An analysis of optimal compression for the advanced baseline imager-based on entropy and noise estimation *

M. Grossberg,[1] S. Gottipati,[1] I. Gladkova,[1]
M. Goldberg,[2] L. Roytman.[1]

[1] CCNY, NOAA/CREST, 138th Street and Convent Avenue, New York, NY 10031
[2] ORA NOAA/NESDIS, E/RA1 5200 Auth Road, Camp Springs, MD 20746

ABSTRACT

As new instruments are developed, it is becoming clear that our ability to generate data is rapidly outstripping our ability to transmit this data. The Advanced Baseline Imager (ABI), that is currently being developed as the future imager on the Geostationary Environmental Satellite (GOES-R) series, will offer more spectral bands, higher spatial resolution, and faster imaging than the current GOES imager. As a result of the instrument development, enormous amounts of data must be transmitted from the platform to the ground, redistributed globally through band-limited channels, as well as archived. This makes efficient compression critical.

According to Shannon's Noiseless Coding Theorem, an a upper bound on the compression ratio can be computed by estimating the entropy of the data. Since the data is essentially a stream, we must determine a partition of the data into samples that capture the important correlations. We use a spatial window partition so that as the window size is increased the estimated entropy stabilizes. As part of our analysis we show that we can estimate the entropy despite the high-dimensionality of the data. We achieve this by using nearest neighbor based estimates. We complement these a posteriori estimates with a priori estimates based on an analysis of sensor noise. Using this noise analysis we propose an upper bound on the compression achievable. We apply our analysis to an ABI proxy in order estimate bounds for compression on the upcoming GOES-R imager.

Keywords: entropy, compression, compression bound, multi-spectral data, density estimation, shot noise

1. INTRODUCTION

The rapid advance of imaging technology is heralding a new age of remote sensing characterized by an unprecedented wealth of data as is represented by the upcoming Advanced Baseline Imager (ABI).[1] This wealth brings with it both opportunities and challenges. In particular, it will be difficult to move, store, and analyze the data due to its enormous volume. For example, the ABI will produce 66.6 Mbps of CCSDS packets, with twice the spatial resolution, six times the scan rate, and more than three times the spectral channels as the current GOES imager.[2]

The packets of imager data will need to be transmitted down to earth on wireless channels which unavoidably contain noise. This large volume of data must be widely distributed to users, government agencies and researchers across the world using band limited radio-frequency spectrum. It must also be archived. In moving, storing, or processing the data, the task requirements determine a notion of optimal compression. For example, when transmitting the data over a wireless channel it is necessary to sacrifice the compression ratio in order to accommodate adaptive channel encoding. This makes it possible to cope with unavoidable errors due to wireless transmission. On the other hand, error robustness is not critical requirements for transmission of an archived file via leased ground-based bandwidth. In this case, the requirement may simply be minimal file size. This case is important because any extra requirements (such as speed, efficient memory utilization, and robustness to errors) can only lower the compression ratio. Estimating the optimal compression ratio is critical in order to access whether it is feasible to trade file size for other requirements.

One approach to finding a bound on the optimal compression ratio is to perform an exhaustive empirical evaluation of all known compression algorithms for each new data set. It is often not practical because there are many algorithms each with a set of parameters, and so the brute force space that must be searched is typically too large. Even when possible, the optimal compression may be considerably higher than the maximum compression ratio obtained by the search. If search does not yield a compression algorithm that meets the necessary requirements, it does not give much insight for developing

*Sponsored by NOAA/NESDIS under Roger Heymann (OSD), Tim Schmit (STAR) Compression Group

a novel algorithm meeting the requirements. To develop such an algorithm, it is necessary to understand the structure of the data.

Clearly it is difficult to explore the structure of data for a future instrument. One approach is to use proxy data which closely matches the characteristics of the future ABI data. For example, ABI-proxy data with the planned spatial and spectral resolution has been created by resampleing high spatial and spectral resolution MODIS data.[3] While this type of proxy data is very important for testing and developing algorithms that are based on calculations depending on radiance, it is not appropriate for compression based on digital counts. In particular, subtle effects of interpolation and averaging can have an enormous impact on compression. Hence it is best to work with raw data from a current instrument that is as similar as possible.

The current GOES imagers have 5 channels, while the ABI will have 16 channels. The European Meteosat-8 imager has 12 channels, 11 of which are similar both spectrally, and in spatial resolution, to channels that the ABI will have. Hence, as suggested by T. Schmit we used this as our ABI-proxy. A false color image combining information from all of the 11 channels we used is shown in Figure 1. We emphasize that our primary goal here is present a methodology for estimating the optimal compression given a valid proxy data set. The actual compression ratios achievable for the ABI can differ significantly than the bounds we have found using the Meteosat-8 proxy data. For example the 11 channels we used from Meteosat-8 have a bit depth of 10, and a ground-sampled distance of approximately 3km at nadir.[4] For the ABI the bit depth will vary from 12-14 bits and have resolution varying between .5km and 3km. Nevertheless, it is the most similar instrument available for this study.

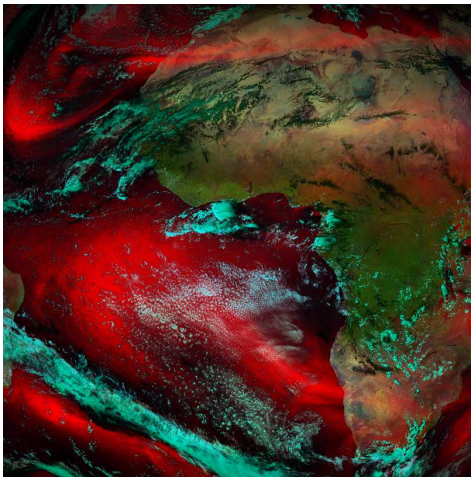


Figure 1. A false color image combining information from the 11 channels of the 12 channel Meteosat-8 imager. The red swirls show data from the water vapor channels.

Formally the structure of the data, as it relates to compression, can be analyzed using information theory. In particular, Shannon's pioneering work demonstrates that the optimal compression of a data stream is its entropy.[5] The entropy can be computed by modeling the data stream as a random process. We will reorganize the ABI proxy data into a data stream. We will show that by estimating the statistical dependencies in this stream, it can be approximated by a random process. We then give statistical estimates of the entropy of this data stream. We also show that these estimation tools reveal dependencies and structure of the data.

The structure of imager data depends on radiance from the scene and the characteristics of the measurement process. Systematic bias in the measurement process can be removed by calibration. Calibration can also quantify sensor noise. Great care is taken to reduce or eliminate sensor noise. However, measurement noise such as shot noise, has a basis in the physics of measurement itself and cannot be eliminated. This noise will be present in even the most structured signal. Thus, we will show that assumptions on the shot noise can give lower bounds on compression independent of the structure of the data.

The analysis we present should give tools enabling the development of custom algorithms tuned to the data, sensor, and mission requirements. It will be useful for cost benefit studies. For example, we can focus our efforts more efficiently if we know which parts of the data stream are more difficult to compress through analysis of the data's structure.

2. IMPACT OF NOISE ON COMPRESSION

There are several factors that contribute noise to digital count. Primary are thermal noise, electronic noise, shot noise, and quantization.[6] There are also related sources of bias to the sensor. For example thermal noise has a non-zero mean related to the temperature of the sensor. Calibration procedures are designed to remove these biases.[7] Even when present, biases do not typically affect the compression ratio. For example, if the temperature of the sensor is uniform, the thermal noise will add an offset to the value at each pixel. This does not change the entropy.

What remains to consider are sources of zero mean noise which can have a dramatic impact on compression. Noise sources are typically modeled as, poisson, gaussian, or uniform noise. All these noise sources have relatively high entropy. High quality manufacturing can reduce electronic noise often below the quantization threshold. For example, sensors are cooled to reduce thermal noise. Channels $7-16$ of the ABI have already been analyzed to determine the appropriate quantization levels due to the thermal and electronic noise.[8]

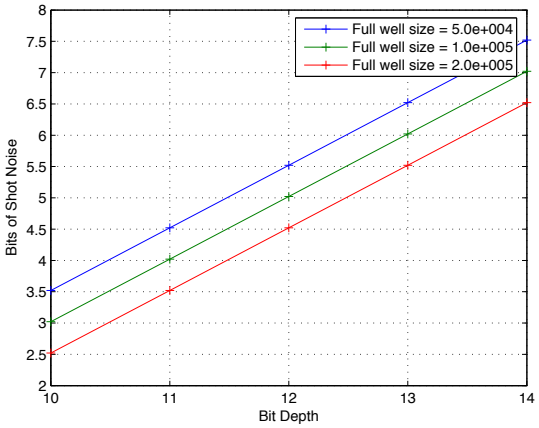


Figure 2. The three curves represent the shot noise in bits as a function of bit depth for three different full well values.

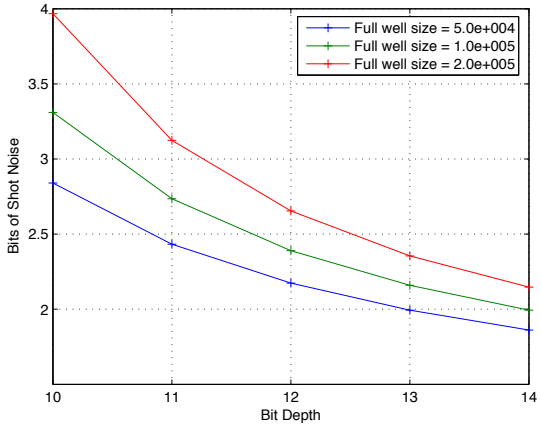


Figure 3. The three curves represent the compression ratio for pure shot noise in bits as a function of bit depth for three different full well values.

Unlike thermal and electronic noise, the standard deviation of the shot noise varies with square root of the scene radiance. Shot noise is not a result of any flaw or property of the sensor, but the natural variations in the arrival of photons. Photon arrivals can be modeled by a poisson process. A poisson process has only one parameter which is the mean μ_{poisson}.[9] If the mean is known and subtracted, the variations are approximately like a gaussian with mean μ_{poisson} and standard deviation $\sqrt{\mu_{\text{poisson}}}$. Hence, the shot noise depends only on the total number of photons the sensor collects. The photons are measured by counting electrons corresponding to photon absorbtion. If the measurements span the range of a linear sensor, this can be determined by knowing the maximum number of electrons the sensor can collect. This number is called the full well size. In Figure 2, the standard deviation in bits for each of three different plausible ABI full well sizes is shown. Figure 3 shows the compression ratio of shot noise for three different full well values at different bit depths. While there may be great scientific value in collecting noisy data we note that even if the thermal and electronic noise are relatively low, the shot noise can have a detrimental impact on compression. This must be considered if high compression rates are required.

3. ENTROPY AND COMPRESSION

A data stream is a sequence of messages $a_1, a_2, \ldots$. For example each a_i may be a granule. Suppose there are J possible messages. We may model the stream as a sequence of samples from a discrete probability p, where p_j is the probability of message a_j. We we will assume that the messages in the stream are mutually independent. In fact the messages need only be effectively independent. For example, it is often a requirement that a compression algorithm be able to compress and decompress granules independently. We may also consider a stream of messages independent if the added benefit of

compressing them together is marginal. From Shannon's theorem,[10] the optimal compression of the stream is $H+1$, where H the information entropy given by

$$H = -\sum_{j=1}^{j=J} p_j \log(p_j). \tag{1}$$

Here the $\log$ is computed in base 2 and thus the entropy is measured in bits.

In principle, it is a simple matter to determine the optimal entropy if one can:

- Represent the data as a stream of independent messages a_i
- Compute the probability p_i for each possible message a_i
- Perform the summation in Equation 1

All of these steps present challenges. It is clear that the sequence of granules from an imager is far from independent. Each image from the ABI will be taken at relatively short intervals in time, for example 15 minutes apart. Still as described earlier, for logistical reasons it is not practical to require that multiple granules must be kept together.

A more serious problem is that the size of the data makes it unreasonable to deal with whole granules at a time. The proposed granule size for the ABI imager is approximately 2^{24} bits. Hence, we must break the data into parts, each of which can be treated as approximately independent to a size smaller than a granule. To maintain this independence we break the data into spatial windows of varying sizes and estimate the independence. In fact the relevant quantity of interest is the mutual information. As this becomes small we may treat the messages independently.

Even by breaking the granule up into smaller chunks we are faced with formidable computational problems. For example, consider 25×25 spatial windows in all 11 channels of the Meteosat-8 imager. Each message then consists of 68750 bits. While this seems small even by today's standards, the number of possible messages is 2^{68750}. Directly computing p_i and performing the sum in Equation 1 are beyond the reach of computation for the foreseeable future.

Rather than compute the entropy in the discreet domain, we will use a continuous proxy for the entropy, called the differential entropy.[5] The differential entropy is given by the formula

$$H = -\int_{x=1}^{x=B^J} p(x) \log(p(x)), \tag{2}$$

where x is taken to be a continuous message (real numbers rather than integers) and B is taken to be the bit depth. In this way we treat each channel of each footprint as its dimension. The dimensionality for 25×25 using the Meteosat-8 imager is 6875 dimensions. While the use of the differential entropy is necessary to reduce the complexity, it is far from sufficient. Estimation of probability distributions in high dimension is difficult in general. Estimation of p in so many dimensions requires more data than can be obtained. In worse case scenarios, it would be necessary to have enough data so that multiple instances of each possible message. Hence, even though a granule provides us tens of thousands of samples, we are always in a "data poor" environment.

Fortunately, the data does not come from an arbitrary source. It comes from a set of physical measurements. This means that the probability distribution is reasonably well behaved. For example, successful algorithms that compress hyperspectral AIRS data exploit the fact that this data clusters around low dimensional hyperplanes. While hyperspectral sounder data has strong spectral corrolations, Imager data is more closely related to monochromatic images in that spatial correlations are responsible for much of the compression. In either case the data can be broken into a structure component which typically is concentrated in a few dimensions, and a noise portion, which is close to uniform across the remaining dimension. By noise here we include that stochastic portion of the data which is independent from between individual measurements. This includes the thermal, and electronic noise of the sensor. It also includes the shot noise due to the random arrival of photons at the sensor.

The weak assumption that our data is concentrated in low dimensions. This assumption makes it possible to overcome the problems described above and estimate the entropy despite the high dimensionality. We have developed a technique which is based on non-parametric estimation of the entropy. The estimator we use is based on the K-L estimator but we have introduced a novel method to reduce the bias for use in high dimensions.[11,12] The technique can also be used to

estimate mutual information in high dimensions. Using mutual information we show that, for imager data, it should be possible to compress the data effectively with relatively small blocks.

In this paper we have developed a new non-parametric technique to estimate the differential entropy of high dimensional data, such as windowed ABI imager data. We have done this by introducing a novel method to remove the bias of the K-L entropy estimator when the bias is asymptotically gaussian. The same methodology allows us to compute the mutual information. Using these tools we show that proxy ABI imager data, such as that from Meteosat-8, can be effectively windowed. We will use our novel tools to evaluate the linearity of the data which is the foundation of transform methods such as wavelets. We will evaluate the relative importance of the spectral and spatial components.

4. ENTROPY ESTIMATION TECHNIQUES

Since the entropy is based on the probability distribution, entropy estimation is related to estimation of the probability. While the imager data is quantized, the bit depth of the data is greater than 10 bit. Typically the noise levels are larger than the quantization. Moreover, because we are averaging over many measurements, we can approximate the discrete entropy by computing the differential entropy of a continuous probability density function.

There are two basic classes of techniques to estimate probability density functions: parametric and non-parametric. When a density function can be assumed to be close to a known family of densities, the density function can be fit to the data. These parametric entropy estimates are called "plug-in" estimates. These are faster to compute, less complex, scale to very high dimension, and not as sensitive to "data poor" environments. Hence we will discuss our analysis of these first.

4.1. Parametric Techniques

The most widely used and simplest model of multi-dimensional data is to assume that it is a multi-dimensional gaussian p. This assumption is the basis of the PCA algorithm which attempts to find a hyperplane that lies close to the data.[13] For example, if we break the data into 10×10 spatial windows of footprints (all channels), then the measurements in this block can be treated as a vector. The data set is then represented as the set of these vectors. The plot in Figure 4 shows the mean of the set of vectors. Each step is a different channel. The flat shape indicates that the average value spatially is consistent for each channel. The images shown in Figure 5 are normalized values, per channel of the first principal component. While some spatial variation is apparent, this is due to the per channel normalization. The primary variation in the data is in the spectral direction rather than across space. The second principal component as is shown in Figure 5 also primarily shows variations in the spectral dimensions. For comparison the smallest principle component is shown in Figure 7. The checkerboard like pattern is similar to that of a high frequency discrete cosine transform.

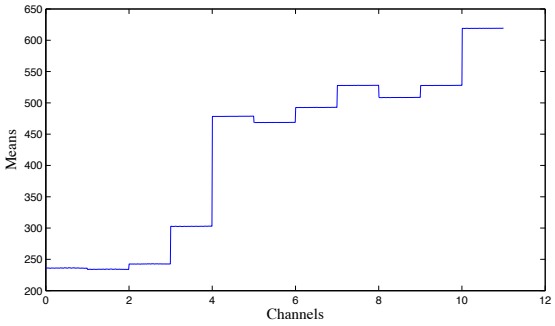


Figure 4. The mean of 10×10 spatial windows of footprints (all channels) of the ABI-proxy data. Note that the mean is constant within each channel.

If the data were actually gaussian, the matrix of eigenvectors provides a rotation which decouples the channels. The eigenvalues are then variances of a set of independent gaussian, one for each channel. A graph of the eigenvalues of the covariance matrix of 10×10 windows from 11 channel Meteosat-8 data is shown in Figure 8. For compression this shows that a good first order approximation can be obtained by projection into the top eigenspaces. In Figure 9 variance of the subspaces is computed from a cumulative sum of the normalized eigenvalues for the same data. Note that 99% of the

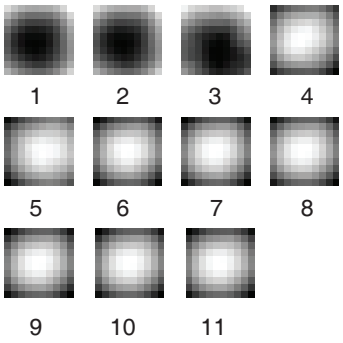


Figure 5. The first principal component of windowed ABI-proxy data as 11 normalized images of size 10×10.

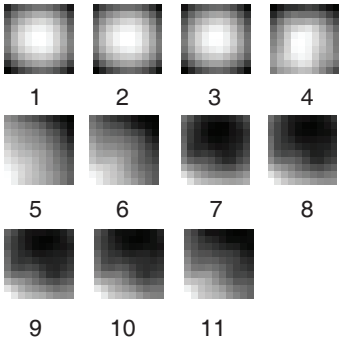


Figure 6. The second principal component of windowed ABI-proxy data. Both the first and second principal components show a dominant variation in the spectral dimensions.

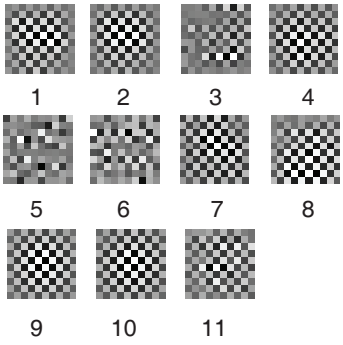


Figure 7. The smallest principal component of windowed ABI-proxy data. Note the pattern is similar to that of a high frequency component of the discrete cosine transform.

variance is explained by just 43 out of the 1100 dimensions. Given that much of the variation is in the spectral dimension, this indicates that treating the channels together is important for compression.

The plot shown in Figure 10 shows the log of the eigenvalues for different window sizes. Here the largest eigenvalues have been aligned to facilitate comparison. Note that after re-scaling the eigenvalue curves are quite similar. From this there does not seem to be a dramatic change in the structure of the space as more spatial correlations are taken into account. This also points to importance of the correlations in the spectral channel. While the eigenvalue curves seem to converge until the window size reaches 10×10 they seem to diverge for larger windows. This can be explained by the fact that as the window size increases, the number of samples decreases. When there are two few samples, the small eigenvalues are underestimated and the large eigenvalues over estimated as shown in Figure 11. This creates a bias which would result in a higher estimate of the optimal compression ratio than is correct.

We determine a compression ratio bound from the computed eigenvalues by noting that the entropy for the gaussian is given by

$$H(p) = \frac{n}{2} \log(2\pi e \det(C)), \tag{3}$$

where $\det(C)$ is the determinant of the covariance matrix. This can be computed as the product of the eigenvalues. We note that a given a fixed covariance, a gaussian distribution has the maximal entropy of all distributions with the same covariance.[5] Hence the compression ratios derived from a gaussian estimate represent *lower bounds* on the compression rate. To obtain this lower bound we must still determine whether the window sizes are independent. The mutual information[5] measures the benefit of compressing neighboring windows together rather than separately. The mutual entropy is plotted in Figure 12 as a function of window size. It decreases until just after a window size of 10×10 and then appears to increase due to undersampling. While 10×10 are not independent, the effective dependency is only about 5% of the entropy.

The plot in Figure 13 gives a compression ratio bound on each window size assuming the distribution of the data to be gaussian. We note that a gaussian has the highest entropy of any distribution with the same covariance. Hence the bound on the optimal compression ratio computed from a gaussian, will be strictly lower if the distribution of the data is not gaussian. For 10×10 windows the lower bound on optimal compression is 2.6. With this window size, we expect that 2.6 to 1 is best that can be obtained from any strictly linear transform such the Discrete Cosine, Wavelet or Karhunen-Loeève (PCA) transform.

One clear weakness of a parametric model based on a single gaussian is that it assumes the data is uni-modal centered about the mean. In hyper-spectral imaging data there are obviously inhomogeneities in the data. Intuitively these should arise from the existence of different classes of cloud cover, as well as land and oceans. For example, 10,000 samples of 10×10 spatial window of footprints is shown in Figure 15 projected from 1100 dimension into 3 dimensions. It is clear from this image that there is an in-homogeneous structure that cannot be captured with a single gaussian. One way

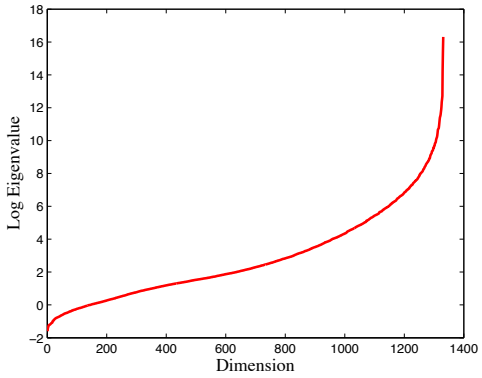


Figure 8. A Log-Eigenvalue curve for the covariance of 10×10 windowed ABI-proxy data using all 11 channels. Note that most of the variation is concentrated in the largest eigenspaces.

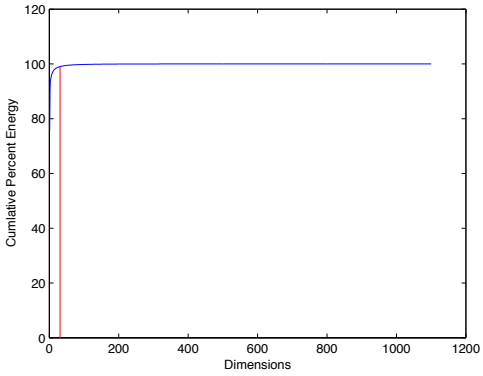


Figure 9. Graph showing the percentage of the total variance explained by subspaces. Note that 99% of the variance is explained by just 43 out of the 1100 dimensions.

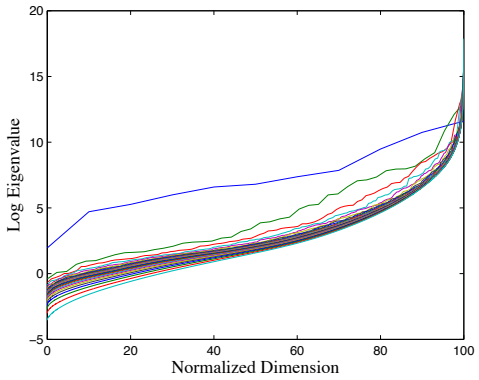


Figure 10. Rescaled log-eigenvalue curves for windowed ABI-proxy data. Spatial window sizes ranging from a single footprint (with all channels) to 25×25 are shown as different curves. The curves converge as the window sizes increase.

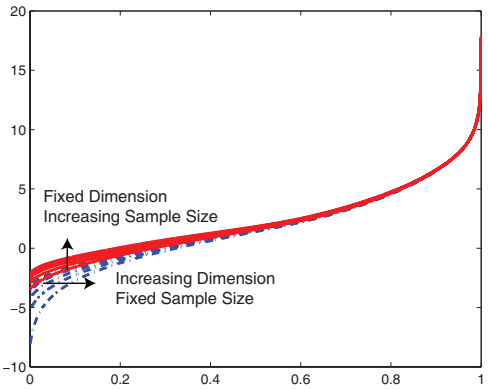


Figure 11. Plot of rescaled log-eigenvalue curves showing the effect of undersampling. Undersampling occurs if the number of samples is small relative to the dimensionality.

to generalize the parametric gaussian model is to capture the inhomogeneity present in the data by model the probability density as a gaussian mixture model (GMM). The density function for this model p_{GMM} is given as

$$p_{GMM}(x) = \sum_{i=1}^{k} \rho_i K_i \exp((x-\mu_i)^T \Sigma_i (x-\mu_i)), \tag{4}$$

where μ_i, Σ_i, are the respective mean and covariance matrix of the i-th mixture, where K_i are the gaussian normalizing constants, and p_i are the mixture proportions. A mixture model simultaneously performs a soft clustering of the data with linear fitting on the clusters.

There are several algorithms to establish a mixture of gaussians, the most popular being Expectation Maximization (EM). This is an iterative algorithm and can be very sensitive to an initialization and the number of clusters chosen. Using an algorithm based on EM we were able to estimate the entropy for a footprint and a 2×2 window of footprints. The estimated entropy gives a considerable improvement over the entropy estimation of a single gaussian. Unfortunately, generalizing this technique to higher dimension requires addressing some of the instabilities that are aggravated by high dimensional parameter estimation.

Parametric techniques to estimate entropy have a number of advantages. With enough mixtures, and sufficient data, an arbitrarily complex density can be approximated. The parametric form also partially compensates for the "data poor"

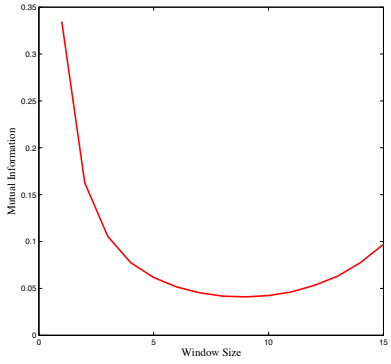


Figure 12. A graph of the mutual information vs window size. The mutual information is only 5% of the entropy for a 10×10 window. The apparent increase in mutual information for larger window sizes is due to undersampling.

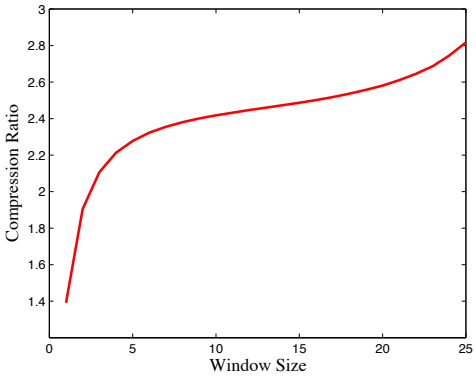


Figure 13. A plot of the compression ratios bound as a function of window size based on a single gaussian. This is a set of lower bounds on the optimal compression ratios.

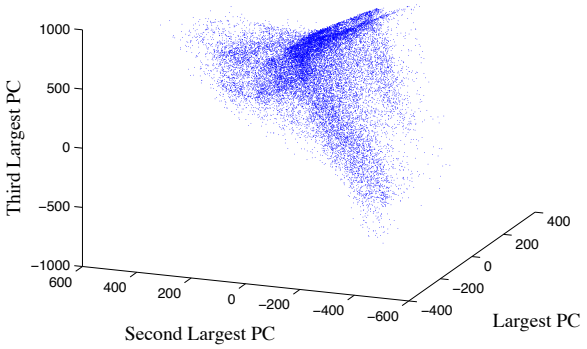


Figure 14. A scatter plot of the 11 spectral channels ABI-proxy data projected into the first three principal components.

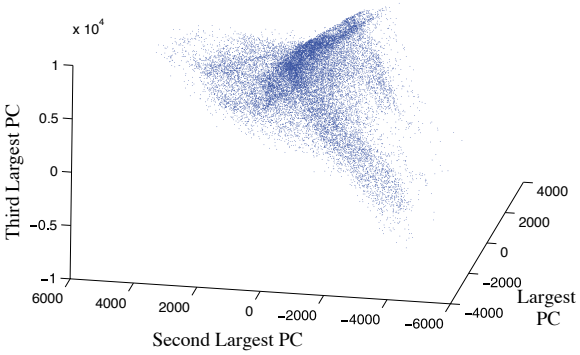


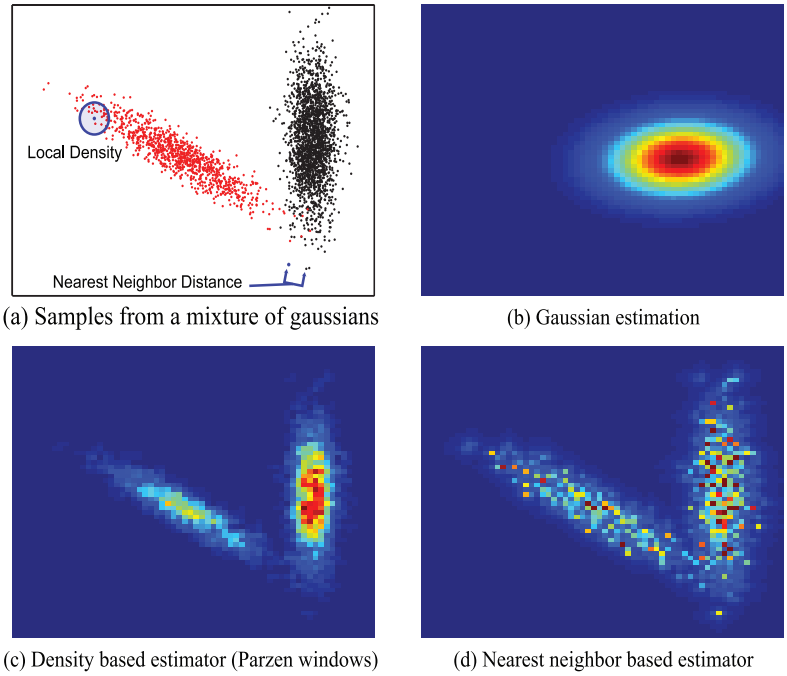
Figure 15. A scatter plot of the 10×10 window of the 11 spectral channels ABI-proxy data projected into the first three principal components.

environment by interpolation. Another advantage of a parametric approach is that the integration needed to compute the entropy from the estimated probability can be computed analytically for any dimension. This completely sidesteps difficulties of high dimensional integration.

Unfortunately, parametric techniques to estimate probabilities can become unstable in high dimensions. Moreover, they impose a prior assumption on smoothness of the data. Hence, even though it is theoretically possible to use a large number of clusters to fit the probability density, in practice GMM works well if the data can be well approximated by a small number of clusters. The last point is that any "plug-in" estimator tries to optimize the fit of the density function rather than optimize the entropy accuracy. Instead we will explore an approach which does not make as many assumptions.

4.2. Non-Parametric Techniques

If there is sufficient data, and the data is one dimensional, a histogram can give a very good non-parametric approximation of the probability density and the entropy. Even in moderately low dimensions the number of bins becomes exponentially large. Thus histograms are not appropriate for estimating the entropy in high dimensions. There are two approaches to estimating the histogram that do not depend strongly on dimension to implement. One is based on a local density estimation, and the other is based on distance between neighboring data points as shown in Figure 16.



(a) Samples from a mixture of gaussians (b) Gaussian estimation

(c) Density based estimator (Parzen windows) (d) Nearest neighbor based estimator

Figure 16. A comparison of density estimators. The upper left figure illustrates a multi-modal data distribution and two methods of non-parametric density estimates. On the upper right the image represents a gaussian density estimate, which does not fit the data well. The lower left image shows the estimate from Parzen windowing and the lower right image shows an estimate based on the K-L nearest neighbor approach.

The density approach is based on the idea that if the probability density function at a point q is high, the number of data points in a fixed volume around q should also be high. This approach is called Parzen Windowing. More generally, rather than counting the number of points in a fixed volume, a kernel function can be used to weight points according to their distance from q. Hence, distant points count less. Although this method is non-parametric, the choice of the kernel imposes a chosen prior on the smoothing. This is both an advantage and a disadvantage of this method. The advantage is that it allows local assumptions on the probability distribution to be incorporated in the estimate. The disadvantage, as with parametric estimation above, an incorrect choice of kernel leads to a biased estimate.

The distance approach is based on the idea that if the probability density function at a point $\mathbf{q}$ is high, the distance to nearest neighbors will be small. More precisely, for a locally constant density p the expected value of the distance d to the next nearest neighbor can be computed. Rather, given the expected distance to the nearest neighbor d, it is possible to compute the value of a locally constant probability p with that expectation. Hence, by computing the distance to nearest neighbors, we are able to obtain a probability density for each point $p(\mathbf{q})$. Unlike the density approach above, this approach does not require any parameters, and the estimate comes solely from the data. On the down side, if the point $\mathbf{q}$ is identical to a data point, the estimate will blow up. This is actually very rare for high dimensional data.

To use either of the above techniques to compute the entropy we have to integrate the estimate over a high dimensional space. In general, such integration is not an easy task. However there is a way to combine the non-parametric estimation with the integration using a Monte-Carlo approach. Recall that for a function $f(x)$, with $x \in \Omega$, its integral with respect to the probability distribution p can be estimated by random sampling. The idea is to first choose a random sequence of points $x_1, ..., x_n$ from Ω using p. For sufficiently many points,

$$\int f(x)p(x)dx \approx (1/N)\sum_{i=1}^{N} f(x_i). \tag{5}$$

Note that the summation does not need to have p in it. That is because if the samples $x_1, ..., x_n$ are chosen from Ω using p, the summation will be automatically weighted to favor high probability locations. If we substitute an estimate obtained for log p, into Equation 5, we get a Monte-Carlo entropy estimate. The non-parametric estimate we have chosen for log p is the $K - L$ estimator, based on a fast nearest neighbor search.[11] The advantage of using nearest neighbor for the local estimation is that since the method does not have any parameters, it is completely driven by the data itself. The $K - L$ estimator we have developed can be computed in arbitrary dimension. For physical data we have developed a nearest neighbor search algorithm that can run very rapidly. It also scales with computational resources in that one can trade accuracy for speed very easily.

5. ENTROPY ESTIMATION IN HIGH DIMENSION WITH BIAS REMOVAL

Since the K-L estimator does not assume a model and is completely data driven it should not introduce a bias. Unfortunately there are some assumptions, such as the local constancy of the probability distribution function, that introduce a bias. This bias disappears with enough samples, see.[11] Unfortunately the number of samples needed to eliminate the bias can become enormous. The surface shown in Figure 17 shows the bias of the K-L estimator with data chosen from an isotropic multi-variate unit gaussian. The bias increases with dimension and decreases with sample size. Even though the bias asymptotically vanishes with increasing sample size, the bias shrinks very slowly. For some distributions, it is possible that number of samples needed to decrease the bias below a threshold is exponential in dimension. In cases it is practically impossible to obtain or even store enough data to eliminate the bias of the K-L estimation in high dimension. Hence in order to estimate the optimal compression using the K-L estimator we need to account for the bias. The plot in Figure 17 includes examples for data where the dimension is in the range 1-1000. If we increase sample size linearly with dimension, the bias remains stable and is possible to estimate.

We do argue that it is possible to estimate the bias for our data set. The reason is that as we have seen above, the data exhibits considerable structure in the first few principal components. However, as we consider components with smaller eigenvalues the eigenvalue curve asymptotically flattens and begins to look more like gaussian noise. Due to the high quality of the imager equipment, and the abundance of reflected photons within the broad bands that the imager measures, this is unlikely to be measurement noise. However, it is consistent enough so that although we cannot treat it as gaussian in computing the entropy, we can use a gaussian proxy for computing the *bias* of the K-L estimator. We have validated this approach on synthetically generated data from a large family of multi-modal distributions that have a similar asymptotic gaussian behavior to the Meteosat imager data. Even though the entropy and parameters of the distributions varied, the K-L estimation bias was largely determined by the bias of a single gaussian fit. The plot in Figure 19 shows that even with the bias correction, the bias still increases with dimension. However, as can be seen from the graph in Figure 18 the percent error in our bias corrected entropy estimate remains fixed at around 4%. This provides evidence that our novel method can give an accurate estimate of the entropy, and hence the compression ratio, even in high dimensions.

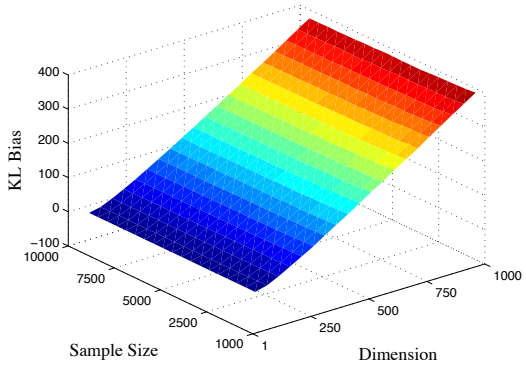


Figure 17. Surface showing the bias for K-L entropy estimator applied to an isotropic unit gaussian distribution as a function of dimension and sample size. Even though the bias vanishes with increasing sample size, the bias shrinks very slowly.

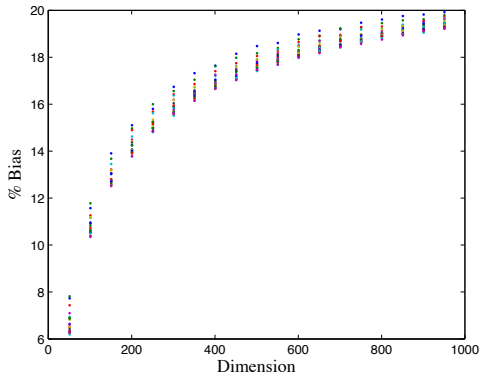


Figure 18. Scatter plot of bias verses dimension for different sample sizes. If the ratio of sample size to dimension is held fixed the bias follows a well behaved curve.

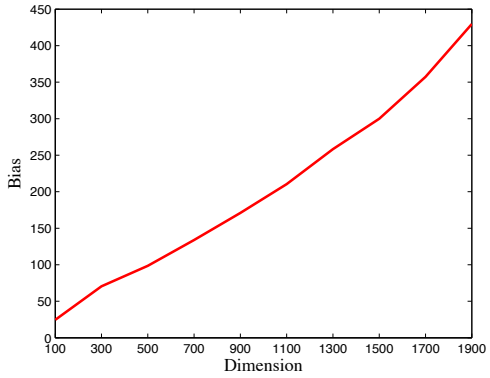


Figure 19. A plot showing the bias in entropy remaining after applying our bias corrected K-L estimator applied to known distributions as a function of dimensions applied to a validation set.

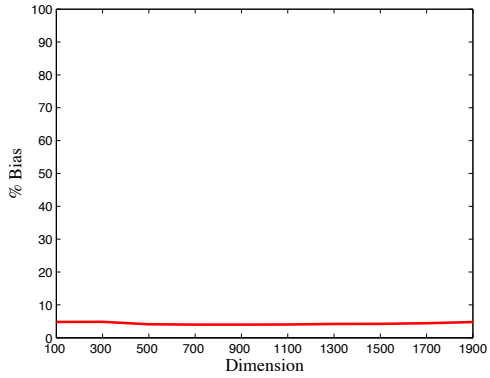


Figure 20. A graph showing the percent bias remaining after applying our bias corrected K-L estimator applied to known distributions for a range of dimensions. For the validation sets, the accuracy is constantly within %4.

We applied our bias-corrected K-L entropy estimator to the Meteosat-8 image data to estimate the compression ratio as a function of window size. As can be seen from the plot in Figure 21, the increase in the estimated upper bound of the compression ratio slopes dramatically although it continues to increase slowly. As was seen in Figures 11-12 the estimates become corrupted due to undersampling for larger windows. At a 10×10 window size there are a sufficient number of samples for this window size for the estimate to be reliable. With these parameters we estimate the 11 channel Meteosat-8 imager data has estimated upper bound compression ratio of 3.1-to-1. Since the ABI differs from Meteosat-8 imager in resolution, spectral sensitivity, and number of channels, these ratios cannot directly be applied to the ABI.

6. CONCLUSION

In this work we have presented a framework for computing bounds on the optimal lossless compression ratios of imaging data. For current instruments such as the imager on the Meteosat-8, our method may be used to evaluate compression algorithms, for archiving, or for data distribution. For the future instruments such as the ABI, our framework can give estimates on the upper bound for compression ratios, to the extent that proxy data that matches the future data is available. We applied our framework to 11-channels of the Meteosat-8 data set. We have estimated a bound on the lossless compression ratio by considering the measurement noise. We concluded that the potential effect of the shot noise can severely limit lossless compression ratios for high bit depth.

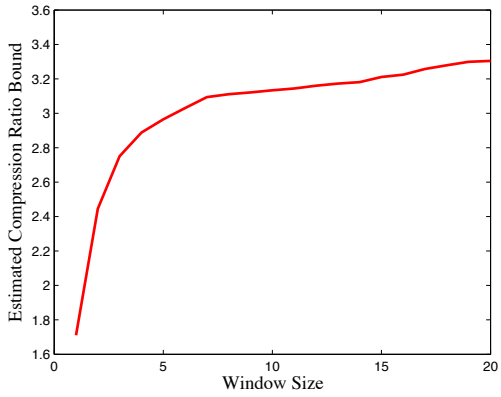


Figure 21. Graph of the compression ratio vs. window size using our bias-corrected K-L entropy estimator. For a 10×10 window the estimated compression ratio is 3.1. The subsequent increase in apparent compression ratio is largely due to undersampling.

We also analyzed the limits of compression based on computing the entropy of the Meteosat-8 imager. In order to capture the important dependencies in this data we broke the data into blocks with all the spectral channels for windows in the spatial domain. We showed from the point of view of compression, the data can be considered independent for large window sizes. Due to the constraint of compression based on a single granule, we found that a 10×10 window of data provided a good compromise between a sufficient number of samples and their independence. We also obtained a lower bound of 2.6-to-1 on the optimal lossless compression using a linear model applied to this data. Using the linear model, a projection in the first three components shows that the data is not distributed as a single gaussian, rather it is multi-modal. It is also clear that although spatial dependencies are important the probability density of windowed and single footprint data are quite similar. This similarity in densities indicates that the variations in the spectral domain are more dominant than those in the spatial domain.

The complexity of the ABI proxy data indicates that the strong assumptions implicit in PCA and other parametric entropy estimates do not hold. To avoid these assumptions we have developed a non-parametric estimator based on the K-L nearest-neighbor estimator. While the K-L estimator may be used in arbitrary dimensions, it suffers from bias which becomes enormous in high dimensions. By estimating the bias for a gaussian proxy we have modified the K-L estimator so we can accurately determine the entropy even in very high dimensions. With this novel tool we have estimated that 10x10 windows of 11 channel Meteosat-8 data have a theoretical compression bound of 3.1 to 1. Although this number cannot be directly applied to the ABI, the similarity of a significant portion of the data sets to 11 channel Meteosat-8 should be helpful in accessing the risk associated with ABI designs dependent on significantly higher lossless compression.

7. ACKNOWLEDGMENTS

This work is being sponsored by NOAA/NESDIS and has been prepared in support of the NOAA/NESDIS satellite earth science remote sensing data compression research group led by Roger Heymann of its Office of Systems Development and Tim Schmit of its Office of Research and Applications.

REFERENCES

1. T. J. Schmit, M. M. Gunshor, W. P. Menzel, J. Gurka, J. Li, and S. Bachmeier, "Introducting the next-generation advanced baseline imager (ABI) on GOES-R," *Bull. Amer. Meteorol. Soc.* **86**, pp. 1079–1096, 2005.
2. ITT Industries, "Next Generation Sensor," tech. rep., http://www.ssd.itt.com/literature/pdf/ABI_imager.pdf, 2006.
3. T. J. Schmit, "http://cimss.ssec.wisc.edu/goes/abi/bitdepthcompression/bitdepthcompression.html," tech. rep., NOAA/STAR, 2006.
4. J. Schmetz, P. Pili, S. Tjemkes, D. Just, J. Kerkmann, S. Rota, and A. Ratier, "An Introduction to Meteosat Second Generation (MSG).," *Bull. Amer. Meteor. Soc.* **83**, pp. 977–992, 2002.
5. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, 1991.
6. G. Healey and R. Kondepudy, "CCD camera calibration and noise estimation," *CVPR* **92**, pp. 90–95, 1994.
7. D. W. Hillger, "GOES imager and sounder calibration, scaling and image quality," tech. rep., NOAA, June 1999. NESDIS 93, 34-p.
8. D. W. Hillger and T. Schmit, "Quantization noise for GOES-R ABI bands," in *Proceedings of the 13th Conference on Satellite Meteorology and Oceanography Norfolk, VA, 20-23*, *American Meteorological Society*, Sept 2004.
9. S. M. Ross, *Introduction to Probability Models*, Academic Press, New York, 8th ed., 2003.
10. C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal* **27**, pp. 379–423, 623–656, 1948.
11. L. F. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problems of Information Transmission* **23**(2), pp. 95–101, 1987.
12. L. Paninski, "Estimation of entropy and mutual information," *Center for Neural Science* **15**, pp. 1191–1253, 2003.
13. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, S.E*, Wiley Interscience, 2000.