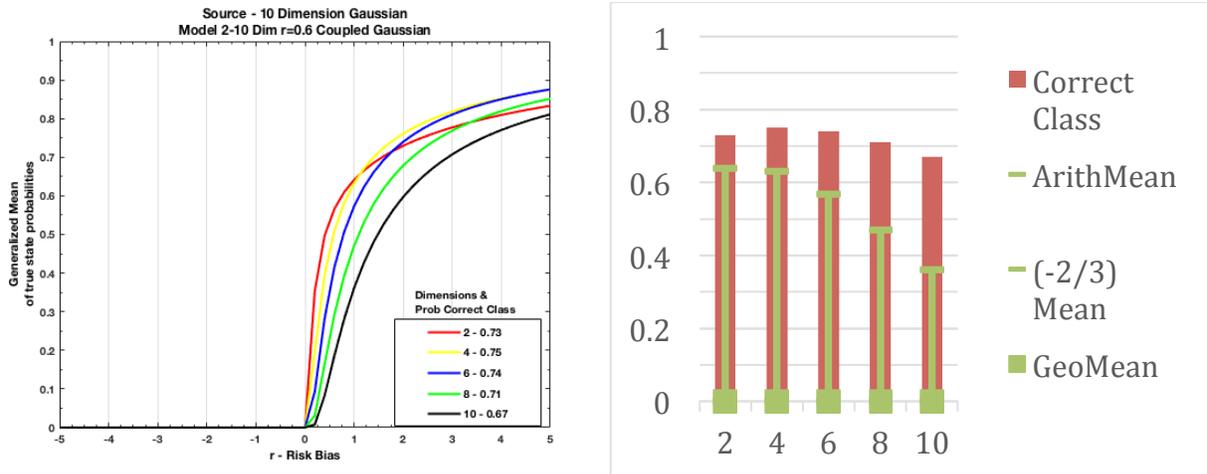


Figure 5 Using a compact-support distribution to model a source of data which is Gaussian results in the probability accuracy being zero. a) The risk profile shows that the model using $r=0.6$ is neither accurate nor robust. b) The classification performance (red bar) is only 75% for the model with four dimensions, but characterization of only the classification performance would not show the severity of the problem with this model.

5 Conclusion

The purpose of this discussion has been to show that translating results of information theory from the entropy domain to the probability domain can simplify and clarify interpretation of important information metrics. In particular, the basic fact that the entropy of a Gaussian distribution when translated to a density as shown in Equation (2.4) is equal to the density of the Gaussian at the mean plus the standard deviation should be a widely understood representation of the relationship between the standard deviation and entropy. Unfortunately, while entropy provides the convenience of an additive information measure, the connection to the underlying probabilities of a distribution is often lost.

There are two areas where the lack of clear intuition between probabilities and their central tendency as expressed by the entropy function has caused significant confusion and misunderstanding. When analyzing the statistical characteristics of complex systems there has been recognition of a need to separate the role of nonlinearity and the measure of central tendency in the uncertainty, but a diversity of perspectives on how to accomplish this. Two important candidates for a generalized entropy (Rényi and Tsallis) are both unified by the weighted generalized mean upon translation to the probability domain. From the perspective of the probability domain, the power of the generalized mean is seen

to account for potentially non-exponential decay in the tail of a distribution. When the power of the generalized mean is matched with the tail decay, then the generalized mean represents a modified expression of the central tendency of the distribution. This is illustrated for the coupled-Gaussians (equivalently the Student's t) in Equation (2.11) and Figure 2.

The second area of confusion and a focus of this discussion is the role of information theory in evaluating the performance of probabilistic forecasts. In most domains of science the 'average' is simply the arithmetic mean of a random variable. The median may be used as a robust alternative to the arithmetic mean to filter the affects of outliers, but nevertheless the connection between the arithmetic mean and the central tendency is understood to be universal. Unfortunately, this is not the case for numbers which are formed by ratios, of which probabilities are a particularly important example. An elementary principal of probability theory is that the total probability of a set of independent probabilities is their product. So why isn't n th root of the total probability or the geometric mean of the independent set of probabilities also recognized and taught to be the average? Or if we take the probabilities to be part of distribution the weighted geometric mean, with the weight set by the probabilities? The answer seems to be both the misconception that the arithmetic mean is always the central tendency and the role that entropy serves in translating probabilities to a domain in which the arithmetic mean is indeed the central tendency.

For the evaluation of probabilistic forecasts, this has created a serious problem, in which a variety of different 'scoring rules' have been developed with the assumption that evaluators are free to choose a metric depending on there own conception of the utility of a forecast. Given this perspective, the log scoring rule, which derives from probability and information theory, has often been rejected as over penalizing because a forecast near zero for an event which occurs tends toward a cost of infinity. In its place, the Brier Score or mean-square average has become a popular alternative despite its lack of rigorous connection to probability and information theory. Examining these issues from the perspective of the generalized mean of the forecasted probabilities provides some needed clarity. The geometric mean is an unbiased measure of the central tendency of the forecast. Other powers are biased and can be associated with a degree of risk aversion. The arithmetic mean is biased toward decisive forecasting. When the bias of the arithmetic mean is compensated for via inclusion of the non-occurring forecasts, the mean-square average is derived. Thus the arithmetic mean or mean-square average should not be used in isolation to evaluate a forecast. Together with a conjugate mean with a negative power which provides a measure of robustness, the arithmetic mean and its conjugate can provide a measure of the degree of fluctuation in the forecast about the central tendency measured by the geometric mean.

Identification of a method for assessing probabilistic forecasts on the probability scale opens up other possibilities for integrating analysis with visual representations of

performance. Recently, it was shown [20] that a calibration curve comparing reported probabilities and the measured distribution of the test samples can be overlaid with metrics using the generalized mean of the reported and measured probabilities. This approach provides insight regarding the sources of uncertainty in a forecast from insufficient models and insufficient features. As the utility of measuring the generalized mean of a set of probabilities is explored, further innovations can be developed for robust accurate probabilistic forecasting. These are particularly important for the development of machine learning and artificial intelligence applications which need to carefully manage uncertain and potentially risky decisions.

Appendix

A Modeling risk as a coupling of statistical states

The text focuses on the role of risk sensitivity r in evaluating the performance of probabilistic forecasts. The model for risk sensitivity derives from a model of complex statistical systems influenced by nonlinear coupling between the states of a system. The nonlinearity κ of a complex systems increases the uncertainty about the long-range dynamics of the system. In [11] the effect of nonlinearity, such as multiplicative noise or variation in the variance, was shown to result in a modification from the exponential family

$f(x) \propto e^{-x^\alpha}$ to the power-law domain with $\lim_{x \rightarrow \infty} f(x) \propto x^{-\frac{\alpha}{r}}$, where the risk sensitivity can be decomposed into the nonlinear coupling and the power and dimension of the argument

$r(\kappa, \alpha, d) = \frac{-\alpha\kappa}{1+d\kappa}$. As the source of coupling κ increases from zero to infinity, the increased nonlinearity results in increasingly slow decay of the tails of the resulting distributions. Negative coupling can also be modeled, resulting in compact-support domain distributions with less variation than the exponential family. The negative domain, which models compact-support distributions, is $-\frac{1}{d} < \kappa < 0$.

The expression for r is also known within the field of nonextensive statistical mechanics as a dual transformation between the heavy-tail and compact support domains.

With the alpha term dropping out the dual has the following relationship $\hat{\kappa} \Leftrightarrow \frac{-\kappa}{1+d\kappa}$. The

dual is used to determine the conjugate to the decisive risk bias of 1. Taking $\alpha = 2$ and

$d = 1$, the coupling for a risk bias of one is $1 = \frac{-2\kappa}{1+\kappa} \Rightarrow \kappa = -\frac{1}{3}$ and the conjugate values are

$$\hat{\kappa} = \frac{\frac{1}{3}}{1-\frac{1}{3}} = \frac{1}{2} \text{ and } \hat{r} = \frac{-2\left(\frac{1}{2}\right)}{1+\frac{1}{2}} = -\frac{2}{3}.$$

The risk bias is closely related to the Tsallis entropy parameter $q=1-r$ [21]–[23]. One of motivating principals of the Tsallis entropy methods was to examine how power law systems could be modeled using probabilities raised to a power p_i^q [24]. As such, q can be thought of as the number of random variables needed to properly formulate the statistics of a complex system, while r represents the deviation from a linear system governed by exponential statistics. When the deformed probabilities are renormalized the resulting distribution can be shown to also represent the probability of a “coupled state” of the system

$$P_i^r = \frac{p_i^{1-r}}{\sum_{j=1}^n p_j^{1-r}} = \frac{p_i \prod_{\substack{k=1 \\ k \neq i}}^n p_k^r}{\sum_{j=1}^n \left(p_j \prod_{\substack{k=1 \\ k \neq j}}^n p_k^r \right)}, \quad (2.21)$$

hence use of the phrase “nonlinear statistical coupling”.

Just as the probabilities are deformed via a multiplicative coupling, the entropy function is deformed via an additive nonlinear coupling term. The non-additivity of the generalized entropy H_κ provides a definition for the degree of nonlinear coupling (neglecting the power α and dimensions d of the variable for the moment). The joint coupled-entropy of two independent systems A and B includes a nonlinear term

$$H_\kappa(A, B) = H_\kappa(A) + H_\kappa(B) + \kappa H_\kappa(A) H_\kappa(B). \quad (2.22)$$

For $\kappa=0$ the additive property of entropy is satisfied by the logarithm of the probabilities. The function which satisfies the nonlinear properties of the generalized entropy is a generalization of the logarithm function referred to as the coupled logarithm

$$\ln_\kappa x \equiv \frac{1}{\kappa} \left(x^{\frac{\kappa}{1+\kappa}} - 1 \right), \quad x > 0 \quad (2.23)$$

In the limit when κ goes to zero the function converges to the natural logarithm. This definition of the generalized logarithm has the property that $\int_0^1 \ln_\kappa p^{-1} dp = 1$, thus the deformation modifies the relative information of a particular probability while preserving the ‘total’ information across the domain of probabilities.

The inverse of this function is the coupled exponential

$$\exp_\kappa x \equiv (1 + \kappa x)^{\frac{1+\kappa}{\kappa}}. \quad (2.24)$$

A distribution of the exponential family will typically include an argument of the form $-x^\alpha/\alpha$ which is generalized by the relationship $\left(\exp_\kappa x^\alpha \right)^{\frac{-1}{\alpha}} = \exp_\kappa^{-1/\alpha} x^\alpha = (1 + \kappa x^\alpha)^{-\frac{1+\kappa}{\alpha\kappa}}$. The rate of

decay for a d -dimensional distribution is accounted for by $\exp_{\kappa,d}^{-1/\alpha} x^\alpha = (1 + \kappa x^\alpha)^{\frac{1+d\kappa}{-\alpha\kappa}}$, neglecting the specifics of the matrix argument. This is the form of the multivariate Student's t distribution with κ equal to the inverse of the degree of freedom. When the generalized logarithm needs to include the role of the power and dimension this is

expressed as $\ln_{\kappa,d} x^{-\alpha} \equiv \frac{1}{\kappa} \left(x^{\frac{-\alpha\kappa}{1+d\kappa}} - 1 \right)$ or alternatively $\left(\ln_{\kappa,d} x^{-\alpha} \right)^{\frac{1}{\alpha}} = \left(\frac{1}{\kappa} \left(x^{\frac{-\alpha\kappa}{1+d\kappa}} - 1 \right) \right)^{\frac{1}{\alpha}}$. The

first expression is used here, though research regarding the later expression has been explored.

There are a variety of expressions for a generalized entropy function which when translated back to the probability domain lead to the generalized mean of a probability distribution. Generalization of the entropy function can be viewed broadly as a modification of the logarithm function and the weight forming the arithmetic mean. The translation back to the probability domain makes use of the inverse of the generalized logarithm, namely the generalized exponential. While a more general express is possible, here will make use of the generalizations defined in the previous paragraph. The general expression for aggregating probabilities is then

$$\begin{aligned}
 P_\kappa(\mathbf{w}, \mathbf{p}; \alpha, d) &= \exp_{\kappa,d}^{-1/\alpha} \left(H_\kappa(\mathbf{p}; \mathbf{w}, \alpha, d) \right) \\
 &= \exp_{\kappa,d}^{-1/\alpha} \left(\sum_{i=1}^N w_i \ln_{\kappa,d} p_i^{-\alpha} \right) \\
 &= \left(1 + \kappa \left(\sum_{i=1}^N \frac{w_i}{\kappa} \left(p_i^{\frac{-\alpha\kappa}{1+d\kappa}} - 1 \right) \right) \right)^{\frac{1+d\kappa}{-\alpha\kappa}} \\
 &= \left(\sum_{i=1}^N w_i p_i^{\frac{-\alpha\kappa}{1+d\kappa}} \right)^{\frac{1+d\kappa}{-\alpha\kappa}}
 \end{aligned} \tag{2.25}$$

where the weights w_i are assumed to sum to one. In the main text the focus is placed on

the risk bias $r = \frac{-\alpha\kappa}{1+d\kappa}$, which forms the power of the generalized mean. The coupled entropy function is defined using the coupled probability (2.21) for the weights. Other generalized entropy functions use different definitions for the weights and generalized logarithm, but as proven in [11] for at least the normalized Tsallis entropy, Tsallis entropy and Rényi entropy they all converge to an expression of the form

$$\begin{aligned}
P_{\kappa}(\mathbf{p};\alpha,d) &= \left(\sum_{i=1}^N \left(\frac{p_i^{1+\frac{\alpha\kappa}{1+d\kappa}}}{\sum_{j=1}^N p_j^{1+\frac{\alpha\kappa}{1+d\kappa}}} \right)^{\frac{-\alpha\kappa}{1+d\kappa}} \right)^{\frac{1+d\kappa}{-\alpha\kappa}} \\
&= \left(\frac{\sum_{i=1}^N p_i}{\sum_{j=1}^N p_j^{1+\frac{\alpha\kappa}{1+d\kappa}}} \right)^{\frac{1+d\kappa}{-\alpha\kappa}} \\
&= \left(\sum_{i=1}^N p_i^{1+\frac{\alpha\kappa}{1+d\kappa}} \right)^{\frac{1+d\kappa}{\alpha\kappa}} .
\end{aligned} \tag{2.26}$$

This express for the central tendency of a set of probabilities, assumes that the probabilities form a distribution which sums to one.

The assessment of a probabilistic forecast treats each test sample as an independent equally likely event. The weights, even using the coupled probability, simplify to one over the number of test samples

$$w_i = \frac{\left(\frac{1}{N}\right)^{1+\frac{\alpha\kappa}{1+d\kappa}}}{\sum_{i=1}^N \left(\frac{1}{N}\right)^{1+\frac{\alpha\kappa}{1+d\kappa}}} = \frac{\left(\frac{1}{N}\right)^{1+\frac{\alpha\kappa}{1+d\kappa}}}{N \left(\frac{1}{N}\right)^{1+\frac{\alpha\kappa}{1+d\kappa}}} = \frac{1}{N} \tag{2.27}$$

Thus the generalized mean used for the Risk Profile has a power with the opposite sign of that used for the average probability of distribution

$$P_{\kappa}(\mathbf{p};\alpha,d) = \left(\frac{1}{N} \sum_{i=1}^N p_i^{1+d\kappa} \right)^{\frac{1+d\kappa}{-\alpha\kappa}} \tag{2.28}$$

where the probabilities in this express are not from a distribution, but rather are test samples of forecasted of events which have occurred.

Acknowledgement

The initial research developing the *Risk Profile* was sponsored by Raytheon IRAD IDS202 2010-2012. Conversations with Ethan Phelps and Herb Landau highlighted the need for the surprisal metric in the assessment of discrimination algorithms.

References

- [1] C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication; Univ. of Illinois," *Urbana*, p. 114, 1949.
- [2] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
- [3] I. J. Good, "Rational decisions," *J. R. Stat. Soc. Ser. B*, pp. 107–114, 1952.
- [4] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007.
- [5] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, no. XIV. Springer Publishing Company, Incorporated, 2010.
- [6] C. Tsallis, *Introduction to nonextensive statistical mechanics: Approaching a complex world*. Springer New York, 2009.
- [7] K. P. Nelson, B. J. Scannell, and H. Landau, "A Risk Profile for Information Fusion Algorithms," *Entropy*, vol. 13, no. 8, pp. 1518–1532, 2011.
- [8] D. McAlister, "The law of the geometric mean," *Proc. R. Soc. London*, vol. 29, no. 196–199, pp. 367–376, 1879.
- [9] A. I. Khinchin, *Mathematical Foundations of Statistical Mechanics*. Courier Corporation, 1949.
- [10] A. I. Khinchin, *Mathematical Foundations of Information Theory*. Courier Corporation, 1957.
- [11] K. P. Nelson, S. R. Umarov, and M. A. Kon, "On the average uncertainty for systems with nonlinear coupling," *Phys. A Stat. Mech. its Appl.*, vol. 468, no. 15 Feb, pp. 30–43, 2017.
- [12] E. T. Jaynes, *Probability theory: the logic of science*. Cambridge, UK: Cambridge University Press, 2003.
- [13] E. T. Jaynes, "The relation of Bayesian and maximum entropy methods BT - Maximum-entropy and Bayesian methods in science and engineering," Springer, 1988, pp. 25–29.
- [14] A. Rényi, "On measures of entropy and information," *Fourth Berkeley Symp. Math. Stat. Probab.*, vol. 1, pp. 547–561, 1961.
- [15] C. Tsallis, *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World*. Springer Verlag, 2009.
- [16] S. I. Amari and H. Nagaoka, *Methods of Information Geometry*, vol. 191. Oxford University Press, 2000.
- [17] "Risk Aversion," *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Risk_aversion. [Accessed: 17-Sep-2017].
- [18] C. P. Simon and L. Blume, *Mathematics for economists*. Vinod Vasishtha for Viva Books, 2006.

- [19] W. Gossett, "The Application of the 'Law of Error' to the Work of the Brewery," Dublin, 1904.
- [20] K. P. Nelson, "Assessing Probabilistic Inference by Comparing the Generalized Mean of the Model and Source Probabilities," *Entropy*, vol. 19, no. 6, p. 286, Jun. 2017.
- [21] K. P. Nelson and S. Umarov, "Nonlinear statistical coupling," *Phys. A Stat. Mech. its Appl.*, vol. 389, no. 11, pp. 2157–2163, 2010.
- [22] Q. A. Wang, "Probability distribution and entropy as a measure of uncertainty," *J. Phys. A Math. Theor.*, vol. 41, p. 65004, 2008.
- [23] A. Souto Martinez, R. Silva González, and A. Lauri Espíndola, "Generalized exponential function and discrete growth models," *Phys. A Stat. Mech. its Appl.*, vol. 388, no. 14, pp. 2922–2930, 2009.
- [24] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *J. Stat. Phys.*, vol. 52, no. 1, pp. 479–487, 1988.