On Model Order Estimation by Information Theoretic Criteria

Bernard Lampe, *Member, IEEE*

***Abstract*—Model order selection is an important first step when fitting a model to observed data. A very prolific method was developed by Wax and Kailath and is used widely for both model-driven and data-driven methods [6]. The solution put forth in their paper assumes a complex multivariate Gaussian model for the signal subspace and an isotropic Gaussian noise model for the noise subspace. In addition, they assume a linear mixture model of the signal subspace and noise subspace. With those assumptions they develop explicit Akaike Information Criteria (AIC) and minimum description length (MDL) and Bayesian Information Criteria (BIC) based estimators for selecting model order. In this study we investigate their formulation through empirical analysis with regard to estimator consistency and robustness to model mismatch.**

***Index Terms*—Akaike Information Criterion, AIC, Minimum Description Language, MDL, Bayesian Information Criteria, BIC, Information Theoretic Criteria, ITC**

I. Introduction

MODEL order selection is the focus of this study. In particular the formulation of information theoretic criteria to detect the number of signals. We follow the developments of Wax and Kailath [6] which is based on the the work of Akaike [1], Schwartz [5] and Rissanen [4]. In these works, they develop the Akaike Information Criteria (AIC), minimum description length (MDL) and Bayesian Information Criteria (BIC) to be used to detect the dimensionality of the signal subspace. We investigate these criteria empirically by implementing each in Matlab and observing the consistency and robustness of the estimators as the number of samples increases. Following what Wax and Kailath developed in their paper [6], we use a stationary, ergodic and zero mean multivariate Gaussian random process for the signal subspace and isotropic multivariate Gaussian random process for the noise subspace.

The first step in the empirical investigation is to choose an appropriate mathematical formulation of the signal subspace and noise subspace. Following the procedure in [3] and [6], we assume a linear mixing model as in equation 1. This equation decomposes the t-th observation, $\boldsymbol{x}(t)$, as a linear combination of the q vectors in the signal subspace $\{\boldsymbol{\alpha}_i\}_{i=1}^{q}$, where the scalars $s_i(t)$ are the coefficients of the t-th signal and the vector $\boldsymbol{n}(t)$ is the noise subspace vector at index t. Where $\boldsymbol{x}(t) \in I\!R^{p\text{x}1}$, $\boldsymbol{\alpha}_i \in I\!R^{p\text{x}1}$, $s_i(t) \in I\!R$ and $\boldsymbol{n}(t) \in I\!R^{p\text{x}1}$.

$$\boldsymbol{x}(t) = \sum_{i=1}^{q} \boldsymbol{\alpha}_i s_i(t) + \boldsymbol{n}(t) \tag{1}$$

We can reformulate equation 1 to represent all N observations as in equation 2. Here the observations are $\boldsymbol{X} = [\boldsymbol{x}(1)|\boldsymbol{x}(2)|\ldots|\boldsymbol{x}(N)]$, the signal subspace is $\boldsymbol{A} = [\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_2|\ldots|\boldsymbol{\alpha}_q]$, the signal coefficients are $\boldsymbol{S} = [\boldsymbol{s}(1)|\boldsymbol{s}(2)|\ldots|\boldsymbol{s}(N)]$ and the noise is $\boldsymbol{N} = [\boldsymbol{n}(1)|\boldsymbol{n}(2)|\ldots|\boldsymbol{n}(N)]$. Where $\boldsymbol{X} \in I\!R^{p\text{x}N}$, $\boldsymbol{A} \in I\!R^{p\text{x}q}$, $\boldsymbol{S} \in I\!R^{q\text{x}N}$ and $\boldsymbol{N} \in I\!R^{p\text{x}N}$. We also assume the system in 2 is overdetermined, therefore $q < p$.

$$\boldsymbol{X} = \boldsymbol{AS} + \boldsymbol{N} \tag{2}$$

Our goal is to accurately determine the dimensionality of the signal subspace q. If we assume that the signal subspace matrix $\boldsymbol{A}$ is of full column rank, we can observe the eigenvalues of the correlation/covariance matrix of the data, denoted $\boldsymbol{\lambda}^T = [\lambda_1, \lambda_2, \ldots, \lambda_p] = \sigma(E\{\boldsymbol{XX}^T\})$, to determine the dimension of the signal subspace. Where $\lambda_1 > \lambda_2 > \cdots > \lambda_p$. The first q eigenvalues attributed to the signal should be much larger than the $p - q$ eigenvalues attributed to the noise if there is good signal to noise ratio in the data (SNR). In practice this approach is made difficult by the additive noise meaning the separation of ”large” eigenvalues from ”small” eigenvalues becomes ambiguous. In order to automate this process, Wax and Kailath purposed changing the eigenvalue detection problem into a model selection problem [6].

A. AIC and MDL/BIC Criteria

In order to convert the eigenvalue separation problem into a model selection problem, we need to select an appropriate statistical model $f(\boldsymbol{X}|\boldsymbol{\Theta})$. Where $\boldsymbol{X}$ is the observation matrix and $\boldsymbol{\Theta}$ is the set of model parameters. Akaike proposed that the best fit model is the one with the minimum AIC [1] as in equation 3. Schwartz and Rissanen proposed the minimum MDL and BIC respectively which turn out to be the same estimator in the limit of number of observations as in equation 4 [4], [5]. Therefore, we consider MDL and BIC the same in this study. In each equation k is the evaluated model order. Both estimators, at the minimum, find the model order with the minimum Kulback-Liebler distance between the model density $f(\boldsymbol{X}|\boldsymbol{\Theta})$ and the estimated model from the data $f(\boldsymbol{X}|\hat{\boldsymbol{\Theta}})$. Each estimator consists of the maximum likelihood estimator (MLE) for the first term and a number of degrees of freedom term for the second term. Therefore, the first term will be relevant when fitting the data, and the second term will penalize a model which is too complex. The AIC tends to overestimate and not be as consistent as MDL/BIC. According to Wax and Kailath, any estimator with the form of $-log(f(\boldsymbol{X}|\boldsymbol{\Theta})) + \alpha(N)k$ is consistent [6]. We can see that the second term of the AIC is not a function of N therefore the AIC tends to overestimate.

$$AIC(k) = -2log(f(\boldsymbol{X}|\hat{\boldsymbol{\Theta}})) + 2k \tag{3}$$

$$MDL/BIC(k) = -log(f(\boldsymbol{X}|\hat{\boldsymbol{\Theta}})) + \frac{1}{2}klog(N) \tag{4}$$

For this study, we follow the procedure in Wax and Kailath, and use a stationary, ergodic, zero-mean multivariate Gaussian distribution for the signal source. We also use a stationary, ergodic, zero-mean, isotropic multivariate Gaussian for the noise. Therefore, the covariance matrix for the noise component is $\sigma^2 I_{pxp}$ meaning the noise has the same variance in all directions and only one parameter is needed to account for the noise. Our sources of data are real numbers and not complex as was the development in Wax and Kailath. In order to appy the AIC and MDL/BIC criteria in equations 3 and 4 we need to count the number of free model parameters for our real valued model. The number of parameters will be the number of degrees of freedom in the $\hat{\boldsymbol{\Theta}}$ vector shown in equation 5. There are k eigenvalues, 1 component for the isotropic noise variance and k eigenvectors. However, each eigenvector is normalized and orthogonal. Therefore, we lose k degrees of freedom for normalization and $1/2k(k-1)$ degrees of freedom for orthogonalization. This means the number of degrees of freedom to estimate using the ITC is $k+1+pk-k-1/2k(k-1) = k(p+1/2(1-k))+1$ degrees of freedom in total. If we apply the AIC and BIC/MDL criteria to the real valued Gaussian sources, we arrive at equations 6 and 7. Assuming the Gaussian model fits the data well, the minimum of the AIC/BIC/MDL functions will estimate the number of non-zero eigenvalues of the data such as equation 8. We implemented both equations in Matlab and as a test arrived at the graphs in figures 1 and 2. This simple test confirmed our implementation by generating a 25 dimensional Gaussian with zero noise, high correlation and 1000 samples. We can see that this test chooses a conservative model order by looking at the eigenvalues of the data as in figure 3. The AIC choose $k^* = 18$ and MDL choose $k^* = 15$. From the eigenvalue plot we can see that choice is just about when the curve flattens out completely.

If we study the equations in 6 and 7, we can see the first term is the log of the Gaussian MLE and the second term is the bias penalty term to keep the model from being too complex. The MLE terms are the ratio of the geometric mean to the arithmetic mean. There is a well known theorem that proves that the geometric mean is always less than or equal to the arithmetic mean with equality holding when all the values of l_i are equal to each other. This means the first term will be zero if the data is isotropic and in the isotropic case, the AIC/MDL/BIC graphs will be parabolically increasing with respect to k giving $k^* = 1$. This is intuitively correct because there is only one non-zero eigenvalue with multiplicity p. In other words, you only need to estimate the variance in one direction to know the variance in all directions in the isotropic case with no data correlation.

$$\hat{\boldsymbol{\theta}} = (\lambda_1, \lambda_2, \ldots, \lambda_k, \sigma^2, \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k) \tag{5}$$

$$\begin{aligned} AIC(k) = -2log\left(\frac{\prod_{i=k+1}^{p} l_i^{1/(p-k)}}{\frac{1}{p-k}\sum_{i=k+1}^{p} l_i}\right)^{(p-k)N} \\ + 2k(p + \frac{1}{2}(1-k)) + 1 \end{aligned} \tag{6}$$

$$\begin{aligned} MDL/BIC(k) = -log\left(\frac{\prod_{i=k+1}^{p} l_i^{1/(p-k)}}{\frac{1}{p-k}\sum_{i=k+1}^{p} l_i}\right)^{(p-k)N} \\ + \frac{1}{2}(k(p + \frac{1}{2}(1-k)) + 1)log(N) \end{aligned} \tag{7}$$

$$\begin{aligned} \hat{k}^* = \min_{0<k<p-1}(AIC/MDL/BIC(k)) \\ \approx |\{\lambda_1, \lambda_2, \ldots, \lambda_q, \sigma^2, \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\}|. \end{aligned} \tag{8}$$

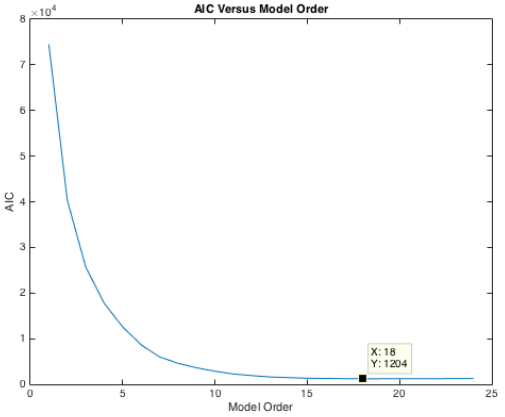

Fig. 1: AIC Curve for multivariate Gaussian $p = 25$, Correlation $\rho = 0.99$ and 1000 Samples. Minimum is annotated at $k^* = 18$.

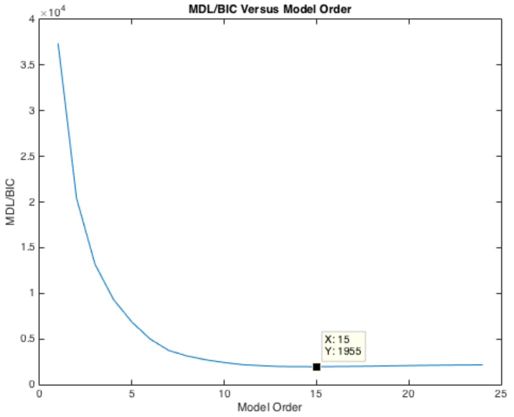

Fig. 2: MDL Curve for multivariate Gaussian $p = 25$, Correlation $\rho = 0.99$ and 1000 Samples. Minimum is annotated at $k^* = 15$.

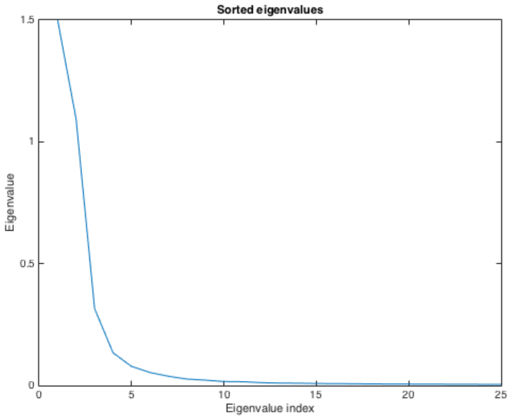

Fig. 3: Zoomed in plot of Eigenvalues for multivariate Gaussian $p = 25$, Correlation $\rho = 0.99$ and 1000 samples.

II. Data Generation

In order to generate test data for our empirical study we use the multivariate generalized power exponential family of distributions from the paper by Gomez, et. al. [2]. We reproduce the equation for data generation from their paper in equations 9 and 10. The equation is parameterized by $\boldsymbol{\Theta} = \{\boldsymbol{u}, \boldsymbol{\Sigma}, \boldsymbol{\beta}\}$ where we set the mean $\boldsymbol{u} = \boldsymbol{0}$. We also have an algorithmic way to compute $\boldsymbol{\Sigma}$ where the covariance values are a function of the distance from the center diagonal which is parameterized by ρ. If $\rho = 0$, then $\boldsymbol{\Sigma} = I_{pxp}$ and there is no correlation in the data and the data is isotropic. If, we increase $\rho > 0$ then the correlation increases with an exponential decay off the diagonal. Figure 4 is a surface plot for $\boldsymbol{\Sigma}$ when the dimension of the data is $p = 50$ and the correlation parameter $\rho = 0.99$. Also, figure 5 shows the resulting Gaussian with correlation $\rho = 0.85$. Finally, if $\beta = 1$, then the data generated is Gaussian distributed. There is an example of this for 2 dimensional data in figure 6. If $\beta < 1$, then the generated data becomes super Gaussian where the kurtosis increases beyond 3. There is an example of super Gaussian for 2 dimensional data in figure 7. If $\beta > 1$, then the data becomes sub Gaussian where the kurtosis decreases below 3. An example of sub Gaussian for 2 dimensional data is in figure 8. For robustness testing, we vary the kurtosis in order to make the data non-Gaussian. In addition, we decided to try two other multivariate distributions. The first was random and the second was a Beta distribution with the parameters $\alpha = \beta = 0.5$. This Beta distribution disperses the density to the tails of the distribution making it highly non-Gaussian. An example of the Beta distribution in 2 dimensions is in figure 9.

$$f(\boldsymbol{x};\boldsymbol{u},\boldsymbol{\Sigma},\boldsymbol{\beta}) = \\ k|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}[(\boldsymbol{x}-\boldsymbol{u})^{H}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{u})]^{\beta}\right\} \qquad (9)$$

$$k = \frac{n\Gamma(\frac{n}{2})}{\pi^{\frac{n}{2}}\Gamma(1+\frac{n}{2\beta})2^{1+\frac{n}{2\beta}}} \qquad (10)$$

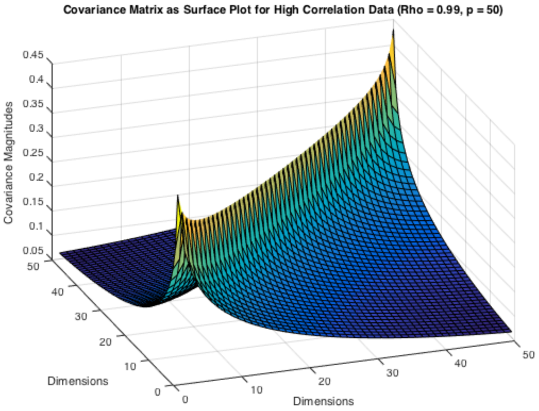

Fig. 4: Surface Plot of Covariance Matrix of High Dimension Gaussian Data Generation Model Using $p = 50, \rho = 0.99, \beta = 1$

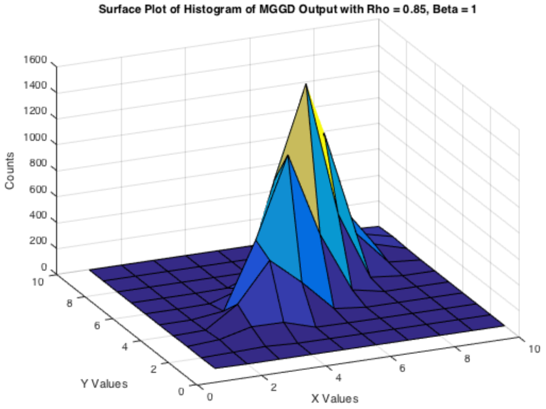

Fig. 5: Surface Plot of Histogram With $p = 2$, $\rho = 0.85$, $\beta = 1$, $10,000$ samples

III. Experiments

For all experiments, we generated the entire data set only once meaning that for the experiment where the sample size N varies, we generate the data and then take subsets of that data. This was done so that we can see how adding IID samples to the same set of samples changes the AIC and MDL/BIC. If you generate the data on each run for increasing samples sizes, the graphs of the model selection criteria are harder to draw conclusions from. This experimental design choice does not change the conclusions which could be drawn, but makes them more apparent in the graphs.

A. Consistency of Criteria

We conducted several experiments to determine how the AIC and MDL/BIC criteria performed when increasing the number of samples. The first experiment was with a model order of $p = 25$ with a minimum of 25 to 25,000 samples. The second experiment was with a model order of $p = 50$ and 50 to 50,000 samples. Both the experiments were conducted with $\beta = 1$ meaning that we generated Gaussian data to match

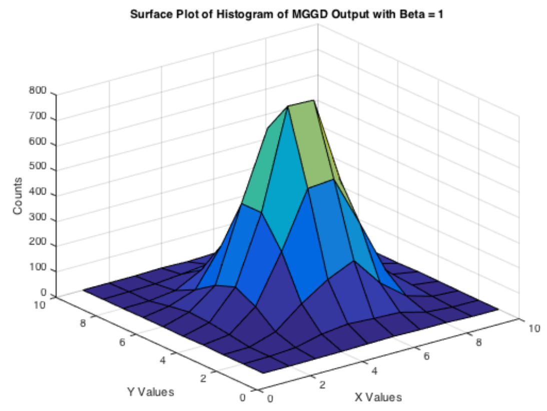

Fig. 6: Surface Plot of Histogram With $p = 2$, $\rho = 0$, $\beta = 1$, $10,000$ samples

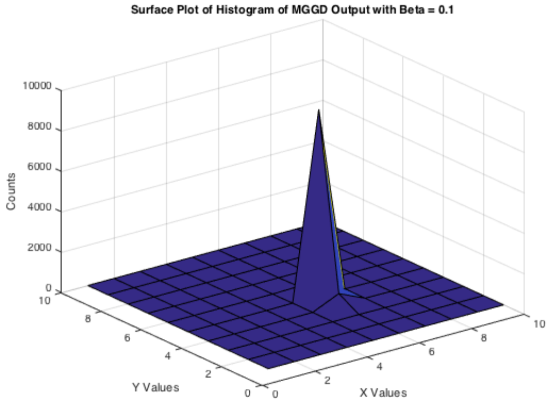

Fig. 7: Surface Plot of Histogram With $p = 2$, $\rho = 0$, $\beta = 0.1$, $10,000$ samples

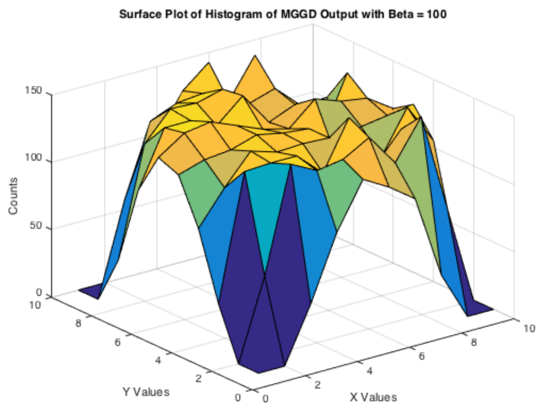

Fig. 8: Surface Plot of Histogram With $p = 2$, $\rho = 0$, $\beta = 100$, $10,000$ samples

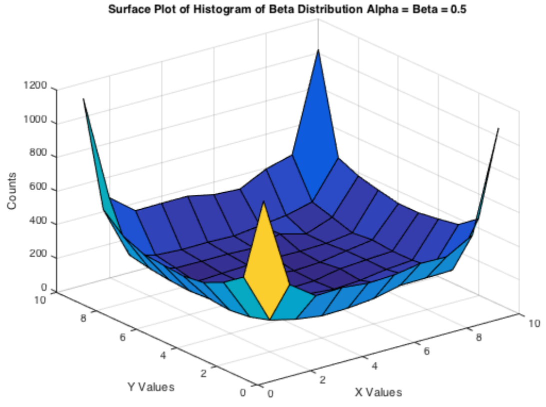

Fig. 9: Surface Plot of Histogram of Beta Distribution $p = 2$, $\alpha = \beta = 0.5$

the assumed model. Our goal is to investigate the consistency of AIC/MDL/BIC as the number of samples increases, how correlation effects the criteria and how the dimensionality of the data effects the criteria. Also, we are concerned with how AIC compares with MDL/BIC in regards to convergence and consistency. The graphs for experiment one are in figures 10 and 11. The graphs for experiment two are in figures 12 and 13.

B. *Robustness from Gaussian Assumption*

We conducted more experiments to determine how the AIC and MDL/BIC criteria performed when the underlying data violates the Gaussian assumed model. In these experiments we fix $p = 25$, the number of samples varies from $N = 25$ to $25,000$ and $\rho = \{0.1, 0.99\}$ for both high and low correlation. We examine the behavior of the AIC/MDL/BIC criteria when the number of samples increases and the correlation parameter for data generation in equation 9, β, is not 1. Therefore, we are introducing higher and lower kurtosis into the generated data. The graphs for these experiments are in figures 14, 15, 16 and 17. We did two different $\rho = \{0.1, 0.99\}$ values to determine if correlation had an effect when the model assumption was violated. In addition, we generated data using a Random distribution and a Beta distribution to study an example of other distributions which are highly non-Gaussian. The Random and Beta distributions we generated were symmetric and therefore the Gaussian it fit was always isotropic, therefore k^* was always 1 meaning there was no correlation in the Gaussian data. This is why we did not graph the results of those test.

IV. Conclusions

A. *Consistency as Number of Samples Increases*

From the first experiment where we generate Gaussian data with parameters of $p = 25, \beta = 1, N = [25, 25000]$ for both AIC and MDL/BIC in figures 10 and 11, we can see that the AIC overestimates when compared to MDL/BIC which was sited as a problem in the Wax and Kailath paper [6]. We also see that the AIC has more variability in the limit as $N \to \infty$. We expect some variability because the estimators converge in probability. Therefore, when we run enough tests we should see some fluctuation. However the AIC seems to fluctuate more with compared to MDL/BIC. In addition, we can clearly see that the lower the correlation in the data, the more samples

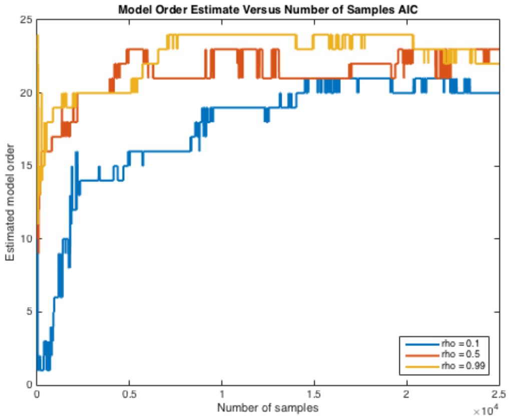

Fig. 10: AIC Model Order Estimate Versus Number of Samples for Lower Dimensional Model $p = 25$, $\beta = 1$, $N = 25$ to $25,000$ and Varying Correlation

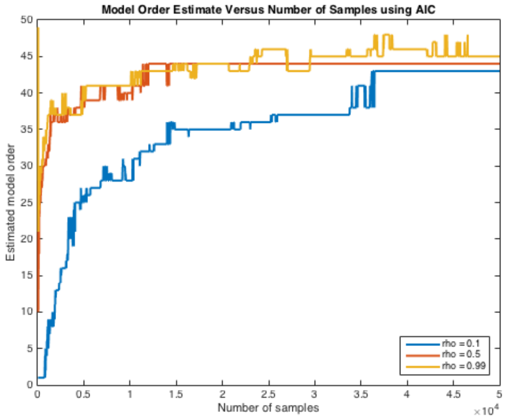

Fig. 12: AIC Model Order Estimate Versus Number of Samples for Higher Dimensional Model $p = 50$, $\beta = 1$, $N = 50$ to $50,000$ and Varying Correlation

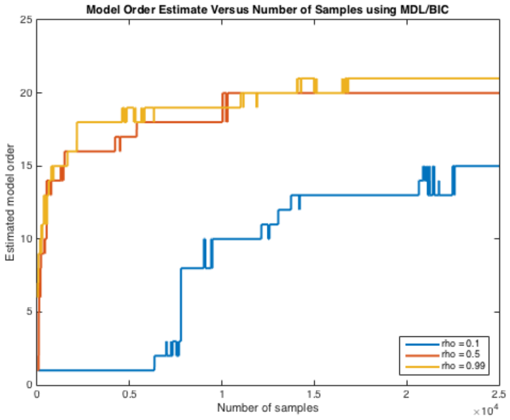

Fig. 11: MDL Model Order Estimate Versus Number of Samples for Lower Dimensional Model $p = 25$, $\beta = 1$, $N = 25$ to $25,000$ and Varying Correlation

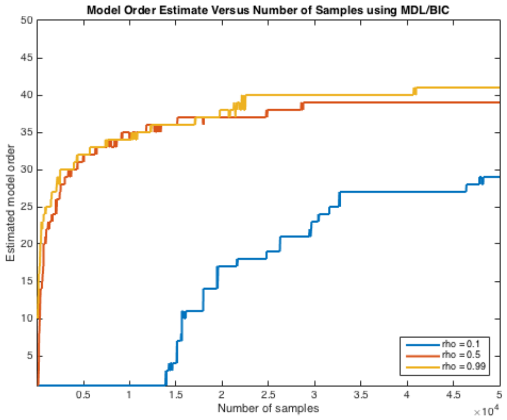

Fig. 13: MDL Model Order Estimate Versus Number of Samples for Higher Dimensional Model $p = 50$, $\beta = 1$, $N = 50$ to $50,000$ and Varying Correlation

that are required for all the criteria to converge. For example, the blue curves in figures 10 and 11 are low correlation and certainly take many more samples to converge and in the MDL case the low correlation data did not converge to the correct solution in $25,000$ samples. This is because the number of samples needed to characterize the low correlation data by the eigenvalues of the covariance matrix is considerable due to the eigenvalues being close in value to each other and random fluctuations will interfere with convergence. In contrast, the highly correlated data converges quite quickly in comparison. We also see the trend that AIC converges faster than the MDL/BIC criteria. However, this is at the cost of variability and overestimation trend mentioned before. AIC captures more subtle correlations in the data faster, but if there are too few samples, the AIC is unreliable which is shown when there are very few samples and the AIC graph peaks to a value of p.

We were also interest in how the dimensionality of the data samples effected the convergence and therefore, we ran the same experiment using $p = 50, \beta = 1, N = [25, 25000]$. The results are graphed in figures 12 and 13. We see the same trends as the first experiment but with one added observation that the number of samples needed for the model selection criteria to converge when doubling the dimensionality increases considerably. This is intuitive and follows the well known curse of dimensionality problem. As we increase the dimension of the data, we need an exponential number of samples for the statistics computed from the data to converge.

B. Robustness as Gaussian Assumption is Violated

The next experiment we ran was to change the generated data away from a Gaussian to a sub and super Gaussian by modifying kurtosis away from 3. The results are graphed for both a low correlation case, in figures 14 and 15, and a high correlation case 16 and 17. From the graphs for the lower correlation, you can see that as we change the β parameter away from 1, the estimates for the model order tend to vary considerably. When $\beta = 1$ the data is Gaussian and we were sure to plot the Gaussian case with the Non-Gaussian cases. As you can see there is more fluctuation, but the estimator does not deviate wildly. This is attributed to the fact that we

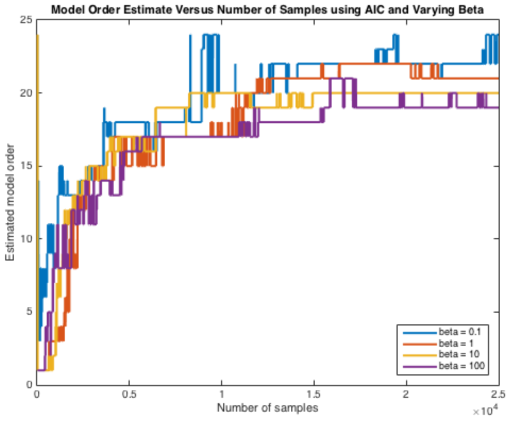

Fig. 14: AIC Model Order Estimate Versus Number of Samples for Lower Dimensional Model $p = 25$, Lower Correlation $\rho = 0.1$, $N = 25$ to $25,000$ and Varying Beta $\beta = 0.1, 1, 10, 100$.

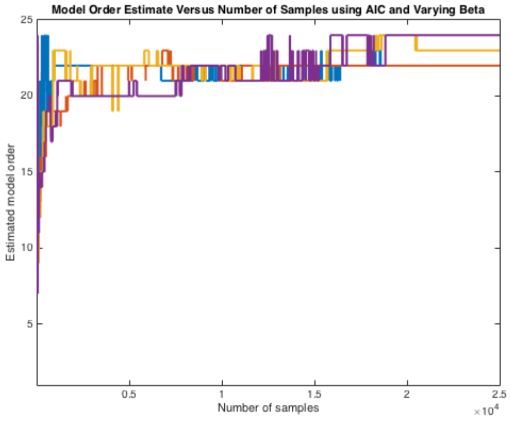

Fig. 16: AIC Model Order Estimate Versus Number of Samples for Lower Dimensional Model $p = 25$, Higher Correlation $\rho = 0.99$, $N = 25$ to $25,000$ and Varying Beta $\beta = 0.1, 1, 10, 100$.

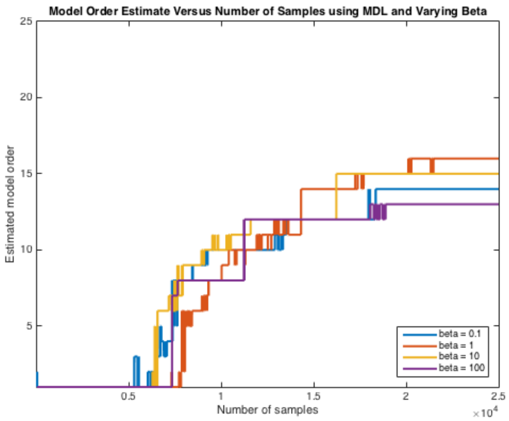

Fig. 15: MDL Model Order Estimate Versus Number of Samples for Lower Dimensional Model $p = 25$, Lower Correlation $\rho = 0.1$, $N = 25$ to $25,000$ and Varying Beta $\beta = 0.1, 1, 10, 100$.

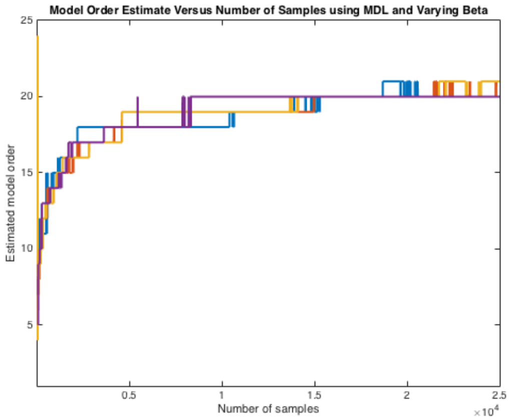

Fig. 17: MDL Model Order Estimate Versus Number of Samples for Lower Dimensional Model $p = 25$, Higher Correlation $\rho = 0.99$, $N = 25$ to $25,000$ and Varying Beta $\beta = 0.1, 1, 10, 100$.

are still generating data which is from the Gaussian family. In the highly correlated data experiment in figures 16 and 17 we see just as before that the model selection criteria converges faster and we again see the fluctuations as in the lower correlation Non-Gaussian experiment. However, there is one new observation in figure 17. If you look at the graph closely, you can see that when there are few samples and high correlation, the MDL/BIC criteria varies widely. This is because we don't have enough samples and the Non-Gaussian data is not generating data that is isotropic. Therefore, the MDL/BIC may compute a very high model order erroneously as it did in this case.

We also compute graphs for the AIC and MDL/BIC criteria with a random distribution and a symmetric Beta distribution. However, we did not provide the resulting graphs here because the output of both AIC and MDL/BIC was a constant 1. This is because the Gaussian model we assumed was trying to fit symmetric data, which means the Gaussian that best fit was isotropic. Clearly this is an inadequate model as the KL distance between a Gaussian and Random or Beta distribution is very high and a model order of 1 would not suffice to capture any of the underlying signals. We conclude that the model selection criteria in this paper is inadequate if the data is significantly different that the assumed model and the criteria can vary wildly if the data is undersampled and this is especially bad in higher dimensional models and samples.

References

[1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, Dec 1974.

[2] E Gomez, MA Gomez-Viilegas, and JM Marin. A multivariate generalization of the power exponential family of distributions. *Communications in Statistics-Theory and Methods*, 27(3):589–600, 1998.

[3] A.P. Liavas and P.A. Regalia. On the behavior of information theoretic criteria for model order selection. *Signal Processing, IEEE Transactions on*, 49(8):1689–1695, Aug 2001.

[4] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[5] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[6] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(2):387–392, Apr 1985.

Bernard Lampe (M'09) became an IEEE Member (M) in 2009 and received his bachelors of science degree from The University of Michigan in Ann Arbor, Michigan, USA in 2009.

Mr. Lampe is also a member of the American Society for Computing Machines (ACM) since 2009.

```
function [aic, mdl] = model_order(data)

% data: observations vectors assumings each ...
    row is a vector
% K: max model order to compute
% return: aic, bic and mdl information ...
    criteria vectors for 1:K model order

% compute the correlation matrix
[p, n] = size(data);
R = data * data'./n;

% compute the eigenvectors
e = sort(eig(R), 'descend');

% compute information criteria for all ...
    possible k = 1:p-1
aic = zeros(p-1, 1);
mdl = zeros(p-1, 1);
for k = 1:p-1
    aic(k) = gauss_aic(e(k+1:p), k, n, p);
    mdl(k) = gauss_mdl(e(k+1:p), k, n, p);
end

% aic function
function aic = gauss_aic(l, k, n, p)
    aic = -2 * (p-k) * n * log(geomean(l) / ...
        mean(l)) + 2 * k * (p + 0.5 * (1 - ...
        k)) + 1
end

% mdl function
function mdl = gauss_mdl(l, k, n, p)
    mdl = -(p-k) * n * log(geomean(l) / ...
        mean(l)) + 0.5 * k * (k * (p + 0.5 * ...
        (1-k)) + 1) * log(n);
end

end
```