Introduction to the Gaussian Information Criterion

Russell Leidich
pkejjy@gmail.com

June 7, 2021

0. Abstract

There are many applications involving physical measurements which are expected to result in a probability density function (PDF) which is asymptotically Gaussian (normal) [0] or lognormal [1]. In the latter case, we can simply take the logs of the (positive) samples in order to obtain the former, so the math in this paper will focus exclusively on Gaussians.

For example, we would expect the distribution of radio power received at a dish to be lognormally distributed, given a sufficiently broad swath of sky to observe for a sufficiently long duration, and in the relative absence of terrestrial radio interference. However, if we were then to focus on a particular star system, the observed "experimental" PDF could substantially deviate from that "background" PDF. It might not even be lognormal if, for example, the star exhibits peaks in radio power at a few distinct frequencies.

It would therefore be useful to have a means to quantify the "surprise" factor of experimental PDFs relative to an established background PDF which is known to be, or be equivalent to, a Gaussian. If a given experimental PDF where also known to be Gaussian, then we could do this by employing the Kullback-Leibler (KL) divergence [2] from one to the other, as Gupta [3] appears to have done for the multidimensional case.

When the experimental PDF is *not* known to be Gaussian (or any PDF archetype, for that matter), the situation is more complicated, mainly because we are forced to deal with a real-valued set of samples ordered by increasing positivity -- a 1D point cloud, to be precise, although "vector" will suffice for brevity -- rather than an analytic function. Ranking the information cost of encoding such a vector, versus others arising from other experiments, under the prior assumption of the same background PDF, is the subject of this paper. We also investigate the question of ascertaining which background PDF is the most useful for the sake of discriminating anomalous from mundane experimental PDFs.

1. Scalar Information with Respect to a Gaussian

Given a Gaussian [2] with mean *μ* and standard deviation *σ*. Its PDF is then given by

$$p(x) \equiv \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where *p*(*x*) is probability *density* -- not probability as such. It is therefore not meaningful to express the probability that this PDF would produce a given real number *x* simply because there are an infinite number of reals from which to choose. But let us consider the literal probability *P*(*x*) in the limit that a histogram slice approaches zero width:

$$P(x) \equiv \lim_{\Delta x \to 0} p(x)\Delta x$$

which is just a verbose expression for zero. By extension, though, it *is* meaningful to express the *ratio* of the probability that *p*(*x*) would output *x* versus *y*:

$$\frac{P(x)}{P(y)} \equiv \lim_{\Delta x \to 0} \frac{p(x)\Delta x}{p(y)\Delta x}$$

$$\frac{P(x)}{P(y)} \equiv \frac{p(x)}{p(y)}$$

Similarly, it is not meaningful to express how much information is required to encode *x*, given *p*(*x*). It is, however, meaningful to express the *change δ'*(*y*, *x*) in encoding cost incurred by replacing *y* with *x*. We can do so in units of nats (natural logs, which are bits times (ln 2)):

$$\delta'(y, x) \equiv - \ ln \frac{p(x)}{p(y)}$$

$$\delta'(y, x) \equiv \ ln\, p(y) - \ ln\, p(x)$$

The above definition (watch the signs!), which emerges from Shannon entropy [4], holds for *all* PDFs *p*(*x*) -- not just Gaussians. But in the latter case, we have

$$- \ ln\, p(x) \equiv - \ ln \frac{1}{\sigma\sqrt{2\pi}} + \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

$$- \ ln\, p(x) \equiv \frac{1}{2}\left[ln\, \sigma^2 + \ ln\, 2\pi + \frac{(x-\mu)^2}{\sigma^2}\right]$$

which is the relative information cost of encoding a real sample *x* with a Gaussian prior, which is only meaningful for the sake of comparison among samples with the same prior. (To be pedantic, this relative cost is only asymptotically accurate; it deviates due to overhead such as length information and digitization. But this is a reasonable approximation for practical purposes involving statistically significant numbers of samples.)

2. Vector Information with Respect to a Gaussian

The ratio of the probability that *K* samples of *p*(*x*) would include all *K* components of a vector ***x*** = (x_0, x_1... x_{K-1}) versus those of a vector ***y*** = (y_0, y_1... y_{K-1}) would just be the ratio of products of scalar probabilities:

$$\frac{P(\vec{x})}{P(\vec{y})} \equiv \lim_{\Delta x \to 0} \frac{K!\left(\Delta x^{K}\right) \prod_{k=0}^{K-1} p\left(x_k\right)}{K!\left(\Delta x^{K}\right) \prod_{k=0}^{K-1} p\left(y_k\right)}$$

where the factorial of *K* is the number of ways (permutations) in which the components of each vector could have been chosen. (This is wrong if there exists a pair of equal components in either vector, but this consideration is generally negligible in practice on account of its rarity and minor impact. Moreover, even if 2 floating-point measurements turned out to be equal, it would be implausible that they resulted from the same analog quantity. Therefore we assume that the permutation terms cancel, leaving just the product series intact.) Thus

$$\frac{P(\vec{x})}{P(\vec{y})} \equiv \prod_{k=0}^{K-1} \frac{p\left(x_k\right)}{p\left(y_k\right)}$$

Thinking again of information, if we now we define some *g'*(***x***) in terms of *p*(*x*) as follows:

$$g'(\vec{x}) \equiv - \sum_{k=0}^{K-1} ln\, p\left(x_k\right)$$

then the information cost change *Δ'*(***y***, ***x***) due to the replacement of ***y*** with ***x***, given *p*(*x*), can be expressed as follows:

$$\Delta'(\vec{y}, \vec{x}) \equiv g'(\vec{x}) - g'(\vec{y})$$
$$\Delta'(\vec{y}, \vec{x}) \equiv \sum_{k=0}^{K-1} \left[ln\, p\left(y_k\right) - ln\, p\left(x_k\right)\right]$$

and furthermore, the identity second above is valid even if the number of components of ***x*** and ***y*** differ because *g'*(***x***) increases monotonically with the number of samples, as one would expect of an encoding cost.

Now let us define the "Gaussian information criterion" (GIC) *g*(***x***) as the Gaussian case of *g*'(***x***):

$$g(\vec{x}) \equiv \sum_{k=0}^{K-1} \frac{1}{2}\left[ln\, \sigma^2 + ln\, 2\pi + \frac{\left(x_k - \mu\right)^2}{\sigma^2}\right]$$
$$g(\vec{x}) \equiv \frac{1}{2}\left[K\left(ln\, \sigma^2 + ln\, 2\pi\right) + \frac{1}{\sigma^2} \sum_{k=0}^{K-1} \left(x_k - \mu\right)^2\right]$$

which is thus the relative information cost of encoding (discrete PDF) ***x*** given a Gaussian prior with mean μ and variance σ^2. (Note that this particular formulation spares us the trouble of computing a square root in order to obtain σ. And we retain the (ln 2π) bias for reasons that will become apparent in Section 5.) "Relative" implies that this quantity only has meaning in the context of the information cost change Δ(***y***, ***x***) due to the replacement of ***y*** with ***x*** under the assumption of a Gaussian background PDF. -- the Gaussian case of Δ'(***y***, ***x***). Another way to conceptualize the GIC is as the negative log of the coefficient of the infinitesimal quantity $((K!)(\Delta x^K))$ under the assumption that $p(x)$ is Gaussian, which is meaningful in an *absolute* sense and supports its utility as a yardstick of information cost.

Now we could expand $g(x)$ as follows:

$$g\left(\vec{x}\right) \equiv \frac{1}{2}\left[K\left(ln\, \sigma^2 + ln\, 2\pi\right) + \frac{1}{\sigma^2}\sum_{k=0}^{K-1}\left(x_k^{\ 2} + \mu^2 - 2\mu x_k\right)\right]$$

$$g\left(\vec{x}\right) \equiv \frac{1}{2}\left[K\left(ln\, \sigma^2 + ln\, 2\pi\right) + \frac{1}{\sigma^2}\left(\sum_{k=0}^{K-1} x_k^{\ 2} - 2\mu\sum_{k=0}^{K-1} x_k + K\mu^2\right)\right]$$

But the components of ***x*** have their *own* mean μ_x and variance $\sigma_x^{\ 2}$:

$$\mu_x \equiv \frac{1}{K}\sum_{k=0}^{K-1} x_k$$

$$\sigma_x^{\ 2} \equiv \frac{1}{K}\sum_{k=0}^{K-1}\left(x_k - \mu_x\right)^2$$

so that

$$\sum_{k=0}^{K-1} x_k \equiv K\mu_x$$

$$\sum_{k=0}^{K-1} x_k^{\ 2} \equiv K\left(\sigma_x^{\ 2} + \mu_x^{\ 2}\right)$$

which means that we can produce the following alternative GIC formulation in terms of the mean μ_x and variance σ_x^2 of the components of ***x***, and with respect to some Gaussian $N(\mu, \sigma)$, simply by substituting the above expressions into the expansion of g(***x***) above:

$$G\left(\mu_x, \sigma_x\right) \equiv \frac{1}{2}\left[K\left(ln\, \sigma^2 + ln\, 2\pi\right) + \frac{1}{\sigma^2}\left(K\left[\sigma_x^{\ 2} + \mu_x^{\ 2}\right] - 2K\mu\mu_x + K\mu^2\right)\right]$$

$$G\left(\mu_x, \sigma_x\right) \equiv \frac{K}{2}\left[ln\, \sigma^2 + ln\, 2\pi + \frac{1}{\sigma^2}\left(\sigma_x^{\ 2} + \left[\mu_x - \mu\right]^2\right)\right]$$

which is rather surprising considering that we never made the assumption that the components of ***x*** were chosen from a Gaussian!

The upshot of all this is that we need only compute 4 real numbers and remember one natural number (*K*, which is often globally constant) -- effectively, just a hash of discarded sample points -- to compute the GIC from a Gaussian to an arbitrary experimental PDF composed of an arbitrary number of samples. For that matter, given the symmetry of its alternative formulation, we could just as well compute the GIC in the reverse direction, under the assumption that the experimental PDF were Gaussian but not necessarily the background PDF.

3. Background PDF Selection

There is also the question of selecting the most accurate background PDF. In reality, of course, such putatively pure Gaussians are polluted with noise such as radio interference, so there is a motivation to find the one which is most useful for the purpose of signal detection among the *experimental* PDFs. Informally, this would likely be the "most Gaussian" (or "most lognormal") of background PDF candidates.

Determining this entails discovering the background PDF which maximizes the GIC "entropy contrast" over a set of experimental PDFs. In turn, determining entropy contrast entails finding the ratio of the maximum to the minimum of a given information measure applied to each item in a dataset -- in this case, the GIC from a candidate background PDF to each experimental PDF. The background PDF which maximizes said ratio is empirically the most useful one for the sake of anomaly discovery, and perhaps the most accurate approximation of ground truth in the absence of interference. (Intuitively, entropy contrast should be a *difference* rather than a ratio, as the former is already exponential in probability. However, due to noise in the data and inevitable inaccuracies in all information measures, it is common to observe large but spurious such differences due to these reasons rather than the presence or absence of a signal. Therefore we consider entropy contrast to be the *ratio* of the extrema of a given information measure, in the same sense that nits of contrast are the ratio of the brightest white to the darkest black on a computer display, which thus worsens (shrinks) as the brightness level increases due to pixel noise.)

4. Gaussian Model Selection

We use the word "criterion" in reference to model selection. In this sense, the GIC is analogous to but less generic than the Akaike and Bayesian information criteria. In particular, the minimum GIC identifies the Gaussian maximum likelihood estimate (that is, the "parameter set" (μ, σ) corresponding to the Gaussian $N(\mu, \sigma)$ which is most likely to have been the hidden "generator" which produced $\boldsymbol{x}$).

By the way, "likelihood" just means "retrocausal probability". And machine learning parlance notwithstanding, maximum likelihood *estimators* do not *estimate* anything; they *identify* the most likely of a set of parameter sets (in this case, an infinite set of parameter sets of the form (μ, σ)), any of which may have given rise to the observed states (in this case, the components of $\boldsymbol{x}$). In other words, we cannot say that "$\boldsymbol{x}$ was produced by function $\boldsymbol{F}$ with a parameter set that we estimate to be $\boldsymbol{y}$" but rather merely "$\boldsymbol{x}$ is more likely to have been produced by function $\boldsymbol{F}$ with parameter set $\boldsymbol{y}$ than any other individual parameter set". (Pedantically, the "parameter set" (*μ, σ*) is a *vector* rather than a *set* because the meanings of its components depend upon their respective positions.)

Now in order to find the parameter set (*μ, σ*), we must minimize the relative information cost *g*(***x***) incurred by instantiating ***x*** in exactly *K* samplings of *N*(*μ, σ*). Because *g*(***x***) is analytic for all ***x*** in $\boldsymbol{R}^K$, calculus tells us that a local extremum will occur at points (*μ, σ*) where the first derivatives of *g*(***x***) with respect to both *μ* and *σ* are zero. But indeed, this method is guaranteed to reveal the *global minimum* because *g*(***x***) diverges as either *μ* or *σ* diverges because, in turn, *p*(*x*) approaches zero in either case. So we need to solve the following pair of equations:

$$1.\ \frac{d}{d\mu} g\left(\vec{x}\right) = 0$$

$$2.\ \frac{d}{d\sigma} g\left(\vec{x}\right) = 0$$

$$1.\ \frac{d}{d\mu}\left(\frac{1}{2}\left[K\left(ln\,\sigma^2 + ln\,2\pi\right) + \frac{1}{\sigma^2}\sum_{k=0}^{K-1}\left(x_k - \mu\right)^2\right] = 0\right)$$

$$2.\ \frac{d}{d\sigma}\left(\frac{1}{2}\left[K\left(ln\,\sigma^2 + ln\,2\pi\right) + \frac{1}{\sigma^2}\sum_{k=0}^{K-1}\left(x_k - \mu\right)^2\right] = 0\right)$$

$$1.\ -\frac{1}{\sigma^2}\sum_{k=0}^{K-1}\left(x_k - \mu\right) = 0$$

$$2.\ \frac{K}{\sigma} - \frac{1}{\sigma^3}\sum_{k=0}^{K-1}\left(x_k - \mu\right)^2 = 0$$

$$1.\ \mu = \frac{1}{K}\sum_{k=0}^{K-1} x_k$$

$$2.\ \sigma = \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left(x_k - \mu\right)^2}$$

which is exactly the textbook method by which Gaussian parameters are derived from a complete population of samples (as opposed to a subpopulation requiring Bessel's correction [5]). (The negative solution for *σ* is omitted because standard deviation is by definition nonnegative.)

This is all to say that, given a vector, the GIC will identify the Gaussian of minimum relative information cost -- in other words, the one with greatest relative probability of having produced the vector. More generally, the lesser the GIC of ***x*** relative to *N*(*μ, σ*), the greater the likelihood of the latter being the generator of the former.

5. Relationship to Gaussian Entropy

The entropy *h*(*σ*) of a Gaussian -- its information content with itself as its own prior -- is stated in [0] as

$$h(\sigma) \equiv \frac{1}{2} ln\ 2\pi\sigma^2 + \frac{1}{2}$$

Given that, and if we take ($\mu = \mu_x$) and ($\sigma = \sigma_x$), then G(μ_x, σ_x) degenerates to G(μ, σ) thusly:

$$G(\mu, \sigma) \equiv \frac{K}{2}\left[ln\ \sigma^2 + ln\ 2\pi + \frac{1}{\sigma^2}\left(\sigma^2 + [\mu - \mu]^2\right)\right]$$
$$G(\mu, \sigma) \equiv \frac{K}{2}\left(ln\ \sigma^2 + ln\ 2\pi + 1\right)$$
$$G(\mu, \sigma) \equiv K\left(\frac{1}{2} ln\ 2\pi\sigma^2 + \frac{1}{2}\right)$$
$$G(\mu, \sigma) \equiv Kh(\sigma)$$

This makes intuitive sense because we expect information cost to increase linearly with the number of samples, and that cost should be based on the optimal Gaussian, that is, the one which minimizes the information cost of explaining the samples from which it was derived in the first place.

6. Relationship to Wideband Signal Discovery

Suppose that we have identified an accurate background PDF which happens to be Gaussian, as might occur with the logs of radio power measurements at a specific frequency (channel) in the absence of significant interference. It would then be possible to measure the GIC from said background PDF to many experimental PDFs under the null hypothesis that each of the latter is simply another sampling of the former. In this case, the experimental PDF with the greatest GIC would be the strongest candidate for containing a signal.

When comparing experimental PDFs with different numbers of components, it might be more useful to first compute their GIC values *per component* (independently of *K*). Bear in mind, however, that doing so runs the risk of showing experimental PDFs consisting of "small" numbers of samples to have high statistical significance merely by chance.

In any case, the putative signal could contain components spread across many different frequencies. Its impact on any one channel might therefore not reach statistical significance. In other words, it might a be a quiet wideband signal. Moreover, although it would potentially reach statistical significance when summed over all involved channels, they would not necessarily occupy a contiguous region of the spectrum.

By computing the GIC from each background PDF to each corresponding experimental PDF in each channel (separately), then ranking them in descending order, we could quickly learn which channels were, jointly, the most likely to be harboring a weak wideband signal. Armed with this indication, we could then proceed to perform a more computationally rigorous search which might not otherwise be justifiable.

7. Normalization, Ranking, and the Wideband Cliff

But there remains the problem of scale! The (ln σ^2) term is a bias which contributes more weight to PDFs with more variance. This is a problem if our objective is to compare channels of different variances on an equivalent basis for the sake of finding a wideband signal. To reiterate:

$$G\left(\mu_x, \sigma_x\right) \equiv \frac{K}{2}\left[ln\,\sigma^2 + ln\,2\pi + \frac{1}{\sigma^2}\left(\sigma_x^2 + \left[\mu_x - \mu\right]^2\right)\right]$$

How can we eliminate this bias? Perhaps we can subtract away the entropy of the background PDF relative to itself as derived above:

$$G(\mu, \sigma) \equiv \frac{K}{2}\left(ln\,\sigma^2 + ln\,2\pi + 1\right)$$

resulting in the the "GIC delta" *d*(μ_K, σ_x):

$$d\left(\mu_x, \sigma_x\right) \equiv G\left(\mu_x, \sigma_x\right) - G(\mu, \sigma)$$
$$d\left(\mu_x, \sigma_x\right) \equiv \frac{K}{2}\left[\frac{1}{\sigma^2}\left(\sigma_x^2 + \left[\mu_x - \mu\right]^2\right) - 1\right]$$

which is now scale-free in the sense of being invariant under uniform scaling of all the means and variances. (Such scaling might occur, for example, if one were to compare the mean and variance of the voltage of a single battery with that of several of them in series.) But for ranking purposes, of course, nothing changes if we employ the more computationally economical "GIC ranker" *r*(μ_x, σ_x) in place of *d*(μ_x, σ_x):

$$r\left(\mu_x, \sigma_x\right) \equiv \frac{K}{\sigma^2}\left(\sigma_x^2 + \left[\mu_x - \mu\right]^2\right)$$

the significance of which extends beyond mere economy. It indirectly expresses -- in *semi*nats -- the amount of information in ***x*** minus (the least possible amount of information that it could contain with respect *N*(μ, σ), which would occur if all components of ***x*** were simply μ). In other words:

$$r\left(\mu_x, \sigma_x\right) \equiv 2\left[G\left(\mu_x, \sigma_x\right) - G(\mu, 0)\right]$$

wherein G(μ, 0) is sort of an information theory equivalent of a zero-point energy in physics.

Clearly, then, the GIC ranker is nonnegative. Furthermore, if in fact a weak wideband signal of sufficient power were hiding throughout a noncontiguous subset of channels, then that entire subset would exhibit greater *r*(μ_x, σ_x) than uninvolved channels. This would motivate us to sort all channels descending by their respective such values.

It would then be reasonable to expect one or more salient cliffs in the list of descendingly sorted *r*(μ_x, σ_x). In other words, we should be able to prioritize candidate channel subsets by their respective likelihoods of containing a wideband signal. Furthermore, given the noise considerations mentioned at the end of Section 3, we would probably do better to sort potential cliff locations by the "GIC ratio" *R*(μ_{x0}, σ_{x0}, μ_{x1}, σ_{x1}) of successive such GIC rankers, rather than their differences:

$$R\left(\mu_{x0},\ \sigma_{x0},\ \mu_{x1},\ \sigma_{x1}\right) \equiv \frac{K_1 {\sigma_0}^2\left[{\sigma_{x1}}^2+\left(\mu_{x1}-\mu_1\right)^2\right]}{K_0 {\sigma_1}^2\left[{\sigma_{x0}}^2+\left(\mu_{x0}-\mu_0\right)^2\right]}$$

such that we always obtain a fraction on the interval [0, (K_1/K_0)] but for a singularity when the numerator and denominator are both zero. In that case, $R(\mu_0,\ \sigma_0, \mu_1,\ \sigma_1)$ is most fairly defined to be (K_1/K_0). (The denominator cannot be zero in isolation because that would violate the assumption of descending rankers.)

The advantage of this approach is that it affords us a fair basis of comparison between a set of experimental PDFs with K_0 samples and another with K_1 samples, namely, their respective *minimum* $R(\mu_0, \sigma_0, \mu_1,\ \sigma_1)$ values. The lower this minimum in any given case, the starker the contrast between the channels potentially harboring a wideband signal, and those not doing so. It would be useful to determine, say, that it is more likely that a set S_0 of log-of-power spectra contain a wideband signal involving a subset C_0 of its channels, than it is that some other set S_1 contain some other such signal involving a subset C_1 of *its* channels -- even if both the number of channels and the number of samples per channel vary among them. In theory, the GIC ratio would facilitate comparisons of this nature.

But this thesis has yet to be tested in any experimental setup...

8. Bibliography

[0] https://en.wikipedia.org/wiki/Normal_distribution
[1] https://en.wikipedia.org/wiki/Log-normal_distribution
[2] https://en.wikipedia.org/wiki/Kullback-Leibler_divergence
[3] https://mr-easy.github.io/2020-04-16-kl-divergence-between-2-gaussian-distributions
[4] https://en.wikipedia.org/wiki/Entropy_(information_theory)
[5] https://en.wikipedia.org/wiki/Bessel%27s_correction