SAND2023-12113C

Sandia National Laboratories

Statistical Properties of Compression Analytics (CA)

Kurtis Shuler*, Alexander Foss*, Travis Bauer+, Richard Field‡, Christina Ting+

*Department of Statistical Sciences, 05573, +Computational Decision Science, 05554, ‡Data Science & Applications 05553

Project Goals:

- Improve CA through better understanding of its underlying statistical properties
- Advance CA through development of "non-local" compression techniques that can identify and utilize nonadjacent structure present within a byte stream for better ML performance

Applications:

- Text analysis, detection of disinformation campaigns, DNA/RNA sequence analysis, cybersecurity

What is Compression Analytics?

- **CA uses compression to achieve machine learning tasks:** Anomaly detection, classification, clustering, etc.
- File compression algorithms used to measure similarity or shared information across items

Advantages of CA

- **CA is "featureless" in the sense that it can be applied to any byte stream**
- Allows rapid model development and prototyping with **minimal analyst intervention, even on unstructured data**
- Many open source and COTS compression tools readily available

CA for Classification

- Consider a classification problem with M different classes
- The goal is to assign class label *m*∈{1,...,M} to a new, unlabeled byte stream
- For each class a different compression model will be produced ("trained") using the labeled training data
- For the new byte stream, assign a class label based on which of the *M* compression models compresses it best

CA and Model Likelihood

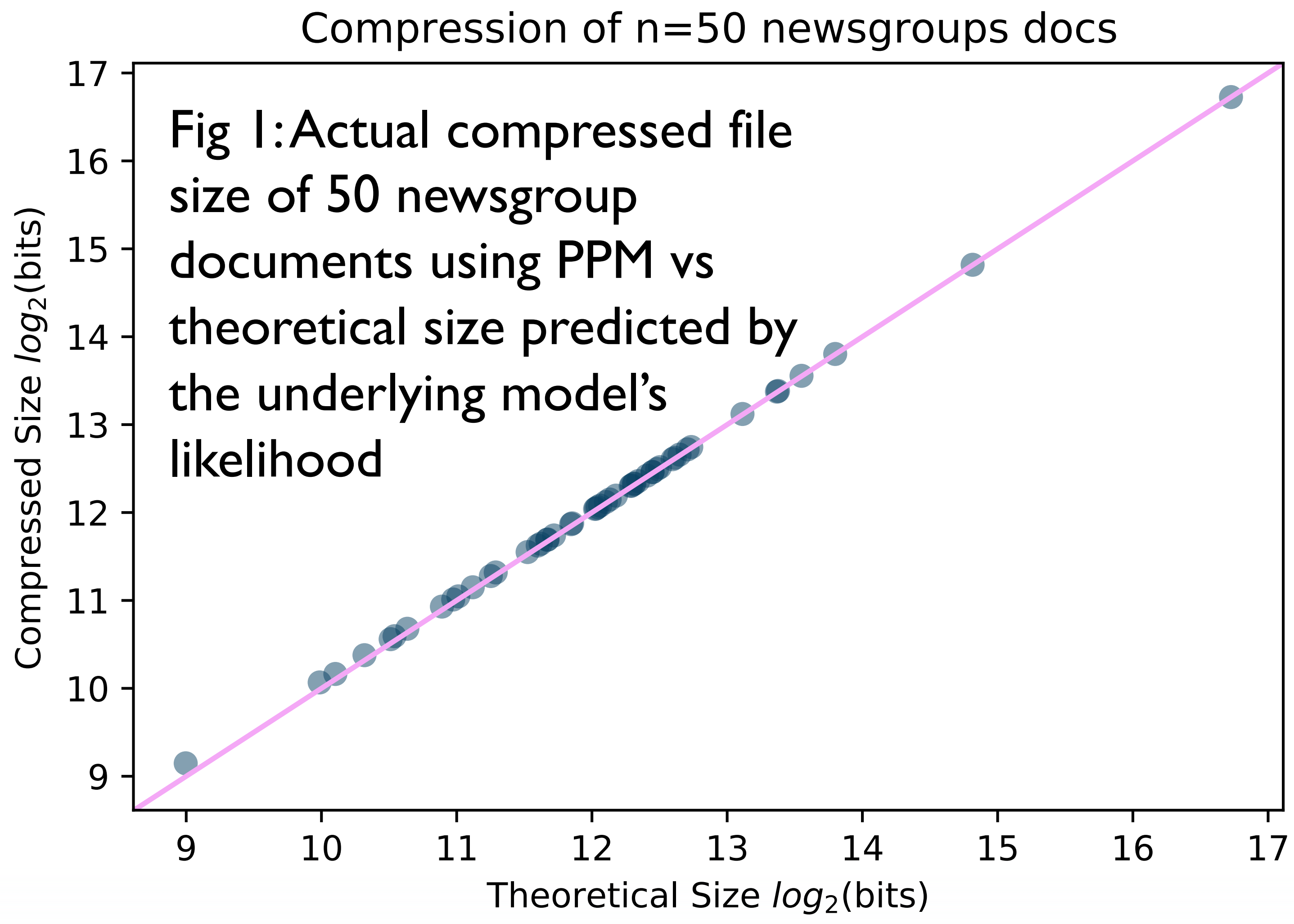

Fig 1: Actual compressed file size of 50 newsgroup documents using PPM vs theoretical size predicted by the underlying model's likelihood

- Entropy encoders use statistical models to compress files by encoding more frequently occurring symbols with a smaller number of bits
- A lower bound for the number of bits required to encode a stream is related to the stream's entropy, and is given by Shannon's source coding theorem
- We show for this class of compression algorithms the **classification rule is equivalent to choosing *m* to be the class corresponding to the compression/statistical model with the highest likelihood**

Proposed Method: Model Selection

- KL divergence gives the expected number of extra bits required to encode a stream when the true (typically unknown) byte stream distribution is f(x), but is encoded using g(x|θ)

$$I(f,g) = \sum_x f(x) \log_2 \left(\frac{f(x)}{g(x|\theta)} \right)$$

- Akaike Information Criterion (AIC) provides an approximately unbiased estimator of expected relative KL information
- **For entropy encoders, model selection criteria can be used to select an optimal compression algorithm**
- Figures 2 and 3 demonstrate this using 100 training documents and 100 test documents from the 20 newsgroups dataset

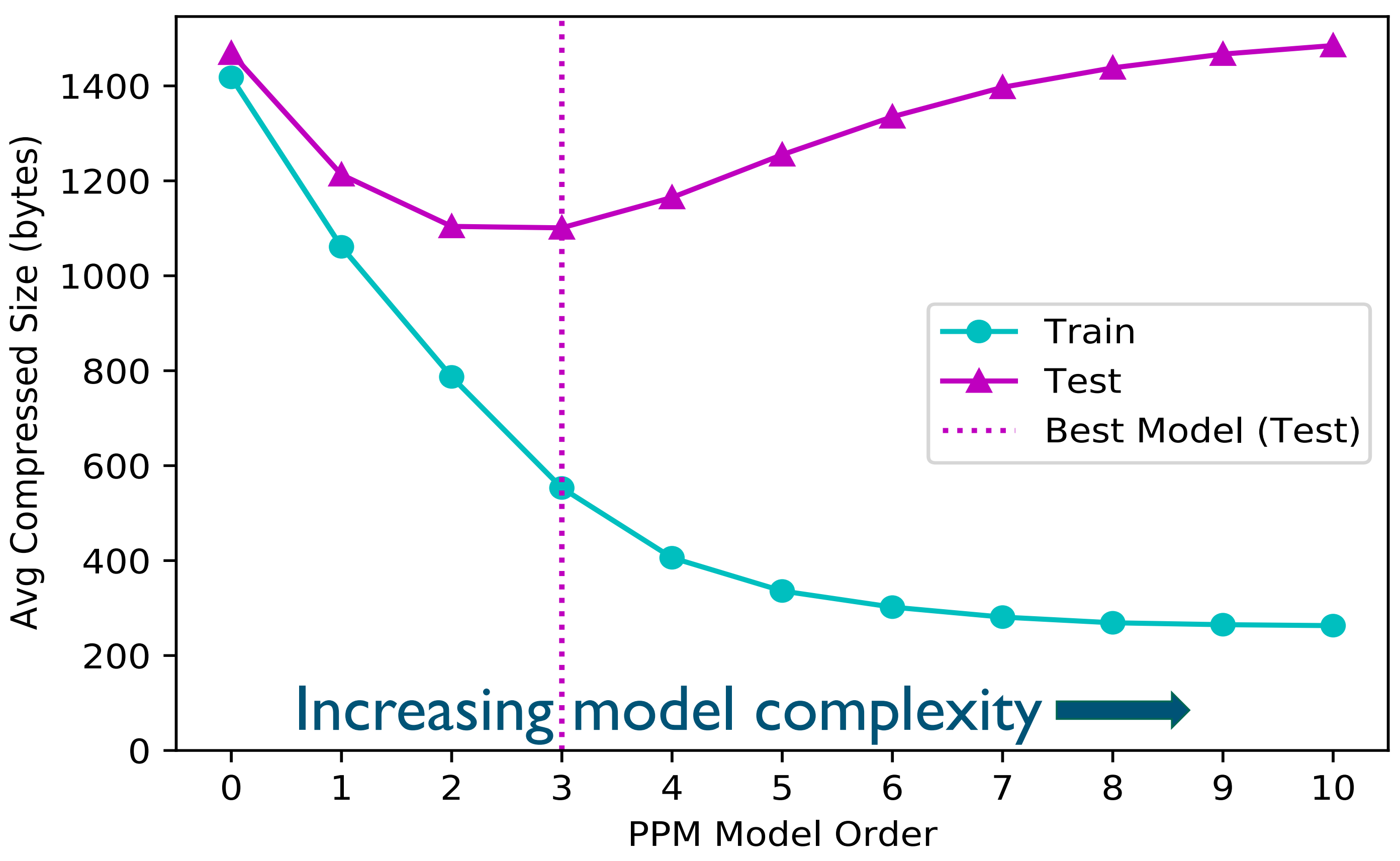

Fig 2: Bias/variance tradeoff in file compression size

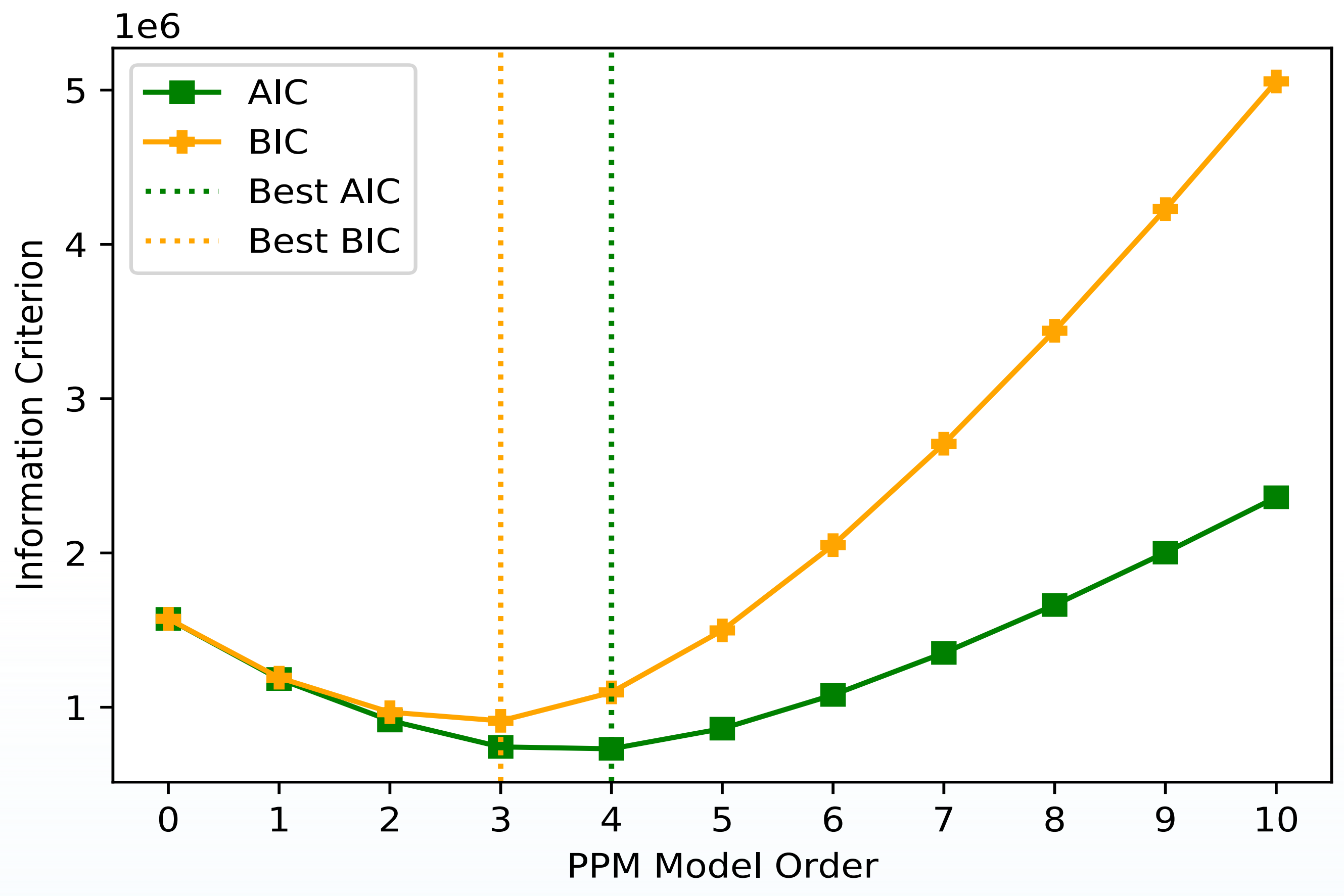

Fig 3: Model selection criterion for different PPM models

Statistical interpretation of CA yielded novel model selection technique with promising initial results on real-world data

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND No. ___________

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.