Bayesian event categorization matrix approach for explosion monitoring

Koermer, Scott Carl
Carmichael, Joshua Daniel
Williams, Brian J.

To be published in: Geophysical Journal International

DOI to publisher's version: 10.1093/gji/ggaf171

Bayesian Event Categorization Matrix Approach for Explosion Monitoring

Scott Koermer *[1], Joshua D. Carmichael[2], and Brian J. Williams[1]

[1] *Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, NM, USA*

[2] *National Security Earth Science Group, Los Alamos National Laboratory, Los Alamos, NM, USA*

June 16, 2025

Abstract

Current efforts to correctly categorize natural events from suspected explosion sources with data that is collected by ground- or space-based sensors presents historical challenges that remain unaddressed by the Event Categorization Matrix (ECM) model. Smaller historical events (lower-yield explosions) may have data available from fewer measurement techniques than are available today, and therefore, a historical event record can lack a complete set of discriminants. The covariance structures can also differ between such observations of event (source-type) categories. Both obstacles are problematic for the classic Event Categorization Matrix model. Our work addresses this gap and presents a Bayesian update to the previous Event Categorization Matrix model, termed the Bayesian Event Categorization Matrix model, which can be trained on partial observations and does not rely on a pooled covariance structure. We further augment the Event Categorization Matrix model with Bayesian Decision Theory so that false negative or false positive rates of an event categorization can be reduced in an intuitive manner. To demonstrate improved categorization rates for the Bayesian Event Categorization Matrix model, we compare an array of Bayesian and classic models with multiple performance metrics using Monte Carlo experiments. We use both synthetic and real data. Our Bayesian models show consistent gains in overall accuracy and lower false negative rates relative to the classic Event Categorization Matrix model. We propose future avenues to improve Bayesian Event Categorization Matrix models' decision making and predictive capability.

Keywords— Bayesian inference, Monte Carlo methods, Probability distributions, Statistical methods

*Corresponding author, skoermer@lanl.gov

1 Introduction

Statistical methods have historically supported monitoring signatures of suspected conventional and nuclear explosions (Bowers and Selby, 2009; Anderson et al., 2010b). Explosion monitoring researchers in particular leverage such methods to confirm or challenge the hypotheses that some geophysical events present evidence of a detonation rather than natural processes (Council et al., 2012; McGrath, 2009). Such methods have been crucial in seismic source identification, that is, statistical methods that screen explosion-sourced records of seismic activity from records expected from earthquakes or other processes.

Early quantitative work (Booker and Mitronovas, 1964) that developed classification methods to separate explosions from earthquake populations (discrimination) in the 1960s justified later efforts to rigorously defend test-ban treaties (Ericsson, 1970). Some of these treaties were only aspirational at the time that the geophysical work was completed (Myers, 1972; Elvers, 1974). More modern efforts from researchers such as Shumway (Shumway and Rivers, 1984; Shumway, 1996; Shumway, 2001), Anderson (Anderson et al., 2010a; Anderson et al., 2010b; Anderson et al., 2014; Anderson, 2009), and Jih (Jih et al., 1990) have further advanced these statistical methods beyond discrimination. Such newer methods often ingest multiple discriminants (Fah and Koch, 2002; Anderson et al., 2007) or other modalities (e.g. infrasound, electromagnetic, and even vegetation (Redman et al., 2019)) in statistical tests that screen all-source explosions from earthquakes. Some recent research has continued the trend to use data integration or fusion methods to more confidently detect (Scoles, 2020; Carmichael et al., 2020; Carmichael et al., 2016), identify (Taylor, Arrowsmith, and Anderson, 2010; Arrowsmith and Taylor, 2013), and characterize (Ford et al., 2021a; Ford et al., 2014; Green et al., 2013; Williams et al., 2021) populations of explosions from other events, and thereby reduce false positive rates. This effort continues to focus on smaller, evasively conducted explosions (Koper, 2020; Rodd et al., 2023).

One such multi-discriminant, statistical method that supports explosion monitoring is called the Event Categorization Matrix (ECM, Anderson et al., 2007). This method has historically been used to test if multiple observations associated with a single event support the hypothesis that their source was a conventional or nuclear detonation (Maceira et al., 2017). The ECM method currently consumes seismic discriminants such as event depth (Arora, Basu, and Krishnan, 1983; Murphy, Barker, and Rodi, 2002), ratios of body-wave and surface-wave magnitudes (Stevens and Day, 1985), so-called depth phases (Warren and Shearer, 2005), and ground motion polarity (Brillinger, Udias, and Bolt, 1980). Some variants of ECM also leverage more novel discriminants, such as infrasound phase arrivals (Park, Hayward, and Stump, 2018) and teleseismic waveform complexity factors (Anderson and Taylor, 2008). Ongoing effort may integrate new discriminants, such as the shape of Rayleigh waveform radiation patterns (Carmichael, 2021) and the presence of high frequency, crustal surface waves like *Rg* (Kafka, 1990) into future variants of ECM.

The ECM method assumes that a vector of discriminant observations can be modeled as a random variable from a multivariate normal distribution, with a mean and covariance matrix specific to a single event category. The ECM method uses previous observations with known event categories (ground truth data) to estimate mean and covariance parameters for each event category distribution, applying regularized discriminant analysis (RDA, Friedman, 1989) to estimate covariance matrices for small data sets. The ECM model then categorizes a new observation with a series of hypothesis tests that are based on typicality indices (McLachlan, 2005) and quantifies the likely set membership of the new, uncategorized observation to each candidate event group. If a new observation is atypical of all categories (with the exception of the detonation category), its source is then categorized as an explosion.

Technologies to monitor for nuclear explosions have historically leveraged multiple sensor net-

works such as the VELA satellites (Wright and De Geer, 2017) and seismic arrays (Ringdal and Husebye, 1982). However, research to routinely fuse such multi-modal, simultaneous observations and improve event categorization accuracy remains ongoing (Herzog, 2017; Kalinowski et al., 2023). Both physical and mathematical issues make it challenging to implement ECM with multiple modalities (Anderson et al., 2007).

Secondly, ECM uses a covariance estimator that can lead to model mis-specification. This is because events that produce multi-modal signals can have drastically different covariance structures. For example, high-yield, aboveground nuclear detonations will produce optical signals with high irradiance that covaries with large-amplitude seismic waveforms (Ford et al., 2021b). This does not imply that a nuisance event that produces a high irradiance signal will also produce seismic waveforms with significant amplitude. Therefore, although a conventional or nuclear detonation may produce covarying discriminants, nuisance events may not.

Lastly, ECM may categorize an event for non-intuitive reasons. This is particularly true when the number of event categories considered grows. When ECM then fails to reject a new observation from multiple event categories, the model declares the event indeterminate, requiring human intervention for categorization (Anderson et al., 2007). A monitoring agent must then explore strategies to empirically reduce the observed number of false negatives without significantly increasing the number of indeterminate categorizations or the overall categorization accuracy. Anderson et al.'s (2007) approach relies on a set of classical statistical methods for hypothesis testing and parameter estimation, and some of these choices can result in deficiencies. Subsequent references to the ECM model provided in Anderson et al. (2007) are termed classical ECM (C-ECM), due to Anderson et al.'s use of classical statistical methods.

To address some of the problems involved in fusing discriminants from multiple modalities, we develop an ECM model that uses Bayesian methods (B-ECM) for parameter estimation and decision criteria. This novel work uses Bayesian decision theory and treats missing data with a matrix t-distribution within a Bayesian Normal mixture model (Stephens and Phil, 1997). We evaluate integrals analytically when possible to avoid the computational expense of multi-dimensional numerical integration. These collective advances then select an event category for a new observation within practical time constraints and without supervision. We utilize a Bayesian typicality index to detect if a new observation is inconsistent with the event categories used for training (under the uncertainty imparted by using training data with missing elements). Lastly, we demonstrate the accuracy of the B-ECM methodology against a carefully prepared, curated dataset with missing entries (Fig. 1), which C-ECM has previously been demonstrated against (Anderson and Taylor, 2008). This establishes a baseline and measures gains in accuracy with the same curated dataset.

We organize this report as follows: Section 2 summarizes the statistical methods assimilated to create the B-ECM methodology. Section 3 explains how these methods are implemented in codes and algorithms. Section 4 details results from a series of Monte Carlo (MC) experiments that compare the performance of the B-ECM models to the C-ECM model. Finally, section 5 provides a discussion of the results, their implications, and potential avenues for further improvements to event categorization models.

2 Statistical Basis for the Decision Framework

B-ECM is a complete framework for exploiting geophysical discriminants to train a model, evaluate a new event with the model, and decide the new event category by combining model output with a formalized system for including subjective reasoning. We use Bayesian inference (Section 2.1) throughout this work to provide a statistically consistent framework for propagating uncertainties

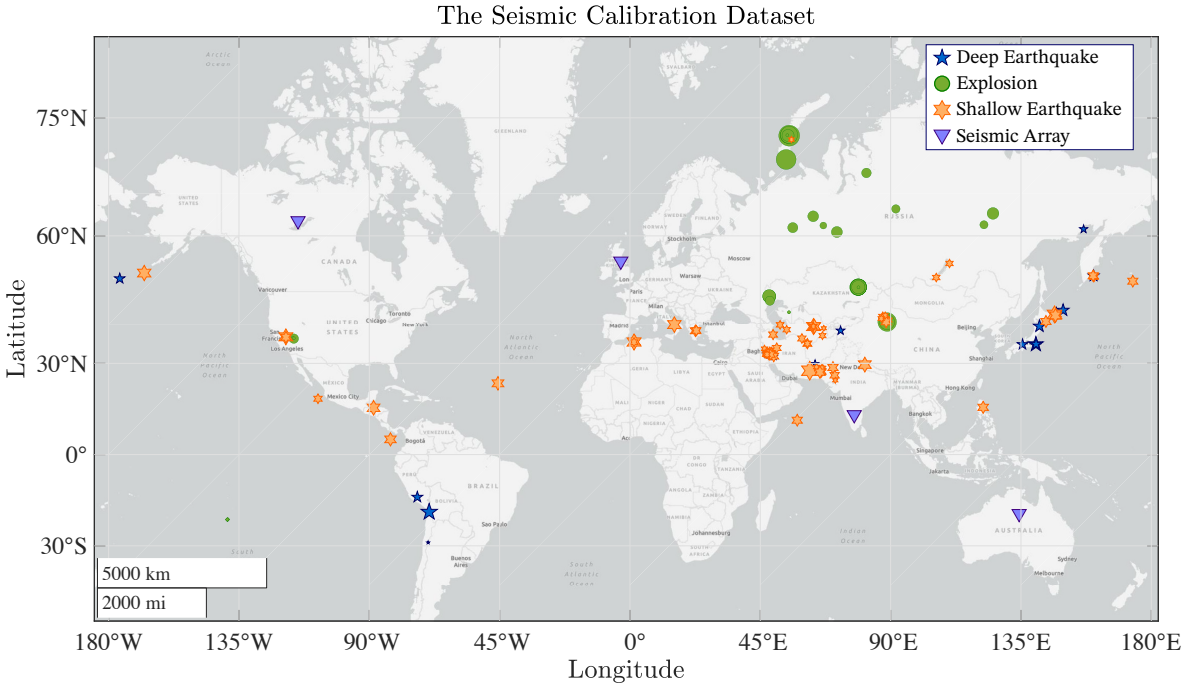


Figure 1: The source locations for deep earthquakes (DEQ), explosions (EX), and shallow earthquakes (SEQ) that we used to calibrate C-ECM and B-ECM. Many sources were recorded from historical nuclear tests, and their size is proportional to an exponent raised to the value of a seismic magnitude (magnitude range: 3 to 6.9). Explosions with entirely missing magnitude data are marked with a "magnitude 3" size marker; this provides a crude estimate for the global magnitude of completeness in most continental regions. The upside triangles mark locations of the four, historical seismic arrays that provided the bulk of the discriminant data.

from training data through the decision to categorize a new event as a detonation or nuisance event. Our statistical model (Section 2.2) leverages Bayesian inference to use training data with missing discriminants. Bayesian decision theory (Section 2.3) enables consistent decision making, taking into account category probabilities, the utility of a correct categorization, and the loss of an incorrect categorization. Notation is detailed in Nomenclature Appendix A.

2.1 Bayesian Inference

Bayesian statistics relies on Bayes' theorem (Equation 1) to infer model parameters. Bayes' rule reads as follows: the posterior distribution of ϕ, given data y; $p(\phi|y)$, is equal to the likelihood of the data $p(y|\phi)$ times the prior distribution on the model parameter $p(\phi)$, divided by the marginal likelihood of the data $p(y)$:

$$p(\phi|y) = \frac{p(y|\phi)p(\phi)}{p(y)} \tag{1}$$

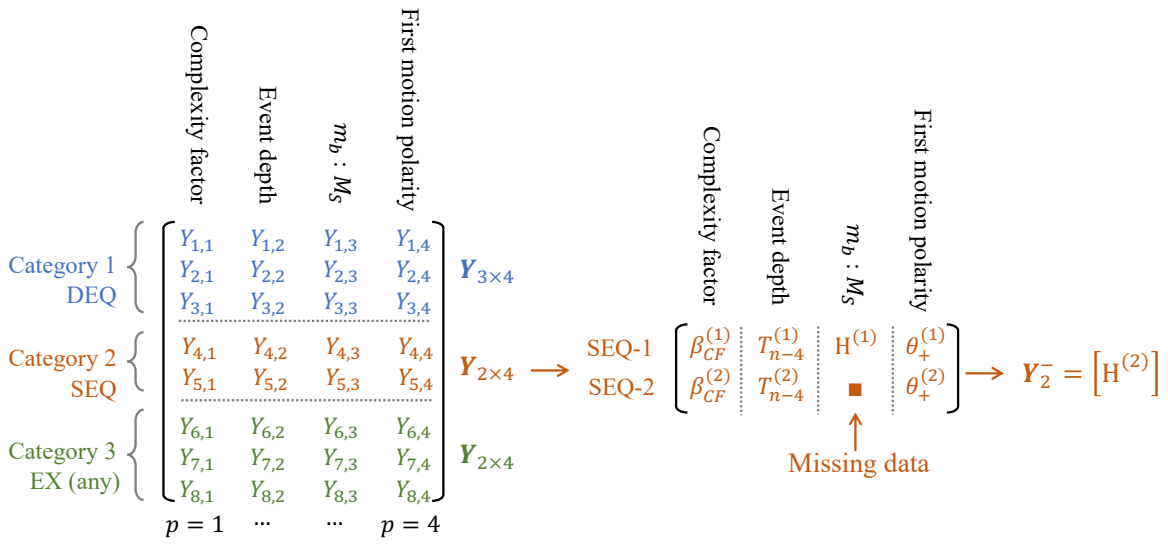
Bayesian inference treats model parameters as random variables, resulting in a probability distribution on ϕ, which quantifies uncertainty. This contrasts to classical methods, which provide a point estimate and quantify uncertainty with confidence regions that are interpreted differently. Computational methods such as the Gibbs sampler often use Markov chain Monte Carlo (MCMC) to infer model parameters by sampling model parameters from the posterior distribution $p(\phi|y)$ (Gelfand et al., 1990; Gelfand and Smith, 1990; Casella and George, 1992; Geman and Geman, 1984). An observer can consider the probabilities for multiple models, given the data, to make data predictions under the Bayesian framework. We use these traits of Bayesian inference to derive B-ECM. Hoff (2009) provides more resources on Bayesian inference, and Robert and Casella (1999) gives more details about MCMC and statistical simulation.

2.2 Bayesian Categorization with Missing Training Data

We choose to categorize data with a methodology that is similar to that in Stephens and Phil (1997). Our method assumes that the event category for each training data event is known, but the group to which a new observation belongs to is unknown. We let $\boldsymbol{Y}_{N\times p}$ be a matrix that contains the entirety of the training data, with N event observations each with p discriminants. Each row of $\boldsymbol{Y}_{N\times p}$ is an observation of the k^{th} event category, where $k \in \{1,\ldots,K\}$ has no uncertainty. We split the training data into the event category-specific matrices $\boldsymbol{Y}_{N_1\times p},\ldots,\boldsymbol{Y}_{N_k\times p},\ldots,\boldsymbol{Y}_{N_K\times p}$. Here, N_k is the number of training observations in the k^{th} event category and $N_1+\cdots+N_k+\cdots+N_K=N$. The p discriminants in each $\boldsymbol{Y}_{N_k\times p}$ are the same and are arranged in the same column order.

We assume that each $\boldsymbol{Y}_{N_k\times p}$ is a realization from a Matrix Normal distribution (Gupta and Nagar, 2018) with an unknown mean $\mathbf{1}_{N_k}\boldsymbol{\mu}_k^\top$, independent rows, and column covariance $\boldsymbol{\Sigma}_k$. Appendix B.1 provides details. A Matrix Normal distributed random variable has infinite support (that is, the interval where the density is not identically zero) on $\mathbb{R}^{N_k\times p}$. Previously, p-values on $(0,1]$ have been used with ECM (Anderson et al., 2007). We choose to transform these values with a logit transformation $\text{logit}(x)=\ln(x)-\ln(1-x)$ so that the transformed values are (instead) unbounded. The arcsine transform used in Anderson et al. (2007) returns values on $[0,1]$ instead of on the entire real number line. This property is a mismatch for the normal distribution, which has support for all real numbers. However, both transforms are available for use in our code.

We integrate the Matrix Normal density of each $\boldsymbol{Y}_{N_k\times p}$ over the prior densities of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ for each k analytically. This reduces the computational burden of inference, as detailed in Appendix B.3. Integration results in the marginal likelihood $p(\boldsymbol{Y}_{N_k\times p}|\boldsymbol{\theta}_k)$, which is a matrix t-distribution



DEQ: Deep Earthquakes
SEQ: Shallow Earthquakes
EX: Explosions

Complexity factor: Taylor and Anderson (2008) (Equation 1)
Event depth: Taylor et al. (2007) (Equation 5)
$m_b : M_S$: Taylor et al. (2007) (Equation 15)
First motion polarity: Taylor et al. (2007) (Equation 17)

Figure 2: A visual description of the training data $\boldsymbol{Y}_{N\times p}$ that populates the B-ECM module. Dimensions $N = 8$ and $p = 4$ in this example. The top, leftmost column matrix organizes data within three categories (deep earthquakes, shallow earthquakes, and explosions) into distinct groups of rows. The four columns enumerate distinct seismic discriminant types (complexity factor, event depth, body wave to surface wave magnitudes, and polarity of first motion). The row matrices $\boldsymbol{Y}_{N_k\times p}$ compartmentalize this training data by category, which we color separately. The sub-matrix $\boldsymbol{Y}_{2\times4}$ $(N_2 = 2, p = 4)$ then stores data observed from shallow earthquakes (SEQ), according to the notation cited at the bottom right. The second shallow earthquake observation lacks a body wave to surface wave magnitude measurement. This missing data then populates the data set $\boldsymbol{Y}_2^-$.

(Gupta and Nagar, 2018) conditioned on a set of prior hyperparameters abbreviated as $\boldsymbol{\theta}_k$. In Bayesian inference, prior hyperparameters specify beliefs about model parameters before data is observed. Hyperparameter values are often chosen such that their influence on posterior inference is significantly reduced when combined with data, but they can be used to strongly encode subjective choices. For brevity, in the main body of this article, we choose to generally refer to hyperparameters for event category k as $\boldsymbol{\theta}_k$ and the set of prior hyperparameters collected over all K event categories as $\boldsymbol{\theta}$. Appendix B.2 details the prior distributions and hyperparameters in $\boldsymbol{\theta}$. Appendix C provides relevant properties of the matrix t-distribution.

Sometimes, an event contained in $\boldsymbol{Y}_{N_k \times p}$ will have less than p recorded discriminants. The unavailable discriminants for such a case are considered *missing data* (Little and Rubin, 2019; Gelman et al., 1995). We let $\boldsymbol{Y}_k^-$ represent the missing data elements from all N_k events in $\boldsymbol{Y}_{N_k \times p}$ and $\boldsymbol{Y}_k^+$ be the recorded elements, respectively. We leave $\boldsymbol{Y}_k^+$ and $\boldsymbol{Y}_k^-$ dimensionless to represent the observed and missing data with generality. We use the chain rule of conditional probability to specify

$$p(\boldsymbol{Y}_{N_k \times p}|\boldsymbol{\theta}_k) = p(\boldsymbol{Y}_k^-, \boldsymbol{Y}_k^+|\boldsymbol{\theta}_k) = p(\boldsymbol{Y}_k^-|\boldsymbol{Y}_k^+, \boldsymbol{\theta}_k)p(\boldsymbol{Y}_k^+|\boldsymbol{\theta}_k). \tag{2}$$

Because of the properties of conditional probability and $p(\boldsymbol{Y}_{N_k \times p}|\boldsymbol{\theta}_k)$, we can obtain a large number of random draws of $\boldsymbol{Y}_k^-$ given $\boldsymbol{Y}_k^+$ and $\boldsymbol{\theta}_k$ from the probability density function $p(\boldsymbol{Y}_k^-|\boldsymbol{Y}_k^+, \boldsymbol{\theta}_k)$ using a combination of conditional formulas (Appendices C and B.4) for the matrix t-distribution and numerical methods. We can think of these random draws, or samples, of $\boldsymbol{Y}_k^-$ from its distribution as similar to obtaining a large number of virtual samples of coin flips from a binomial distribution. We use these samples of $\boldsymbol{Y}_k^-$, combining each sample obtained with observations $\boldsymbol{Y}_k^+$ to exploit partial observations in the training data set and thereby compute more accurate decisions (instead of discarding partial observations).

Next, we pivot to categorizing a new event into an event category. To simplify notation here, we provide results for conditioning densities on the full data set $\boldsymbol{Y}_{N_k \times p}$. If data with missing elements was used for training, we can now think of $\boldsymbol{Y}_{N_k \times p}$ as the combination of $\boldsymbol{Y}_k^+$ with a single draw of $\boldsymbol{Y}_k^-$. However, important details and final derivations for using partial observations in training data are provided in the appendices. Let $\tilde{\boldsymbol{y}}_{\tilde{p}}$ be a new event recorded as a vector of discriminants of length $\tilde{p} \leq p$. Vector $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is associated with the random vector of length K that we call $\tilde{\boldsymbol{z}}_K^\top = [\tilde{z}_1, \ldots, \tilde{z}_k, \ldots, \tilde{z}_K]$. A random realization of $\tilde{\boldsymbol{z}}_K$ is equivalent to a draw from a categorical distribution, where a single element is equal to 1 and the remaining elements are zero. The index, $k \in \{1, \ldots, K\}$, of the non-zero value of $\tilde{\boldsymbol{z}}_K$, corresponds to $\tilde{\boldsymbol{y}}_{\tilde{p}}$ belonging to the k^{th} training event category. The conditional expected value of $\tilde{\boldsymbol{z}}_K$, given all the data and prior parameter specifications, is then equivalent to

$$\mathbb{E}[\tilde{\boldsymbol{z}}_K|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N \times p}, \boldsymbol{\theta}] = [p(\tilde{z}_1 = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N \times p}, \boldsymbol{\theta}), \ldots, p(\tilde{z}_K = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N \times p}, \boldsymbol{\theta})]. \tag{3}$$

Equation 3 is equivalent to a vector that specifies the probability of $\tilde{\boldsymbol{y}}_{\tilde{p}}$ belonging to each of the K event categories. For the k^{th} event category, $p(\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N \times p}, \boldsymbol{\theta})$ can be evaluated using Bayes' rule.

$$p(\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N \times p}, \boldsymbol{\theta}) = \frac{p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k \times p}, \boldsymbol{\theta}_k)p(\tilde{z}_k = 1|\boldsymbol{Y}_{N \times p}, \boldsymbol{\theta})}{p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\boldsymbol{Y}_{N \times p}, \boldsymbol{\theta})} \tag{4}$$

All densities on the right hand side are available in closed form. Appendix B.5.1 details $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k \times p}, \boldsymbol{\theta}_k)$, the predictive density for category k. Appendix B.5.2 details $p(\tilde{z}_k = 1|\boldsymbol{Y}_{N \times p}, \boldsymbol{\theta})$, the probability that $\tilde{z}_k = 1$ prior to observing $\tilde{\boldsymbol{y}}_{\tilde{p}}$. Appendix B.5.3 details $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\boldsymbol{Y}_{N \times p}, \boldsymbol{\theta})$, the predictive density for $\tilde{\boldsymbol{y}}_{\tilde{p}}$ marginalized over all event categories.

2.3 Bayesian Decision Theory

We now focus on event categorization, that is, placing $\tilde{\boldsymbol{y}}_{\tilde{p}}$ into one of the K training event categories, using Bayesian Decision Theory (Robert et al., 2007; Berger, 2013).

We call a_k , $k \in \{1, \ldots, K\}$ the action of placing $\tilde{\boldsymbol{y}}_{\tilde{p}}$ into the k^{th} event category. Given the data and prior specifications, the choice of each action a_k has an associated loss. Loss is unavoidably subjective and specified by a loss function that can be evaluated for each action. The action with the minimum expected loss is considered the lowest risk in this Bayesian setting. We choose a simple loss function that specifies a loss matrix of constants $\boldsymbol{C}$

$$\boldsymbol{C}_{K\times K} = \begin{array}{c} \\ \tilde{z}_1 = 1 \\ \vdots \\ \tilde{z}_K = 1 \end{array} \begin{array}{c} \begin{array}{ccc} a_1 & \ldots & a_K \end{array} \\ \begin{bmatrix} C_{1,1} & \ldots & C_{1,K} \\ \vdots & \ddots & \vdots \\ C_{K,1} & \ldots & C_{K,K} \end{bmatrix} \end{array}, \tag{5}$$

where each element of $\boldsymbol{C}$ corresponds to the loss for each action that is indexed by the columns, for a value of the random variable $\tilde{\boldsymbol{z}}_K$ indexing the rows. The elements of $\boldsymbol{C}$ are ideally chosen with some thoughtful use of utility theory (Robert et al., 2007; Berger, 2013). For a draw of $\tilde{\boldsymbol{z}}_K$ from the categorical distribution, the loss function is evaluated as $\boldsymbol{L}(\tilde{\boldsymbol{z}}_K, \boldsymbol{a}) = \tilde{\boldsymbol{z}}_K^\top \boldsymbol{C}$, producing a vector of the same length as the number of actions. A simple loss function uses $\boldsymbol{C}_{K\times K} = \boldsymbol{J}_K - \boldsymbol{I}_K$, a matrix of ones minus the identity matrix such that the diagonal of $\boldsymbol{C}_{K\times K}$ is equal to zero. Under this loss function, and using a draw of $\tilde{\boldsymbol{z}}_K$ with $\tilde{z}_2 = 1$, the loss of taking action a_2 and placing $\tilde{\boldsymbol{y}}_{\tilde{p}}$ in category two is equal to zero, while the loss of each of the remaining actions is equal to one. Taking the expectation of the loss function, with respect to the posterior distribution of random vector $\tilde{\boldsymbol{z}}_K$, is therefore $\mathbb{E}\left[\boldsymbol{L}(\tilde{\boldsymbol{z}}_K, \boldsymbol{a})\right] = \mathbb{E}\left[\tilde{\boldsymbol{z}}_K^\top | \tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta}\right] \boldsymbol{C}$.

This Bayesian decision criterion is adaptable to an array of detonation detection scenarios. For binary decisions, an observer decides only whether $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is a detonation or not. Such cases are common in explosion monitoring. Then, $\boldsymbol{C}$ is a 2×2 matrix,

$$\boldsymbol{C}_{2\times 2} = \begin{array}{c} \\ \tilde{z}_1 = 1 \\ \tilde{z}_1 \neq 1 \end{array} \begin{array}{c} \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{bmatrix} C_{1,1} & C_{1,2} \\ C_{2,1} & C_{2,2} \end{bmatrix} \end{array}, \tag{6}$$

where $\tilde{z}_1$ corresponds to the event of interest. When $C_{1,1} = C_{2,2} = 0$ and $C_{1,2} = C_{2,1} = 1$, the matrix represents the 0-1 loss function in classical hypothesis testing (Robert et al., 2007). In Section 4, we primarily investigate binary categorization to showcase how changing an element of $\boldsymbol{C}_{2\times 2}$ allows one to intuitively target false negatives or false positives. Appendix D.1 details how reducing the action space to binary categorization allows several simplifications.

When training data has missing entries, the posterior expected loss is evaluated as $\mathbb{E}\left[\tilde{\boldsymbol{z}}_K^\top | \tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}^+, \boldsymbol{\theta}\right] \boldsymbol{C}$. Appendix D.2 details the practical use of Monte Carlo integration to approximate this marginal expectation.

A critique of this categorization method is that if $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is not from one of the K training event categories and is a true outlier, there will still be an action with the lowest expected loss, and $\tilde{\boldsymbol{y}}_{\tilde{p}}$ will be placed into the wrong category. Our early work on this methodology considered probabilistically allowing for a $(K+1)^{\text{th}}$ category. However, we found the results of our efforts highly sensitive to prior selection. Instead, we address this issue by utilizing a Bayesian typicality index in conjunction with

Bayesian Decision Theory. Our Bayesian typicality index places a Bayesian twist on the typicality index (McLachlan, 2005), which is used in C-ECM, by utilizing the multivariate t predictive distribution of $\tilde{\boldsymbol{y}}_{\tilde{p}}$ and Bayesian Decision Theory to decide on rejection in the event that elements of the training data are missing. In the event that $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is rejected, via the typicality index, from the event category selected as the minimum expected loss action, $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is considered an outlier, possibly belonging to an event category not included in the K training event categories. Appendix D.3 details use of the typicality index in a Bayesian setting.

3 Implementation

An R package titled `ezECM` implements the model; it is a "living package" under active improvement. The `ezECM` package provides functions for loading data, training a B-ECM model, saving and loading training results, predicting the category of a new observation, and making decisions using Bayesian Decision Theory, as well as summarizing and plotting results. The `ezECM` package also includes an implementation of C-ECM, which provides a baseline for comparing empirical results between models.

When there are no missing training data, evaluation of each $p(\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta})$ is available in closed form. When there are missing data entries, we implement a Gibbs sampler (Casella and George, 1992) in `ezECM` to generate samples from each $p(\boldsymbol{Y}_k^-|\boldsymbol{Y}_k^+, \boldsymbol{\theta}_k)$. The Gibbs sampler is the only computationally intensive aspect of training our B-ECM model. Pseudocode for this operation is provided as Algorithm 1. A user of the algorithm either supplies prior parameters $\boldsymbol{\theta}$ (see Appendix B.2) or uses the default values in the package function. The user also provides the total number of samples T to take of $\boldsymbol{Y}^-$ and the number of burn-in samples B, which are discarded under the assumption that the Markov chain has not converged to the target distribution within the first B iterations. $T - B$ total samples are obtained at the end of Algorithm 1. Later in Algorithm 2, the user can choose to use thinning, which trades a reduction in autocorrelation between the Monte Carlo samples of $\boldsymbol{Y}^-$ for an increase in training computation time. Thinning the samples by integer factor q then utilizes only every q^{th} sample of $\boldsymbol{Y}^-$ in the predict function of `ezECM`. The size of the integer set $\{B, \ldots, T\}$ used as values for the index t and the divisor for computing the mean of $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_k^+, (\boldsymbol{Y}_k^-)^{(t)}, \boldsymbol{\theta}_k)$ in Algorithm 2 are adjusted accordingly.

Initial values for the missing entries $(\boldsymbol{Y}^-)^{(1)}$ must be set at the start of Algorithm 1. In `ezECM`, the initialized missing elements $(\boldsymbol{Y}_k^-)^{(1)}$ are taken to be the column mean of the observed elements. The missing entries from each column of each event category are drawn (conditional on the remaining entries). For the $\ell^{\text{th}} \in \{1, \ldots, p\}$ column of $\boldsymbol{Y}_{N_k\times p}$, a column permutation matrix $\boldsymbol{P}_{k,\ell}^C$ swaps the columns of $\boldsymbol{Y}_{N_k\times p}$, and a row permutation matrix $\boldsymbol{P}_{k,\ell}^R$ swaps the rows of $\boldsymbol{Y}_{N_k\times p}$ such that $\underline{\boldsymbol{Y}}_{N_k\times p} = \boldsymbol{P}_{k,\ell}^R \boldsymbol{Y}_{N_k\times p} \boldsymbol{P}_{k,\ell}^C$. The missing data in the ℓ^{th} column of $\boldsymbol{Y}_{N_k\times p}$ is found in the first elements of the first column of $\underline{\boldsymbol{Y}}_{N_k\times p}$, so that the missing data in column ℓ can be drawn from the conditional distributions found in Appendix B.4; Appendix B details the notation found in Algorithm 1. Our algorithm saves the draw and updates the corresponding elements of $\boldsymbol{Y}_{N_k\times p}$ with these values. We repeat this process for all columns of $\boldsymbol{Y}_{N_k\times p}$ with missing values in the original training data set, and then the process is repeated for T iterations. The process approximates draws from the joint distribution $p(\boldsymbol{Y}_k^-|\boldsymbol{Y}_k^+, \boldsymbol{\theta}_k)$ (Robert and Casella, 1999) for each k.

Once we make draws of $\boldsymbol{Y}^-$, we then approximate $\mathbb{E}[\tilde{\boldsymbol{z}}_K|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}^+, \boldsymbol{\theta}]$ with a new observation $\tilde{\boldsymbol{y}}_{\tilde{p}}$ with functions in `ezECM`, and we use the results within a Bayesian Decision Theory framework (see Section 2.3). Algorithm 2 documents pseudocode for this process. Using the $t \in \{1, \ldots, T-B\}$ draws of $(\boldsymbol{Y}^-)^{(t)}$ Algorithm 1 outputs, we must evaluate the expected predictive category probability (4) for each $k \in \{1, \ldots, K\}$ by first joining $(\boldsymbol{Y}^-)^{(t)}$ with observations $\boldsymbol{Y}^+$ and evaluating the multivariate

Algorithm 1 Joint Monte Carlo samples of $\boldsymbol{Y}_k^-; \forall k \in \{1, \ldots, K\}$

Require: $\boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{Y}^+, \boldsymbol{P}_{k,\ell}^R, \boldsymbol{P}_{k,\ell}^C$
  Initialize:
    $(\boldsymbol{Y}^-)^{(1)}, (N_k^m)^\ell \times T$ Matrices to store samples $(\boldsymbol{y}_{k,\ell}^-)^{(t)}; \forall\ k \in \{1, \ldots, K\}, \ell \in \{1, \ldots, p\}, t \in \{1, \ldots, T\}$
  for $t \in \{1, \ldots, T\}$ **do**
    for $k \in \{1, \ldots, K\}$ **do**
      for $\ell \in \{1, \ldots, p\}$ **do**
        $\underline{\boldsymbol{Y}}_{N_k \times p} \leftarrow \boldsymbol{P}_{k,\ell}^R \boldsymbol{Y}_{N_k \times p} \boldsymbol{P}_{k,\ell}^C$
        $\underline{\boldsymbol{\eta}}_k \leftarrow (\boldsymbol{P}_{k,\ell}^C)^\top \boldsymbol{\eta}_k$
        $\underline{\boldsymbol{\Psi}}_k \leftarrow (\boldsymbol{P}_{k,\ell}^C)^\top \boldsymbol{\Psi}_k \boldsymbol{P}_{k,\ell}^C$
        $\boldsymbol{y}_{k,\ell}^- \sim p(\underline{\boldsymbol{y}}_{N_k^m \times 1}^- | \underline{\boldsymbol{y}}_{N_k^o \times 1}^+, \underline{\boldsymbol{Y}}_{N_k^m \times (p-1)}, \underline{\boldsymbol{Y}}_{N_k^o \times (p-1)}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$
        $(\boldsymbol{y}_{k,\ell}^-)^{(t)} \leftarrow \boldsymbol{y}_{k,\ell}^-$
        $\underline{\boldsymbol{y}}_{N_k^m \times 1}^- \leftarrow \boldsymbol{y}_{k,\ell}^-$
        $\boldsymbol{Y}_{N_k \times p} \leftarrow (\boldsymbol{P}_{k,\ell}^R)^\top \underline{\boldsymbol{Y}}_{N_k \times p} (\boldsymbol{P}_{k,\ell}^C)^\top$
      end for
    end for
  end for

t-distribution density $p(\tilde{\boldsymbol{y}}_{\tilde{p}} | \tilde{z}_k = 1, \boldsymbol{Y}^+, (\boldsymbol{Y}^-)^{(t)}, \boldsymbol{\theta}_k)$, detailed in Appendix B.5.1, for all k. Then, the integral over $\boldsymbol{Y}^-$ to find each density $p(\tilde{\boldsymbol{y}}_{\tilde{p}} | \tilde{z}_k = 1, \boldsymbol{Y}_k^+, \boldsymbol{\theta}_k)$ is approximated (Ulam and Metropolis, 1949) as the mean over all t and used with the result in Appendix D.2 to evaluate the expected category probabilities for a $\tilde{\boldsymbol{y}}_{\tilde{p}}$. In the case where $\tilde{p} < p$, we use the properties of the marginal matrix t-distribution (Appendix C) to evaluate $p(\tilde{\boldsymbol{y}}_{\tilde{p}} | \tilde{z}_k = 1, \boldsymbol{Y}_k^+, (\boldsymbol{Y}_k^-)^{(t)}, \boldsymbol{\theta}_k)$.

Lastly, we provide a function in `ezECM` to evaluate the loss function and to find the minimum loss action, given the loss matrix $\boldsymbol{C}_{2\times 2}$ or $\boldsymbol{C}_{K \times K}$. We also specify a category of importance. If the minimum loss action is to categorize $\tilde{\boldsymbol{y}}_{\tilde{p}}$ as the specified category, then we calculate the typicality index of that category for a significance level $\tilde{\alpha}$. If the algorithm deems $\tilde{\boldsymbol{y}}_{\tilde{p}}$ as atypical of the category, then we consider $\tilde{\boldsymbol{y}}_{\tilde{p}}$ to be an outlier, pending further analysis.

4 Experiments

We performed two Monte Carlo experiments to compare B-ECM to C-ECM, the first using synthetic statistically generated data and the second using real ground-based data. In each experiment, we used a set of training data to fit the models, and we used a set of testing data with a known truth to measure accuracy, false negatives, and false positives. We fit the five models that include (1) C-ECM, which can only utilize complete data records, (2) B-ECM with only complete data records, (3) B-ECM with all data records (M-B-ECM), (4) M-B-ECM with a loss function chosen to reduce false negatives (M-B-ECM $C_{1,2} = 2$), and (5) M-B-ECM for specific event categorization (M-B-ECM Cat). Decision criteria in models 1, 2, 3, and 4 were set up for binary categorization, while decision criteria for model 5 was set up for categorization into the specific source event type. We used 0-1 loss for B-ECM variants, except model 4. In this case, we utilized the loss matrix that

Algorithm 2 $\mathbb{E}[\tilde{\boldsymbol{z}}_K|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}^+, \boldsymbol{\theta}]$

Require: $\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{\theta}, \boldsymbol{Y}^+, (\boldsymbol{Y}^-)^{(B,\ldots,T)}$
  Initialize:
    Data Matrix $\boldsymbol{p}_{K\times(T-B)}$ to store samples
    $p_k^{(t)} = p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, (\boldsymbol{Y}_{N_k\times p})^{(t)}, \boldsymbol{\theta}_k)$
  for $k \in \{1, \ldots, K\}$ **do**
    for $t \in \{B, \ldots, T\}$ **do**
      $(\boldsymbol{Y}_{N_k\times p})^{(t)} \leftarrow \boldsymbol{Y}_k^+ \cup (\boldsymbol{Y}_k^-)^{(t)}$
      $p_k^{(t)} \leftarrow p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, (\boldsymbol{Y}_{N_k\times p})^{(t)}, \boldsymbol{\theta}_k)$
    end for
    $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_k^+, \boldsymbol{\theta}_k) \leftarrow \frac{1}{T-B}\sum_{t=B}^{T} p_k^{(t)}$
  end for
  $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\boldsymbol{Y}^+, \boldsymbol{\theta}) \leftarrow \sum_{k=1}^{K} p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_k^+, \boldsymbol{\theta}_k)p(\tilde{z}_k = 1|\boldsymbol{Y}_{N\times p}, \boldsymbol{\theta})$
  for $k \in \{1, \ldots, K\}$ **do**
    $\mathbb{E}[\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}^+, \boldsymbol{\theta}] = \frac{p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k=1, \boldsymbol{Y}_k^+, \boldsymbol{\theta}_k)p(\tilde{z}_k=1|\boldsymbol{Y}_{N\times p}, \boldsymbol{\theta})}{p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\boldsymbol{Y}^+, \boldsymbol{\theta})}$
  end for

Equation 7 shows.

$$\boldsymbol{C}_{2\times 2} = \begin{array}{c} \\ \tilde{z}_1 = 1 \\ \tilde{z}_1 \neq 1 \end{array} \begin{array}{c} \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix} \end{array} \tag{7}$$

We chose a level of significance for all typicality indices to be $\tilde{\alpha} = 0.05$ for both C-ECM and B-ECM. Decisions using any B-ECM model first require evaluating the posterior expected loss for each action considered. For binary decisions, the minimum expected loss action is used to place $\tilde{\boldsymbol{y}}_{\tilde{p}}$ in a presumptive category. If the presumptive category is a detonation, the Bayesian typicality index (Appendix D.3) is used to check if $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is rejected as an extreme value for detonations. Then, if $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is rejected from the detonation distribution, $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is categorized as not a detonation. Conversely, if there is failure to reject from the detonation distribution, $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is categorized as a detonation. If the presumptive category is not a detonation, no typicality index is used, and the presumptive category is the final categorization.

When using a B-ECM model for full K categorization, we calculated the typicality index for the minimum expected posterior loss action for all categories. In the case of rejection, $\tilde{\boldsymbol{y}}_{\tilde{p}}$ was categorized as an outlier. No outlier categories were included in the data for our experiments. This means that categorization of $\tilde{\boldsymbol{y}}_{\tilde{p}}$ as an outlier has a detrimental effect on accuracy and provides false negatives when the true category for $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is detonations.

We define accurate binary detection as the correct categorization of a true detonation as a detonation or any non-detonation as not a detonation. All non-detonation groups are combined into a single group. A binary categorization is still considered accurate when a non-detonation would be placed in the wrong subgroup but is correctly categorized as not a detonation. There is no penalty in accuracy for categorizing $\tilde{\boldsymbol{y}}_{\tilde{p}}$ as a deep earthquake if in reality $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is a shallow earthquake. However, for full categorization (as in M-B-ECM Cat), accuracy requires that the exactly correct source event type is chosen. The false positive rate is analogous to the empirical probability of mistakenly categorizing a $\tilde{\boldsymbol{y}}_{\tilde{p}}$ as a detonation when it is not. False negatives can be thought of as an empirical probability that a $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is categorized as a non-detonation when $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is a detonation. False

positives and false negatives are calculated in the same way for all binary detection and M-B-ECM Cat. Accuracy, false positive, and false negative rates are data dependent. However, we see trends over the various data that we utilize.

Using the code provided in the R package `ezECM`, we implemented B-ECM for both sets of experiments, with $p(\tilde{\boldsymbol{z}}_K|\boldsymbol{Y}_{N\times p}, \boldsymbol{\theta})$ informed by the data rather than equally weighted over K. The logit function, $\text{logit}(p) = \ln(p) - \ln(1-p)$, was used to map p-values to $(-\infty, \infty)$. The default prior parameters in the `ezECM` package, selected to allow for wide-ranging data observations, were used. Details on the prior parameters can be found in Appendix B.2.

The B-ECM models that utilized data with missing entries, in which Monte-Carlo was required for inference, generated 50,500 draws of each $\boldsymbol{Y}_k^-$ and discarded the first 500 draws as burn-in. These same implementations of B-ECM "thinned" the draws when we made predictions and only utilized every 5^{th} draw, or 10,000 draws in total.

Our implementation of C-ECM in the experiments was identical to our implementation of C-ECM in the `ezECM` function. In particular, C-ECM fit the RDA model using the `klaR::rda()` function from the `klaR` package (Weihs et al., 2005) in R. We only used C-ECM to form binary decisions. We calculated typicality indices from the Mahalanobis distance, using the methodology in Anderson et al. (2007). We structured the decision framework such that indeterminate and undefined categorizations could not occur because we felt that these categorizations would unfairly diminish performance metrics for C-ECM. When there is a failure to reject $\tilde{\boldsymbol{y}}_{\tilde{p}}$ from multiple event categories, the categorization is indeterminate. If $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is rejected from all event categories, the categorization is undefined. First, the typicality indices were calculated for all event categories. If $\tilde{\boldsymbol{y}}_{\tilde{p}}$ was rejected from the detonation category, then $\tilde{\boldsymbol{y}}_{\tilde{p}}$ was declared to be not a detonation. To have been declared a detonation, $\tilde{\boldsymbol{y}}_{\tilde{p}}$ must not have been rejected from the detonation category and must have been rejected from all other categories. Alternatively, if $\tilde{\boldsymbol{y}}_{\tilde{p}}$ was not rejected from the detonation category and one or more additional categories, the observation was declared to be not a detonation.

4.1 Synthetic Data

Our first experiment used synthetically generated data in 250 independent MC experiments for each of $p \in \{4, 6, 8, 10\}$. The data-generating mechanism was designed to mimic what we would expect to see when fusing ground and discriminants recorded in space, such as event depth and the number of satellites reporting an event. Equation (8) shows that we expect that for a $\boldsymbol{y}_p$ from a given event category, some discriminants will only have correlations within the space (S) and ground (G) modalities, while other event categories will have full correlation across discriminants. Additionally, we expect the number of discriminants used in practice to be of moderate size but not extremely large (fewer than 10). The values used for p reflect this choice:

$$\mathbb{C}\text{ov}(\boldsymbol{y}_{p\times 1}) \stackrel{?}{=} \begin{bmatrix} \boldsymbol{\Sigma}_{\text{S}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{\text{G}} \end{bmatrix} \stackrel{?}{=} \begin{bmatrix} \boldsymbol{\Sigma}_{\text{S}} & \boldsymbol{\Sigma}_{\text{S,G}} \\ \boldsymbol{\Sigma}_{\text{G,S}} & \boldsymbol{\Sigma}_{\text{G}} \end{bmatrix}. \tag{8}$$

We used $K = 3$ event categories to generate data. For each experiment, each category had a unique, randomly generated mean and covariance. The number of events in the data set with all discriminants observed used for training (N^{train}), the number of events with missing discriminants used for training (N^{train^-}), and the number of events used for testing (N^{test}) was held constant throughout the experiment. However, the number of each events partitioned into each event category was randomly selected in each MC iteration with a multinomial distribution and equal event probabilities (1/3 for each category). We chose randomly a subset of the event categories to

have a block covariance structure with a random block size. Using the set of event category-specific randomly generated parameters, we generated data from the draws of the multivariate normal distribution, and then we transformed the data onto $[0,1]^p$ using the logistic function. As a result, the data mimics the p-values used in application. The training data set contained full observations of 25 events. An additional 125 events, for which 50% of the $125 \times p$ elements were unobserved, were included in the training data set. Our testing data set included 100 events with 50% of the $p \times 100$ entries missing. Algorithm 3 of Appendix E.1 details the data generating mechanism.

Fig. 3 shows as a box plot the distribution of the observed accuracy for each of the 250 MC iterations. For C-ECM, B-ECM, M-B-ECM, and M-B-ECM, $C_{1,2} = 2$ models accuracy is defined as the correct placement of a new observation into the detonation group (or correctly deciding the new observation is not a detonation). For M-B-ECM Cat, accurate categorization requires correctly identifying the exact source category. Here, the binary categorization B-ECM models often perform better than C-ECM, even without taking advantage of partial observations for training. This difference is more pronounced as p increases. We hypothesize that C-ECM returned little change in median accuracy for increasing p because the decision criterion is relatively stringent, often leading to indeterminate or undefined categorizations. A C-ECM model applied on a similar data set would require much additional human intervention. Additionally, as p increases, there is a growing gap in the variability of the observed accuracy between all B-ECM models and C-ECM. From this observation, we infer that B-ECM can better leverage larger numbers of discriminants to deliver more consistent results than C-ECM.

Accurate categorization of $\tilde{\boldsymbol{y}}_{\tilde{p}}$ into a specific event source type requires correctly choosing one of K bins $\{\text{Detonation}, \text{ Not Detonation}_1, \ldots, \text{ Not Detonation}_{K-1}\}$. Detection only requires the correct choice between two bins, i.e., {Detonation, Not Detonation}. Detection is therefore simpler than categorization and can be thought of as nested within the categorization problem. This increased complexity and difficulty of categorization is reflected in the results for the M-B-ECM Cat model. For $p = 4$, median accuracy of M-B-ECM Cat is lower than that of C-ECM. However, relative accuracy improves for increasing p, and M-B-ECM Cat clearly performs better than C-ECM for $p = 8$ and 10. The result of M-B-ECM Cat having a higher accuracy than C-ECM for $p = 8$ and 10, when C-ECM is judged on an easier task, shows that the utility of using partial observations can be high enough to automate more difficult tasks.

Table 1 provides results on the accuracy rate, false negative rate, and false positive rate calculated over the entirety of the experiment. Values shown in parentheses indicate the benefit or reduction in benefit from using a B-ECM model over C-ECM in this specific problem. C-ECM has a relatively high false negative rate and a low false positive rate. B-ECM trained on the same data provides slight improvements in accuracy, with the magnitude of improvement for increasing values of p. B-ECM significantly improves upon the false negative rate of C-ECM and has a slightly worse false positive rate.

M-B-ECM has much larger improvements in accuracy, especially when $p = 10$. For $p = 10$, the false positive rate of M-B-ECM is similar to that of C-ECM, and the false negative rate is much improved over C-ECM. Table 1 shows that using the same M-B-ECM fit, but changing the loss function to target a reduction in false negatives, M-B-ECM $C_{1,2} = 2$ trades a small reduction in accuracy and increase in the false positive rate for a further decrease in the false negative rate over M-B-ECM. For $p = 4$, M-B-ECM $C_{1,2} = 2$ has a significantly lower false negative rate. For smaller values of p and constant N, the problem is more difficult for all models. Changing the loss function to reduce false negatives allows us to be more cautious given the increased uncertainty, with little penalty to overall accuracy.

For full categorization using M-B-ECM Cat, the threshold for the value of $p(\tilde{z}_k = 1 | \tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N \times p}, \boldsymbol{\theta})$

to categorize $\tilde{\boldsymbol{y}}_{\tilde{p}}$ in the category of interest indexed as k is lower than that of binary categorization for M-B-ECM. Naturally, this reduction in the threshold reduces the false negative rate while raising the false positive rate when compared to M-B-ECM using 0-1 loss. For $p = 10$, values for all performance metrics are similar for both M-B-ECM and M-B-ECM Cat. The relative performance improvements of these models over C-ECM illustrates the flexibility of B-ECM for adapting to data-fusion applications with larger number of discriminants and covariance matrices with no inter-category dependence.

Table 1: Empirical results from synthetic data experiments. A relatively higher accuracy, lower false negative rate, or lower false positive rate indicate an improvement in performance. Values in parenthesis are the mean result for a B-ECM model minus the corresponding result for C-ECM.

	$p = 4$	$p = 6$	$p = 8$	$p = 10$
C-ECM				
Accuracy	0.73	0.75	0.76	0.75
False Negative	0.74	0.67	0.64	0.69
False Positive	0.04	0.04	0.04	0.03
B-ECM				
Accuracy	0.77 (Δ0.04)	0.82 (Δ0.07)	0.86 (Δ0.1)	0.9 (Δ0.15)
False Negative	0.46 (Δ-0.28)	0.35 (Δ-0.32)	0.25 (Δ-0.39)	0.18 (Δ-0.51)
False Positive	0.11 (Δ0.08)	0.1 (Δ0.06)	0.09 (Δ0.04)	0.06 (Δ0.03)
M-B-ECM				
Accuracy	0.79 (Δ0.06)	0.85 (Δ0.1)	0.89 (Δ0.13)	0.92 (Δ0.17)
False Negative	0.45 (Δ-0.29)	0.31 (Δ-0.36)	0.21 (Δ-0.43)	0.15 (Δ-0.53)
False Positive	0.08 (Δ0.05)	0.08 (Δ0.03)	0.06 (Δ0.02)	0.04 (Δ0.01)
M-B-ECM $C_{1,2} = 2$				
Accuracy	0.76 (Δ0.03)	0.83 (Δ0.08)	0.87 (Δ0.12)	0.92 (Δ0.17)
False Negative	0.23 (Δ-0.51)	0.17 (Δ-0.5)	0.14 (Δ-0.5)	0.11 (Δ-0.58)
False Positive	0.24 (Δ0.21)	0.17 (Δ0.13)	0.12 (Δ0.08)	0.07 (Δ0.04)
M-B-ECM				
Accuracy	0.67 (Δ-0.06)	0.76 (Δ0.01)	0.83 (Δ0.07)	0.89 (Δ0.14)
False Negative	0.35 (Δ-0.39)	0.26 (Δ-0.41)	0.19 (Δ-0.45)	0.14 (Δ-0.54)
False Positive	0.15 (Δ0.12)	0.11 (Δ0.07)	0.08 (Δ0.04)	0.05 (Δ0.02)

4.2 Seismic Discriminant Data

We now implement our competing versions of ECM against real observations that researchers collected from ground-deployed sensors (seismometers) (Anderson and Taylor, 2008). Observers computed the discriminants in this dataset entirely from seismic waveforms that were sourced by real events, including deep earthquakes, shallow earthquakes, and various explosions 1. The size of these events (seismic magnitude or explosion yield) largely determined whether a given distribution

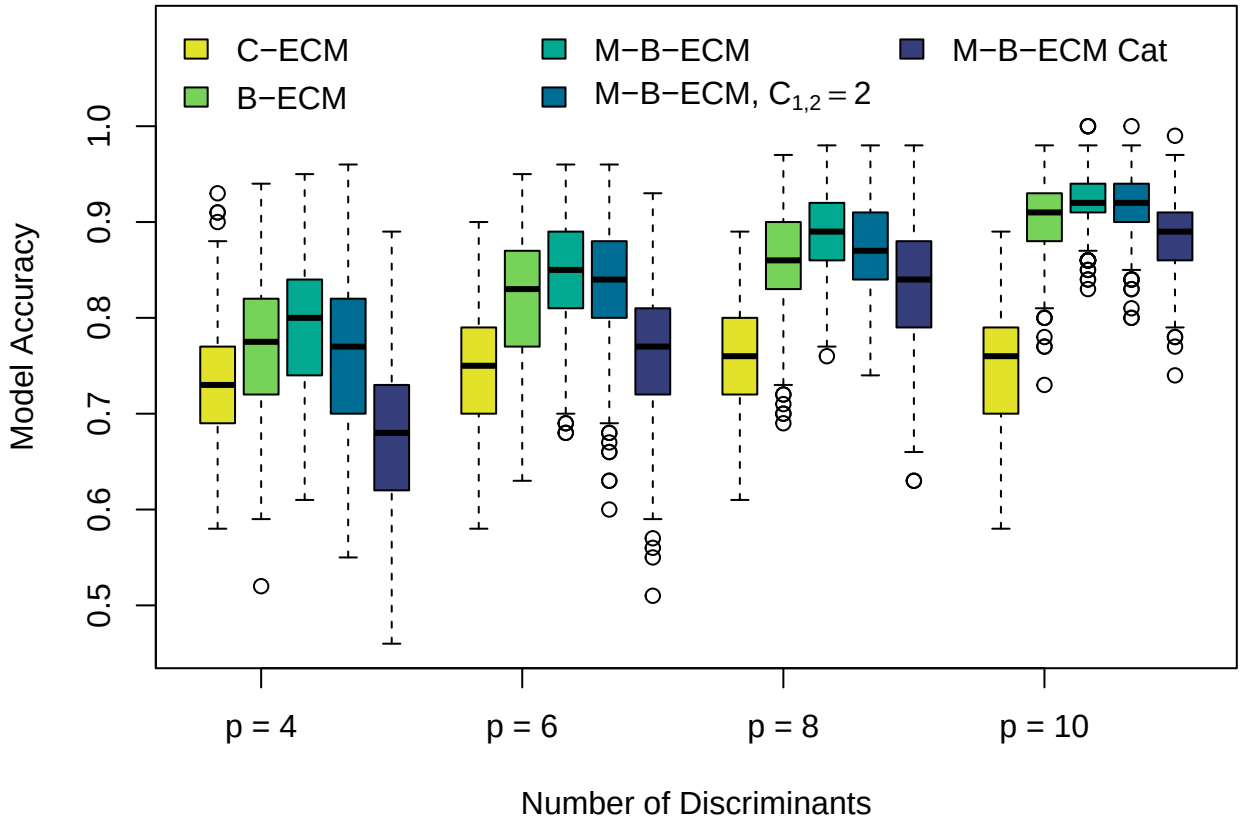


Figure 3: Box plots showing the distribution of observed accuracy over the MC iterations for the synthetic data-generating mechanism.

of seismometers could record signals over the ambient noise environment well enough to estimate source-type discriminants.

In numerous cases, the data set omits discriminants when signals are absent from a sufficient number of seismometer observations; therefore, there are entries missing from the data set. One of the most common discriminants in this dataset is the so-called signal complexity factor, or β_{CF} (see Fig. 2). This factor measures the log-ratio $\log\left(\frac{E_c}{E_s}\right)$ between seismic waveform coda energy E_c and direct (non-scattered) seismic signal energy E_s. Here, the coda wave energy E_c is measured over a window that begins 5 seconds after the first compressional wave's arrival and ends 25 seconds after its arrival. The signal energy E_s is measured in a 5-second window that begins immediately after the compressional wave's arrival. Researchers in the United Kingdom's (UK's) Atomic Weapons Establishment (AWE) formed these measurements from seismic array beams in the UK (station code EKA), India (station code GBA), Australia (station code WRA), and Canada (station code YKA). These beams were filtered over a passband of 0.25 to 4 Hz. Observations demonstrate that waveforms sourced by both nuclear explosions and deep earthquakes show relatively simple waveforms and less scattered energy; these produce a negative value of β_{CF}.

The other discriminants present in our AWE dataset are less populous but more conventional; they include earthquake depth estimates (64 percent) and body wave versus surface wave magnitudes (33 percent). Source depths provide discrimination power between earthquakes and underground explosions because present-day drilling technology limits the emplacement depth of any practical test to $\leq$ 10km; most tests are substantially more shallow. These discriminants can include substantial uncertainty when output by even advanced location algorithms, so that uncertainties in both earthquakes and explosions depths can be on the order of $\sim$ 10km when recorded by sparse networks (Arrowsmith et al., 2021). Seismic sources that locate near test sites and that undergo less than routine analyses (e.g., Voytan et al., 2019) show reduced uncertainties of a few km. We generally conclude that depth discriminants provide the greatest discrimination power between deeper earthquakes and detonations, when used alone.

Ratios between body wave magnitudes and surface wave magnitudes sourced by the same event (often called $m_b : M_S$) historically provided the most routine discrimination test between earthquakes and large ($m_b > 4.5$) nuclear explosions (Heyburn, 2014), although underground detonations in North Korea required modification of prior discrimination lines (Selby, Marshall, and Bowers, 2012). Physical explanations for the success of the $m_b : M_S$ discriminant include the seismic source spectra, focal mechanisms, interference between depth phases, and elastic properties near the source (Stevens and Day, 1985). Spectral differences account for substantial separation between earthquakes (one-half of a body wave magnitude unit) for large, but not smaller explosions. Focal mechanism differences introduce similar discrimination power, but also inflate intra-population scatter in earthquake data. Interference between depths phases improves discrimination rates mostly for deeply buried, large explosions. Differences in near-source elastic properties may provide the most reliable explanation for the discrimination power of $m_b : M_S$. This may be explained because earthquakes tend to occur in dense crustal rock with high velocities, whereas shallow explosions are often emplaced in lower velocity materials. Collectively, $m_b : M_S$ works best to separate large explosions and deep earthquakes.

Our dataset therefore includes discriminants that each show some error in their categorization rates between explosions and earthquakes. Such errors suggest that multiple discriminants that we propose under B-ECM may provide more reliable categorization rates. We refer readers to Anderson and Taylor (2008) and references therein for a summary of other discriminants' descriptions and mathematical forms.

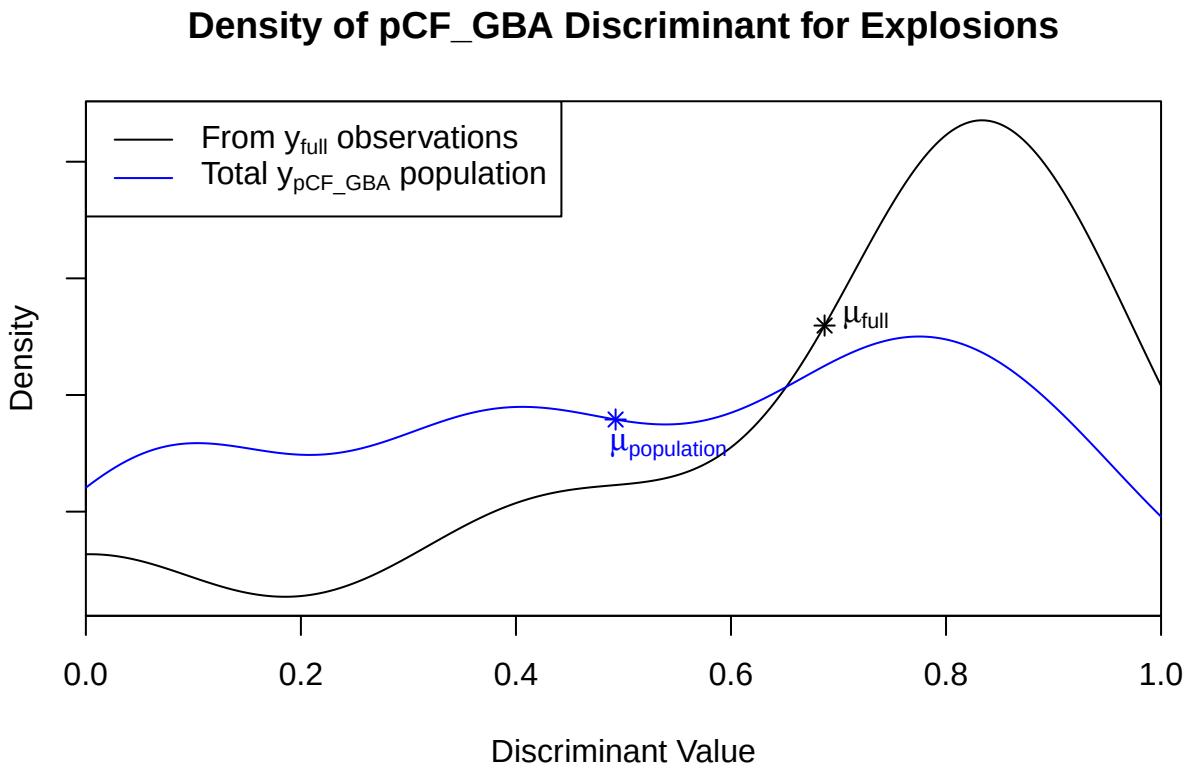


Figure 4: Density fits of subsets of the pCF_GBA discriminant recorded from explosions. The density fits are from the events for which the full $p = 5$ discriminants were recorded (shown in black) and the total population of the pCF_GBA data discriminant for explosions (shown in blue). The mean values, noted as μ, were taken after the logit transform and transformed back to $[0, 1]$ for plotting.

4.3 Application to the Seismic Discriminant Data

The categorization problem that we consider thereby includes only three event-type categories: explosions, deep earthquakes, and shallow earthquakes. The dataset groups nuclear and conventional explosion events together because seismic data cannot generally be used to discriminate between explosion source types (although body wave magnitudes from conventional explosions are usually less than those of nuclear explosions). We then reduce the total number of available discriminants to a subset that includes a sufficient number of discriminants (five) to train the C-ECM model. The resulting data set thereby contains five discriminants computed from 280 observations composed of 155 explosions, 26 deep earthquakes, and 99 shallow earthquakes. Our reduction in discriminants still retains a dataset that has 54% of the $280 \times 5 = 1{,}400$ of its elements missing. A small number of observations (25) contain data for all 5 discriminants: 12 explosions, 2 deep earthquakes, and 11 shallow earthquakes. We were unable to collect a combination of $p > 5$ that increased the number of full observations. We therefore used $p = 5$ within the MC experiment.

Missing entries within the dataset do not appear to be missing at random. Fig. 4 illustrates this property with p-values computed from the signal complexity factors from explosions at the GBA array (written as pCF_GBA). Density fits of this discriminant from the total population and the subset of the data that is associated with full $p = 5$ observations show that for this case, the mean of the data taken from full $p = 5$ observations is shifted significantly from the population

mean. This implies that the distributions of values obtained from events with different degrees of data "missingness" might not be the same.

The experiment consisted of 250 MC iterations. Subsets of the data were sampled for each iteration without replacement for training and testing. Because there were only two deep-earthquake observations with full observations, and because the functions used for C-ECM require at least two full observations for each category, both of these observations were included in every training data set. The 18 additional full observations were randomly sampled from the data set to include in the training data. Of the remaining 255 partial observations, we used $\lfloor 255 \times \sqrt{p}/(1+\sqrt{p}) \rfloor = 176$ (Joseph, 2022) for training. We used the remaining 84 observations, 79 partial observations, and 5 full observations for testing.

The distribution of the observed model accuracy over each MC iteration is shown as box plots in Fig. 5. There is relatively less variability in the results from this real data set compared to Fig. 3. The median accuracy of M-B-ECM, M-B-ECM $C_{1,2} = 2$, and M-B-ECM Cat is clearly better than that of C-ECM and B-ECM. If discriminants are not missing at random, the methods that utilize partial observations for training can get a better representation of the mean of the population, which is useful for predicting the category of $\tilde{\boldsymbol{y}}_{\tilde{p}}$ for testing data where $\tilde{p} \in \{1, \ldots, 5\}$.

The results of the total accuracy, false negative rate, and false positive rate calculated over the entirety of data collected are shown in Table 2. The results largely echo those in Table 1. B-ECM has a slightly higher overall accuracy than C-ECM for this moderate $p = 5$ problem. C-ECM has a much higher false negative rate than all B-ECM comparators and a lower false positive rate.

Including missing data to train a M-B-ECM model generally resulted in improved accuracy, a greatly reduced false negative rate, and a worse false positive rate. For many of the explosion discriminants, the population has a higher variance than the portion that is part of a full $\tilde{p} = 5$ event. We hypothesize that the increase in variance for these discriminants within the training data translated to more true explosion $\tilde{\boldsymbol{y}}_{\tilde{p}}$ being captured as more probable. Conversely, this increase in variance could also lead to more $\tilde{\boldsymbol{y}}_{\tilde{p}}$, which are not truly explosions to be categorized as such. The deep-earthquake and shallow-earthquake discriminant populations often have a lower variance than their explosion counterparts.

M-B-ECM $C_{1,2} = 2$ resulted in a further reduction in the false negative rate, with little trade-off in overall accuracy. Similar to what can be noted from Table 1, the lower threshold for categorization as an explosion results in a lower false negative rate and higher false positive rate than binary categorization with M-B-ECM and 0-1 loss.

5 Discussion and Conclusions

Results consistently suggest that a B-ECM model trained on observations with no missing data has accuracy similar to or greater than a C-ECM model, a much lower false negative rate, and a slightly higher false positive rate. The inclusion of additional training events where some data is missing for a B-ECM model consistently results in additional improvements in categorization accuracy and reductions in the false negative rate. Changes in the false positive rate can depend on how much the distributions of discriminant values overlap between event. These findings indicate that B-ECM is a worthy competitor for event identification in explosion monitoring. The experiments in Sections 4.1 and 4.2 highlight these differences in performance. Although C-ECM has relatively low false positive rates for all data used and all sizes of p tested, B-ECM typically performs better in all other aspects for these difficult problems, and B-ECM has a similar false positive rate for some problems. Adjusting the values of the loss matrix for applications where a lower false positive rate is desirable facilitates such a reduction.

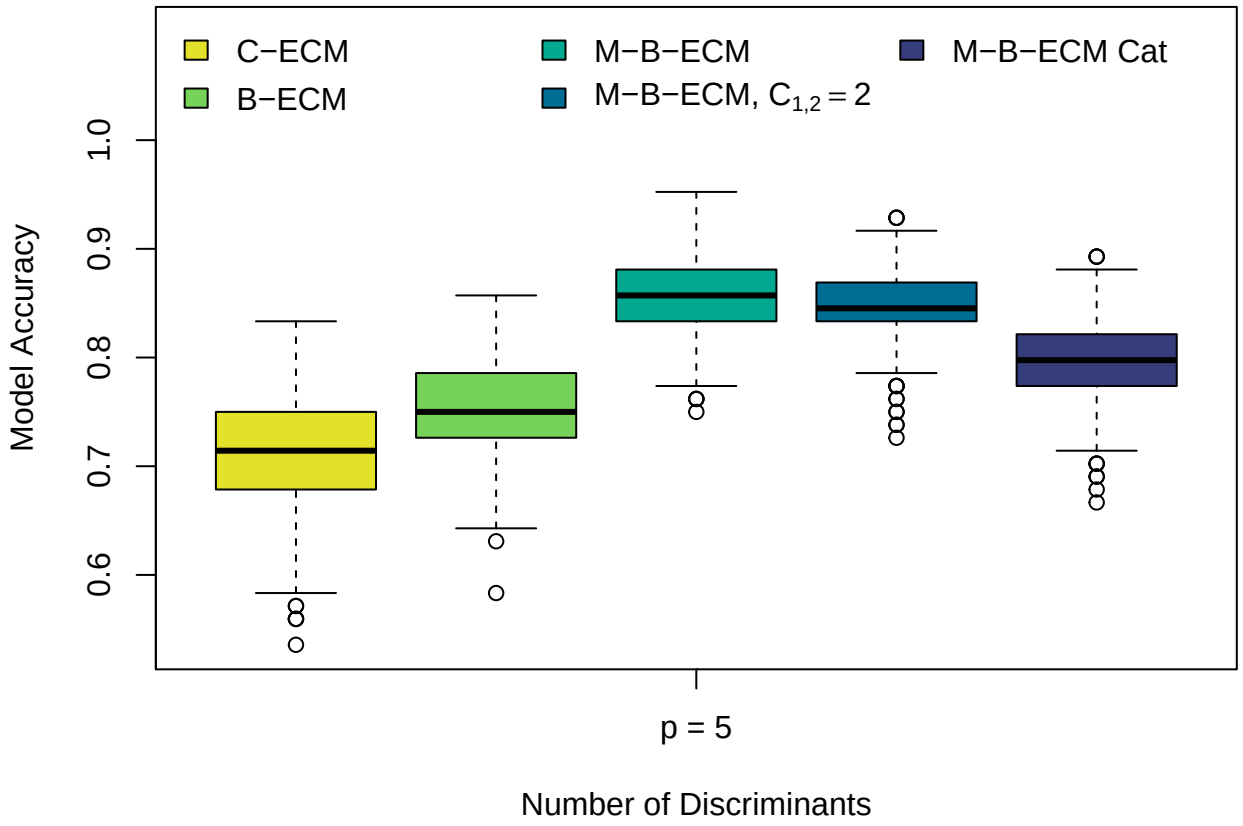


Figure 5: Box plots showing the distribution of observed accuracy over the MC iterations for the seismic discriminant data set.

Table 2: Empirical results from seismic discriminant experiments. A relatively higher accuracy, lower false negative rate, or lower false positive rate indicate an improvement in performance. Values in parenthesis are the mean result for a B-ECM model minus the corresponding result for C-ECM.

	$p = 5$
C-ECM	
Accuracy	0.71
False Negative	0.51
False Positive	0.01
B-ECM	
Accuracy	0.75 (Δ0.04)
False Negative	0.41 (Δ-0.09)
False Positive	0.03 (Δ0.02)
M-B-ECM	
Accuracy	0.85 (Δ0.14)
False Negative	0.12 (Δ-0.39)
False Positive	0.18 (Δ0.17)
M-B-ECM $C_{1,2} = 2$	
Accuracy	0.85 (Δ0.14)
False Negative	0.09 (Δ-0.42)
False Positive	0.24 (Δ0.22)
M-B-ECM Cat	
Accuracy	0.79 (Δ0.08)
False Negative	0.1 (Δ-0.41)
False Positive	0.21 (Δ0.2)

A B-ECM model that can handle training data with missing elements assumes the data is missing completely at random. A combination of intuition and the evidence displayed in Fig. 4 leads us to believe that missing data elements are not truly random. An intuitive hypothetical would be a low-yield weapon not meeting certain thresholds to start automated data recording, resulting in some missing data for such an event. A critic could reasonably call the B-ECM misspecified because of its missing, completely at random assumption. However, just as B-ECM has built upon and addressed deficiencies in C-ECM to improve performance, we expect accounting for a missing-not-at-random mechanism (Little and Rubin, 2019) to result in additional performance gains for ECM models.

The decision-theoretic framework for B-ECM provides intuition to tune "knobs" by changing values within the loss matrix, as in Equation (7). Values within the loss matrix could be chosen subjectively, as is done in Section 4, or tuned empirically given a large enough data set, in order to target a reduction in false negatives or false positives. When we increase the value of element $C_{1,2}$, logically we are increasing the loss associated with erroneously choosing to categorize $\tilde{\boldsymbol{y}}_{\tilde{p}}$ as not a detonation when the event truly is a detonation. The intuitive relationship between the values of the elements of a loss matrix has utility in operations. Just as important, changing the values of

the loss matrix impacts the results as intended. In our experiments, increasing $C_{1,2}$ to a value of 2 in order to reduce the false negative rate did result in the intended reduction by roughly a factor of two, with only slight reductions in overall accuracy.

The B-ECM model is flexible enough to adapt to both detection and categorization applications. Sets of elements from the K length vector from the predictive category distribution $p(\tilde{\boldsymbol{z}}_K|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta})$ can be summed to group event categories together, so that B-ECM can be used to categorize $\tilde{\boldsymbol{y}}_{\tilde{p}}$ into anywhere from two to K groups with a corresponding adjustment of the loss matrix. A larger number of actions to choose from corresponds with a decrease in accuracy. This increase in the complexity of the problem results in the observed decrease in accuracy for both data sets explored in experimentation. However, this effect appears to diminish as p increases, which we hypothesize is due to a decrease in the overlap of $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k\times p}, \boldsymbol{\theta}_k)$ in higher dimension for all k. Our testing with event categorization utilizes data with missing entries. Even though the categorization problem is more difficult, at times, event categorization had higher accuracy than methods that did not take advantage of missing data. This was particularly true in Section 4.2. These results illustrate how powerful utilizing all data can be for more difficult problems.

We introduced a novel decision framework for the Bayesian typicality index, which is able to detect outliers under the uncertainty imparted by using partial data observations. The use of the Bayesian typicality index ensures that a $\tilde{\boldsymbol{y}}_{\tilde{p}}$ which was not generated from one of the K events, might be categorized as an outlier. Without the typicality index, a new observation must be categorized under the finite set of event categories used for training. The Bayesian typicality index is a B-ECM model-specific twist on established methodology, which can take into account uncertainty related to using training data with missing entries.

With the ability to handle partial observations and the use of Bayesian Decision Theory, B-ECM has been shown to be an improvement over C-ECM, and there are many avenues for continuing improvement beyond the current state of the art. Values of the matrix $\boldsymbol{C}$ and typicality index significance level $\tilde{\alpha}$ can be tuned to meet an application-specific objective. The utility of loss matrices where the elements are functions instead of constants can be investigated (Robert et al., 2007; Berger, 2013). We believe taking advantage of Bayesian model selection procedures would produce further improvements in accuracy, especially in applications where data is fused over multiple modalities. In a preliminary investigation, we did not see consistent improvements in accuracy for the the logit function over the transform used in Anderson et al. (2007). The preferable transform depends on the data itself, and the data can inform which transform to use as part of a Bayesian model selection procedure using Bayes Factors (Kass and Raftery, 1995). Additionally, model-selection procedures can be used to allow uniquely sparse covariance matrices (Jordan, 1999; Roverato, 2002) for each category and models that assume missing data is not at random (Little and Rubin, 2019). For event categories where there is no correlation between particular discriminants, such a model may be a better fit to the data.

Author Contribution Statement

- Scott Koermer: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing
- Joshua D. Carmichael: Data curation, Supervision, Methodology, Visualization, Writing - original draft, Writing - review & editing
- Brian J. Williams: Methodology, Supervision, Writing - original draft, Writing - review & editing

Data Availability

Code and data to generate the results can be found within the `ezECM` package for the R programming language. Further details are provided in Appendices E.1.1 and E.2.1.

Acknowledgement

This manuscript has been authored with number LA-UR-24-30189 by Triad National Security under Contract with the U.S. Department of Energy, Office of Defense Nuclear Nonproliferation Research and Development. This research was funded by the National Nuclear Security Administration, Defense Nuclear Nonproliferation Research and Development (NNSA DNN R&D). The authors acknowledge important interdisciplinary collaboration with scientists and engineers from LANL, LLNL, MSTS, PNNL, and SNL. Los Alamos National Laboratory is supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. DE-AC52-06NA25396. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

Research presented in this article was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number 20220188DR. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001).

References

Anderson, Dale N (2009). "Sources of error: regional amplitude and teleseismic magnitude discriminants". In: *Monitoring Research Review: Ground-Based Nuclear Explosion Monitoring Technologies*.

Anderson, Dale N and Steven R Taylor (2008). "Rediscovering signal complexity as a teleseismic discriminant". In: *Pure and Applied Geophysics* 166.LA-UR-08-06754; LA-UR-08-6754.

Anderson, Dale N et al. (2007). "A mathematical statistics formulation of the teleseismic explosion identification problem with multiple discriminants". In: *Bulletin of the Seismological Society of America* 97.5, pp. 1730–1741.

Anderson, Dale N et al. (2010a). "Seismic event identification". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4, pp. 414–432.

Anderson, Dale N et al. (2010b). "Statistical methods in seismology". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.3, pp. 303–316.

Anderson, DN et al. (2014). "Sources of error and the statistical formulation of Ms: mb seismic event screening analysis". In: *Pure and Applied Geophysics* 171, pp. 537–547.

Arora, SK, TK Basu, and CA Krishnan (1983). "Source depth as a useful parameter in the discrimination of earthquakes and underground explosions". In: *Tectonophysics* 91.1-2, pp. 29–52.

Arrowsmith, Stephen et al. (2021). "Event location with sparse data: When probabilistic global search is important". In: *Seismological Society of America* 92.2A, pp. 976–985.

Arrowsmith, Stephen J and Steven R Taylor (2013). "Multivariate acoustic detection of small explosions using Fisher's combined probability test". In: *The Journal of the Acoustical Society of America* 133.3, EL168–EL173.

Berger, James O (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Booker, Aaron and Walter Mitronovas (1964). "An application of statistical discrimination to classify seismic events". In: *Bulletin of the Seismological Society of America* 54.3, pp. 961–971.

Bowers, David and Neil D Selby (2009). "Forensic seismology and the comprehensive nuclear-test-ban treaty". In: *Annual Review of Earth and Planetary Sciences* 37, pp. 209–236.

Brillinger, DR, A Udias, and BA Bolt (1980). "A probability model for regional focal mechanism solutions". In: *Bulletin of the Seismological Society of America* 70.1, pp. 149–170.

Carmichael, Joshua et al. (2020). "A method to fuse multiphysics waveforms and improve predictive explosion detection: theory, experiment and performance". In: *Geophysical Journal International* 222.2, pp. 1195–1212.

Carmichael, Joshua D (2021). "Hypothesis tests on Rayleigh wave radiation pattern shapes: A theoretical assessment of idealized source screening". In: *Geophysical Journal International* 225.3, pp. 1653–1671.

Carmichael, Joshua D et al. (2016). "Fusing geophysical signatures of locally recorded surface explosions to improve blast detection". In: *Geophysical Journal International* 204.3, pp. 1838–1842.

Casella, George and Edward I George (1992). "Explaining the Gibbs sampler". In: *The American Statistician* 46.3, pp. 167–174.

Council, National Research et al. (2012). "The comprehensive nuclear test ban treaty: Technical issues for the United States". In.

Elvers, Eva (1974). "Seismic event identification by negative evidence". In: *Bulletin of the Seismological Society of America* 64.6, pp. 1671–1683.

Ericsson, Ulf A (1970). "Event identification for test ban control". In: *Bulletin of the Seismological Society of America* 60.5, pp. 1521–1546.

Fah, Donat and Karl Koch (2002). "Discrimination between earthquakes and chemical explosions by multivariate statistical analysis: A case study for Switzerland". In: *Bulletin of the Seismological Society of America* 92.5, pp. 1795–1805.

Ford, Sean R et al. (2014). "Partitioning of seismoacoustic energy and estimation of yield and height-of-burst/depth-of-burial for near-surface explosions". In: *Bulletin of the Seismological Society of America* 104.2, pp. 608–623.

Ford, Sean R et al. (2021a). "Joint Bayesian inference for near-surface explosion yield and height-of-burst". In: *Journal of Geophysical Research: Solid Earth* 126.2, e2020JB020968.

— (2021b). "Joint Bayesian inference for near-surface explosion yield and height-of-burst". In: *Journal of Geophysical Research: Solid Earth* 126.2, e2020JB020968.

Friedman, Jerome H (1989). "Regularized discriminant analysis". In: *Journal of the American statistical association* 84.405, pp. 165–175.

Gelfand, Alan E and Adrian FM Smith (1990). "Sampling-based approaches to calculating marginal densities". In: *Journal of the American statistical association* 85.410, pp. 398–409.

Gelfand, Alan E et al. (1990). "Illustration of Bayesian inference in normal data models using Gibbs sampling". In: *Journal of the American Statistical Association* 85.412, pp. 972–985.

Gelman, Andrew et al. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Geman, Stuart and Donald Geman (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on pattern analysis and machine intelligence* 6, pp. 721–741.

Green, David N et al. (2013). "Hydroacoustic, infrasonic and seismic monitoring of the submarine eruptive activity and sub-aerial plume generation at South Sarigan, May 2010". In: *Journal of Volcanology and Geothermal Research* 257, pp. 31–43.

Gupta, Arjun K and Daya K Nagar (2018). *Matrix variate distributions*. Chapman and Hall/CRC.

Harville, David A (1998). *Matrix algebra from a statistician's perspective*.

Herzog, Stephen (2017). "The nuclear test ban: Technical opportunities for the new administration". In: *Arms Control Today* 47.1, pp. 26–32.

Heyburn, Ross (2014). "Forensic Seismology". In: *SECED Newsletter* 25.2.

Hoff, Peter D (2009). *A first course in Bayesian statistical methods*. Vol. 580. Springer.

Jih, Ron-Song et al. (1990). "Magnitude: Yield Relationship at Various Nuclear Test Sites–A Maximum-Likelihood Approach Using Heavily Censored Explosive Yields". In.

Jordan, Michael Irwin (1999). *Learning in graphical models*. MIT press.

Joseph, V Roshan (2022). "Optimal ratio for data splitting". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 15.4, pp. 531–538.

Kafka, Alan L (1990). "Rg as a depth discriminant for earthquakes and explosions: A case study in New England". In: *Bulletin of the Seismological Society of America* 80.2, pp. 373–394.

Kalinowski, Martin B et al. (2023). "Innovation in Technology and Scientific Methods for Nuclear Explosion Monitoring and Verification: Introduction". In: *Pure and Applied Geophysics* 180.4, pp. 1227–1234.

Kass, Robert E and Adrian E Raftery (1995). "Bayes factors". In: *Journal of the american statistical association* 90.430, pp. 773–795.

Koper, Keith D (2020). "The importance of regional seismic networks in monitoring Nuclear Test-Ban Treaties". In: *Seismological Research Letters* 91.2A, pp. 573–580.

Kotz, Samuel and Saralees Nadarajah (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

Little, Roderick JA and Donald B Rubin (2019). *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.

Maceira, Monica et al. (2017). *Trends in Nuclear Explosion Monitoring Research & Development - A Physics Perspective* -. Tech. rep. LA-UR-17-21274. Sponsor: USDOE. URL: `http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-17-21274`.

McGrath, Kaegan (2009). "Verifiability, Reliability, and National Security: The Case for US Ratification of the CTBT". In: *Nonproliferation Review* 16.3, pp. 407–433.

McLachlan, Geoffrey J (2005). *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons.

Murphy, John R, Brian W Barker, and William L Rodi (2002). "Improved focal depth determination for use in nuclear-explosion monitoring". In: *Proc. of 24th Seismic Research Review—Nuclear Explosion Monitoring: Innovation and Integration*, pp. 522–529.

Myers, Henry R (1972). "Comprehensive Test Ban Treaty: Grounds for Objection Diminish". In: *Science* 175.4019, pp. 283–286.

Ng, Kai Wang, Guo-Liang Tian, and Man-Lai Tang (2011). "Dirichlet and related distributions: Theory, methods and applications". In.

Park, Junghyun, Chris Hayward, and Brian W Stump (2018). "Assessment of infrasound signals recorded on seismic stations and infrasound arrays in the western United States using ground truth sources". In: *Geophysical Journal International* 213.3, pp. 1608–1628.

Redman, Brian J et al. (2019). "Hyperspectral vegetation identification at a legacy underground nuclear explosion test site". In: *Chemical, Biological, Radiological, Nuclear, and Explosives (CBRNE) Sensing XX*. Vol. 11010. SPIE, pp. 145–154.

Ringdal, F and ES Husebye (1982). "Application of arrays in the detection, location, and identification of seismic events". In: *Bulletin of the Seismological Society of America* 72.6B, S201–S224.

Robert, Christian P and George Casella (1999). *Monte Carlo statistical methods*. Vol. 2. Springer.

Robert, Christian P et al. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Vol. 2. Springer.

Rodd, Rebecca L et al. (2023). "A Multimodal Event Catalog and Waveform Data Set That Supports Explosion Monitoring from Nevada, USA". In: *Bulletin of the Seismological Society of America*.

Roverato, Alberto (2002). "Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models". In: *Scandinavian Journal of Statistics* 29.3, pp. 391–411.

Scoles, Sarah (2020). "A model to detect explosions big and small". In.

Selby, Neil D, Peter D Marshall, and David Bowers (2012). "mb: M s event screening revisited". In: *Bulletin of the Seismological Society of America* 102.1, pp. 88–97.

Shumway, Robert H (1996). "Statistical approaches to seismic discrimination". In: *Monitoring a comprehensive test ban treaty*. Springer, pp. 791–803.

— (2001). "Classical and Bayesian seismic yield estimation: The 1998 Indian and Pakistani tests". In: *pure and applied geophysics* 158, pp. 2275–2290.

Shumway, Robert H and DW Rivers (1984). *Testing the hypothesis of TTBT (Threshold Test Ban Treaty) compliance, and magnitude-yield regression for explosions in granite. Final technical report, 15 November 1982-30 December 1984*. Tech. rep. Teledyne Geotech, Alexandria, VA (USA).

Stephens, Matthew and D Phil (1997). "Bayesian methods for mixtures of normal distributions". PhD dissertation. University of Oxford Oxford.

Stevens, Jeffry L and Steven M Day (1985). "The physical basis of mb: Ms and variable frequency magnitude methods for earthquake/explosion discrimination". In: *Journal of Geophysical Research: Solid Earth* 90.B4, pp. 3009–3020.

Taylor, Steven R, Stephen J Arrowsmith, and Dale N Anderson (2010). "Detection of short time transients from spectrograms using scan statistics". In: *Bulletin of the seismological society of America* 100.5A, pp. 1940–1951.

Ulam, S and Nicholas Metropolis (1949). "The Monte Carlo Method". eng. In: *Journal of the American Statistical Association* 44.245, p. 335. ISSN: 0162-1459.

Voytan, Dimitri P et al. (2019). "Yield estimates for the six North Korean nuclear tests from teleseismic P wave modeling and intercorrelation of P and Pn recordings". In: *Journal of Geophysical Research: Solid Earth* 124.5, pp. 4916–4939.

Warren, Linda M and Peter M Shearer (2005). "Using the effects of depth phases on P-wave spectra to determine earthquake depths". In: *Bulletin of the Seismological Society of America* 95.1, pp. 173–184.

Weihs, Claus et al. (2005). "klaR Analyzing German Business Cycles". In: *Data Analysis and Decision Support*. Ed. by D. Baier, R. Decker, and L. Schmidt-Thieme. Berlin: Springer-Verlag, pp. 335–343.

Williams, Brian J et al. (2021). "Multiphenomenology explosion monitoring (MultiPEM): a general framework for data interpretation and yield estimation". In: *Geophysical Journal International* 226.1, pp. 14–32.

Wright, Christopher M and Lars-Erik De Geer (2017). "The 22 September 1979 Vela Incident: The Detected Double-Flash". In: *Science & Global Security* 25.3, pp. 95–124.

Yang, Ruoyong and James O Berger (1996). *A catalog of noninformative priors*. Vol. 1018. Institute of Statistics and Decision Sciences, Duke University Durham, NC, USA.

A Nomenclature

Algorithms

$(\boldsymbol{Y}_{N_k\times p})^{(t)}$ The union of the observed data $\boldsymbol{Y}_k^+$ and the t^{th} Markov chain Monte Carlo random sample of missing data $(\boldsymbol{Y}_k^-)^{(t)}$ for the k^{th} event category.

$(x)^{(t)}$ The t^{th} iterative sample of random variable x obtained in a Markov chain Monte Carlo algorithm.

$\boldsymbol{P}_{k,\ell}^C$ Matrix that permutes the ℓ^{th} column of $\boldsymbol{Y}_{N_k\times p}$ to the first column.

$\boldsymbol{P}_{k,\ell}^R$ Matrix that permutes rows containing the missing entries in the ℓ^{th} column of $\boldsymbol{Y}_{N_k\times p}$ to the top N_k^m rows.

$\boldsymbol{X}_{[\boldsymbol{a},\boldsymbol{b}]}$ Subset arbitrary matrix $\boldsymbol{X}$ as indicated by vectors $\boldsymbol{a}$ and $\boldsymbol{b}$.

$\boldsymbol{y}_{k,\ell}$ Missing data elements of column ℓ in $\boldsymbol{Y}_{N_k\times p}$.

$\underline{\boldsymbol{X}}$ Permuted matrix or vector, for arbitrary matrix or vector $\boldsymbol{X}$, rearranged for conditional sampling.

$\underline{\boldsymbol{Y}}_{N_k\times p}$ Permuted matrix of training data for the k^{th} event category. Equal to $\boldsymbol{P}_{k,\ell}^R\boldsymbol{Y}_{N_k\times p}\boldsymbol{P}_{k,\ell}^C$ for a given column ℓ.

$\underline{\boldsymbol{Y}}_{N_k^m\times(p-1)}$ Subset of the first N_k^m rows of $\underline{\boldsymbol{Y}}_{N_k\times p}$, with the first column removed.

$\underline{\boldsymbol{y}}_{N_k^m\times 1}^-$ Vector subset of the missing entries within the first column of $\underline{\boldsymbol{Y}}_{N_k\times p}$.

$\underline{\boldsymbol{Y}}_{N_k^o \times (p-1)}$ Subset of rows $N_k^m + 1$ through N_k of $\underline{\boldsymbol{Y}}_{N_k \times p}$, with the first column removed.

$\underline{\boldsymbol{y}}^+_{N_k^o \times 1}$ Vector subset of the observed entries within the first column of $\underline{\boldsymbol{Y}}_{N_k \times p}$.

B Number of burn-in samples of a random variable discarded immediately after starting the Markov chain.

q Integer parameter for thinning samples from the Markov Chain where after burn-in, only every q^{th} sample is saved.

q Integer thinning factor used to specify the retention of every q^{th} sample in Markov chain Monte Carlo in an effort to reduce auto-correlation between what would ideally be independent samples.

T Total number of iterative samples taken of a single random variable in Markov chain Monte Carlo.

t Indexes the T Monte Carlo samples of a random variable.

Model parameters

$\boldsymbol{\mu}_k$ Mean vector of length p, a multivariate normal distribution for the k^{th} event category.

$\boldsymbol{\pi}$ Vector of weights for the mixture of distributions. $\mathbf{1}_K^\top \boldsymbol{\pi} = 1$

$\boldsymbol{\Sigma}_k$ $p \times p$ covariance matrix from the multivariate normal distribution of the k^{th} event category. Parameterizes the covariance between the columns of $\boldsymbol{Y}_{N_k \times p}$.

π_k Weight of the k^{th} event category distribution within the mixture of distributions.

$\tilde{\boldsymbol{z}}_K$ Vector of latent variables of length K used to model the event category of $\tilde{\boldsymbol{y}}_{\tilde{p}}$. A single k^{th} element is equal to one, indicating inclusion into the k^{th} group, with the remaining $K - 1$ elements equal to zero.

$\tilde{z}_k$ The k^{th} element of $\tilde{\boldsymbol{z}}_K$

Discriminants

$\bar{\boldsymbol{y}}_k$ $p \times 1$ vector containing the mean of each column of $\boldsymbol{Y}_{N_k \times p}$; $\bar{\boldsymbol{y}}_k = \boldsymbol{Y}_{N_k \times p}^\top \mathbf{1}_{N_k} / N_k$.

$\boldsymbol{Y}^+$ The set of elements from $\boldsymbol{Y}_{N \times p}$ which are not missing. Equivalent to $\{\boldsymbol{Y}_1^+, \ldots, \boldsymbol{Y}_K^+\}$.

$\boldsymbol{Y}^-$ The set of missing elements from $\boldsymbol{Y}_{N \times p}$. Equivalent to $\{\boldsymbol{Y}_1^-, \ldots, \boldsymbol{Y}_K^-\}$.

$\boldsymbol{Y}_k^+$ The set of elements where data is not missing from $\boldsymbol{Y}_{N_k \times p}$.

$\boldsymbol{Y}_k^-$ The set of elements where data is missing from $\boldsymbol{Y}_{N_k \times p}$.

$\boldsymbol{Y}_{N \times p}$ Totality of the training data.

$\boldsymbol{Y}_{N_k \times p}$ Observations used for training the k^{th} event category.

$\tilde{\boldsymbol{y}}_p$ Specifies a $\tilde{\boldsymbol{y}}_{\tilde{p}}$ with no missing values, where $\tilde{p} = p$.

$\tilde{\boldsymbol{y}}_{\tilde{p}}$ Vector of observations from a new event with unknown category of length $\tilde{p}$.

$\boldsymbol{Y}_{N_k^m \times (p-1)}$ Subset of the first N_k^m rows of $\boldsymbol{Y}_{N_k \times p}$, with the first column removed.

$\boldsymbol{y}^-_{N_k^m \times 1}$ Vector subset of the missing entries within the first column of $\boldsymbol{Y}_{N_k \times p}$.

$\boldsymbol{Y}_{N_k^o \times (p-1)}$ Subset of rows $N_k^m + 1$ through N_k of $\boldsymbol{Y}_{N_k \times p}$, with the first column removed.

$\boldsymbol{y}^+_{N_k^o \times 1}$ Vector subset of the observed entries within the first column of $\boldsymbol{Y}_{N_k \times p}$.

Dimension and Indexing

$\boldsymbol{N}$ A vector containing all values $[N_1, \ldots, N_K]$.

ℓ Indexes the columns of $\boldsymbol{Y}_{N_k \times p}$.

$\tilde{p}$ Dimension of $\tilde{\boldsymbol{y}}_{\tilde{p}}$. $\tilde{p}$ must be an integer that satisfies $0 < \tilde{p} \leq p$.

K Number of event groups.

k Index of event groups.

N Total number of event observations in the training data.

N^{test} Total number of events used for testing to measure performance for the Monte Carlo experiments. Includes event observations that have no missing data and some missing data.

N^{train^-} Total number of training events that have one or more missing values. Use to describe the Monte Carlo experiments.

N^{train} Total number of training events that have zero missing values. Used to describe the Monte Carlo experiments.

N_k Number of event observations for the k^{th} event group.

N_k^m The number of missing entries in the ℓ^{th} column of $\boldsymbol{Y}_{N_k \times p}$. There are p values of N_k^m associated with each $\boldsymbol{Y}_{N_k \times p}$, with the index often implied to simplify notation but referred to as $(N_k^m)^\ell$ when necessary.

N_k^o The number of available entries in the ℓ^{th} column of $\boldsymbol{Y}_{N_k \times p}$. There are p values of N_k^o associated with each $\boldsymbol{Y}_{N_k \times p}$, with the index often implied to simplify notation but referred to as $(N_k^o)^\ell$ when necessary.

p Total number of discriminants considered for each event category.

Mathematical concepts

$\approx$ Approximately equal to.

$\boldsymbol{A} \otimes \boldsymbol{B}$ Kronecker product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

$\cup$ Union between two sets, herein often variables which do not have explicit dimensions.

$\forall$ For all elements, typically with respect to all elements in a set.

$\Gamma(x)$ The gamma function evaluated at x.

$\Gamma_p(x)$ The multivariate gamma function evaluated at x.

$\lfloor x \rfloor$ The value of the random variable x rounded down to the nearest integer.

$\mathbb{C}\text{ov}(\boldsymbol{x})$ Covariance matrix between elements of random variable $\boldsymbol{x}$.

$\mathbb{E}[x]$ Expected value of random variable x.

$\mathbb{R}^p$ The space containing all p dimensional real numbers.

$\text{logit}(x)$ Logit function on x, equal to $\ln(x) - \ln(1-x)$ where $x \in (0, 1)$.

$\text{logit}^{-1}(x)$ Logistic function on x, equal to $1/(1 + \exp\{-x\})$ for $x \in (-\infty, \infty)$.

$\text{tr}(\boldsymbol{X})$ Matrix trace of arbitrary square matrix $\boldsymbol{X}$.

$\text{vec}(\boldsymbol{X})$ Vectorization of matrix $\boldsymbol{X}$, where columns of $\boldsymbol{X}$ are sequentially arranged as the elements of a vector.

ϕ Statistical parameter to be inferred. Used in notation for Bayes' rule.

$I[\cdot]$ Indicator function, equal to one if the expression in place of $\cdot$ is true.

$I_x(a, b)$ The regularized incomplete beta function.

$x \gtrapprox y$ x greater than approximately y.

$x \lessapprox y$ x less than approximately y.

Distributions

$\boldsymbol{\Lambda}_{11}, \boldsymbol{\Lambda}_{12}, \boldsymbol{\Lambda}_{21}, \boldsymbol{\Lambda}_{22}$ Short hand for linear combinations used for the conditional matrix t-distribution equations.

$\boldsymbol{\mu}$ Mean vector of a multivariate normal distribution or multivariate t-distribution.

$\boldsymbol{\Omega}_{p\times p}$ $p \times p$ column spread matrix for a $N \times p$ matrix t-distributed random variable.

$\boldsymbol{\Sigma}$ Depending on context, either the covariance parameter of a multivariate normal distribution, scale matrix of a multivariate t-distribution, or row spread matrix of a matrix t-distribution.

$\boldsymbol{M}_{N\times p}$ $N \times p$ mean matrix for a $N \times p$ matrix t-distributed random variable.

$\boldsymbol{T}_{N\times p}$ Matrix variate t-distributed random variable with N rows and p columns.

$\boldsymbol{t}_{p\times 1}$ Multivariate t-distributed random variable.

$\mathbb{F}(\tilde{p}, \tilde{\nu})$ F-distribution with degrees of freedom $\tilde{p}$ and $\tilde{\nu}$.

$\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_{p\times p})$ Multivariate normal distribution of dimension p with mean $\boldsymbol{\mu}_p$ and covariance $\boldsymbol{\Sigma}_{p\times p}$.

$\mathcal{N}_{N,p}(\boldsymbol{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$ Matrix Normal distribution for a random $N \times p$ matrix with mean $\boldsymbol{M}$, row scale matrix $\boldsymbol{\Sigma}$, and column scale matrix $\boldsymbol{\Psi}$.

$\mathcal{W}^{-1}(a, \boldsymbol{B})$ Inverse Wishart distribution with degrees of freedom a and scale matrix $\boldsymbol{B}$.

$\tilde{\boldsymbol{\mu}}_k$ Predictive mean of $\tilde{\boldsymbol{y}}_{\tilde{p}}$ in the multivariate t-distribution.

$\tilde{\boldsymbol{\Sigma}}_k$ Predictive scale matrix of $\tilde{\boldsymbol{y}}_{\tilde{p}}$ in the multivariate t-distribution.

$F(\cdot)$ F-distribution cumulative distribution function.

m Degrees of freedom for a Matrix t-distribution.

$t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ Multivariate t-distribution with ν degrees of freedom, location parameter of $\boldsymbol{\mu}$, and scale matrix of $\boldsymbol{\Sigma}$. Dimension can be defined according to the dimension of $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$.

$T_{N\times p}(m, \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ Matrix variate t-distribution with degrees of freedom m, mean matrix $\boldsymbol{M}_{N\times p}$, row spread matrix $\boldsymbol{\Sigma}_{N\times N}$, and column spread matrix $\boldsymbol{\Omega}_{p\times p}$.

Other

$\mathbf{0}_{a\times b}$ Vector or matrix of zeros with a rows and b columns.

$\mathbf{1}_N$ Vector of ones of length N.

$\boldsymbol{I}_N$ Identity matrix of dimension N.

$\boldsymbol{J}_N$ Square matrix of ones with dimension N.

$\boldsymbol{J}_{a\times b}$ Rectangular matrix of ones with a rows and b columns.

Prior hyperparameters

α_k Prior concentration parameter for π_k.

$\boldsymbol{\alpha}$ Vector of prior concentration parameters for $\boldsymbol{\pi}$.

$\boldsymbol{\eta}$ Collection of all K values of $\boldsymbol{\eta}$.

$\boldsymbol{\eta}_k$ Prior mean of $\boldsymbol{\mu}_k$.

$\boldsymbol{\nu}$ Collection of all K values of ν_k.

$\boldsymbol{\Psi}$ Collection of all K values of $\boldsymbol{\Psi}_k$.

$\boldsymbol{\Psi}_k$ Prior covariance of $\boldsymbol{\mu}_k$ and prior scale matrix for $\boldsymbol{\Sigma}_k$.

$\boldsymbol{\theta}$ The full collection of prior hyperparameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$.

$\boldsymbol{\theta}_k$ Shorthand for the group hyperparameters $\{\boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k, \boldsymbol{\alpha}\}$ related to the prior distributions of parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$.

ν_k Prior degrees of freedom for $\boldsymbol{\Sigma}_k$.

Decision theory

$\boldsymbol{a}$ The set of actions $\{a_1, a_2\}$ for binary categorization or $\{a_1, \ldots, a_K\}$ for full K categorization.

$\boldsymbol{C}$ The loss matrix specifying the loss associated with each a_k along the columns and the true value of $\tilde{z}_1$ or $\tilde{\boldsymbol{z}}_K$ along the rows. Of dimension 2×2 for binary categorization and $K \times K$ for categorization into the specific categories used in training.

$\boldsymbol{L}(\tilde{\boldsymbol{z}}_K, \boldsymbol{a})$ Vector loss evaluated for each action considered. Vector values of length equal to the number of elements in $\boldsymbol{a}$.

$\delta(a_1)$ Indicator function for binary decisions. Equal to 1 when a_1 is the minimum expected loss action.

Φ A p-value calculated for the typicality index, with variability resulting from random draws of $\boldsymbol{Y}_k^-$. Conditioning on $k, \boldsymbol{Y}_k^+, \boldsymbol{\theta}_k$ implied.

$\tilde{\alpha}$ Significance level used for hypothesis testing.

a_k The k^{th} categorization action.

$C_{a,b}$ The element in the a^{th} row and b^{th} column of $\boldsymbol{C}$.

$F^{\Phi}(\tilde{\alpha})$ Cumulative distribution function of Φ evaluated at $\tilde{\alpha}$.

$p(\Phi)$ Distribution of random p-values, with variability resulting from random draws of $\boldsymbol{Y}_k^-$. Conditioning on $k, \boldsymbol{Y}_k^+, \boldsymbol{\theta}$ implied.

Z Shorthand for the test statistic $(\tilde{\boldsymbol{y}}_{\tilde{p}} - \tilde{\boldsymbol{\mu}}_k)^\top \tilde{\boldsymbol{\Sigma}}_k^{-1} (\tilde{\boldsymbol{y}}_{\tilde{p}} - \tilde{\boldsymbol{\mu}}_k)/\tilde{p}$.

B The Statistical Model

B.1 Data Likelihood

Let $(\boldsymbol{y}_k^i)_{p\times 1}$ be the i^{th} of N_k observations of p discriminants from the k^{th} event where $k \in \{1, \ldots, K\}$. Each observation in the training data can be indexed as such. There is a total of N_k observations for each event category, making the total number of event data $N = \sum_{k=1}^K N_k$. We assume that each $(\boldsymbol{y}_k^i)_{p\times 1}$ is obtained from a multivariate normal distribution with a mean vector $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. Any random vector $\boldsymbol{y}_p$ marginally drawn from the data is assumed to be sampled from a mixture of normal distributions,

$$\boldsymbol{y}_p \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{9}$$

where π_k is a marginal probability of an event category and $\sum_{k=1}^K \pi_k = 1$. Let $\boldsymbol{Y}_{N\times p}$ denote all data observations, $\boldsymbol{\mu} \equiv \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$, $\boldsymbol{\Sigma} \equiv \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$, and $\boldsymbol{\pi} \equiv \{\pi_1, \ldots, \pi_K\}$. Because the event category for each $\boldsymbol{y}_k$ is already known, the likelihood is proportional to

$$\begin{aligned} p(\boldsymbol{Y}_{N\times p} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) &\propto \prod_{k=1}^{K} \prod_{i=1}^{N_k} \pi_k |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left(\boldsymbol{y}_k^i - \boldsymbol{\mu}_k\right)^\top \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{y}_k^i - \boldsymbol{\mu}_k\right) \right\} \\ &\propto \prod_{k=1}^{K} \pi_k^{N_k} |\boldsymbol{\Sigma}_k|^{-\frac{N_k}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N_k} \left(\boldsymbol{y}_k^i - \boldsymbol{\mu}_k\right)^\top \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{y}_k^i - \boldsymbol{\mu}_k\right) \right\} \end{aligned} \tag{10}$$

B.2 Prior Distributions

We choose conjugate prior (Hoff, 2009) distributions for their mathematical properties, advantageous for later derivations. A prior distribution is required for the parameters $\pi_k, \boldsymbol{\Sigma}_k$, and $\boldsymbol{\mu}_k$ associated with each event category. We specify the conditional priors

$p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k, \boldsymbol{\eta}_k)p(\boldsymbol{\Sigma}_k|\boldsymbol{\Psi}_k, \nu_k)$. $p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k, \boldsymbol{\eta}_k)$ is a multivariate normal distribution with a mean vector of $\boldsymbol{\eta}_k$ and a covariance of $\boldsymbol{\Sigma}_k$, while $p(\boldsymbol{\Sigma}_k|\boldsymbol{\Psi}_k, \nu_k)$ is an inverse Wishart distribution (Hoff, 2009; Gupta and Nagar, 2018) with a scale matrix of $\boldsymbol{\Psi}_k$ and degrees of freedom ν_k. The prior on the vector of mixture parameters $p(\boldsymbol{\pi}|\boldsymbol{\alpha})$ is a Dirichlet distribution (Ng, Tian, and Tang, 2011) with a parameter vector $\boldsymbol{\alpha}$ of length K.

The choice of these prior distributions allows for closed-form posterior distributions of the parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$, as well as closed form integrals for the predictive distributions $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\boldsymbol{Y}_{N_k\times p}, \boldsymbol{\Psi}_k, \boldsymbol{\eta}_k, \nu_k, \tilde{z}_k = 1)$ and $p(\tilde{z}_k = 1|\boldsymbol{Y}_{N\times p}, \boldsymbol{\alpha})$.

In our experiments, we use the default prior parameter settings from the `R` package `ezECM` for a logit transformation of the data. The default for $p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k, \boldsymbol{\eta}_k)$ is $\boldsymbol{\eta}_k = \mathbf{0}_{p\times 1}$. In combination with the values for $p(\boldsymbol{\Sigma}_k|\boldsymbol{\Psi}_k, \nu_k)$ of $\boldsymbol{\Psi}_k = \boldsymbol{I}_p$ and $\nu_k = p$, the marginal likelihood of the data becomes fairly diffuse and is in fact matrix variate Cauchy (Gupta and Nagar, 2018), where the expectation and the variance are undefined. For prediction, the degrees of freedom for category k then becomes $N_k + 1$. The setting of $\boldsymbol{\alpha}$ in the prior distribution $p(\boldsymbol{\pi}|\boldsymbol{\alpha})$ is $\boldsymbol{\alpha} = \mathbf{1}_K \times 1/2$, equivalent to a fairly non-informative Jeffreys and reference prior (Yang and Berger, 1996).

B.3 Marginal Likelihood of the Event Data

To reduce computation time, we choose when reasonable to utilize analytical solutions to integrals instead of numerical solutions. Deriving the marginal likelihood of the k^{th} event data $p(\boldsymbol{Y}_{N_k\times p}|\boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$ is useful for later deriving the predictive distribution $p(\tilde{\boldsymbol{y}}|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k\times p}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$ in Appendix B.5.1.

Each $p(\boldsymbol{Y}_{N_k\times p}|\boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$ requires integration over $\boldsymbol{\mu}_k \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_k$ over a space of symmetric positive definite matrices such that

$$p(\boldsymbol{Y}_{N_k\times p}|\boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k) = \int_{\boldsymbol{\Sigma}}\int_{\mathbb{R}^p} p(\boldsymbol{Y}_{N_k\times p}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)d\boldsymbol{\mu}_k d\boldsymbol{\Sigma}_k. \tag{11}$$

The stepwise solution to this integral requires the trace identities $tr(\boldsymbol{ABC}) = tr(\boldsymbol{CAB})$ for matrices conformal for multiplication (Harville, 1998), as well as the matrix determinant lemma

(Harville, 1998).

$$
\begin{aligned}
&p(\boldsymbol{Y}_{N_k\times p}|\boldsymbol{\eta}_k,\boldsymbol{\Psi}_k,\nu_k)\\
&=\int_{\boldsymbol{\Sigma}_k}\int_{\boldsymbol{\mu}_k} p(\boldsymbol{Y}_{N_k\times p}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k,\boldsymbol{\eta}_k)p(\boldsymbol{\Sigma}_k|\boldsymbol{\Psi}_k,\nu_k)d\boldsymbol{\mu}_k d\boldsymbol{\Sigma}_k\\
&\propto\int_{\boldsymbol{\Sigma}_k}\int_{\boldsymbol{\mu}_k}\left[|\boldsymbol{\Sigma}_k|^{-\frac{N_k}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{N_k}\left(\boldsymbol{y}_k^i-\boldsymbol{\mu}_k\right)^\top\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{y}_k^i-\boldsymbol{\mu}_k\right)\right\}|\boldsymbol{\Sigma}_k|^{-\frac{1}{2}}\right.\\
&\qquad\left.\exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}_k-\boldsymbol{\eta}_k\right)^\top\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{\mu}_k-\boldsymbol{\eta}_k\right)\right\}|\boldsymbol{\Sigma}_k|^{-\frac{(\nu_k+p+1)}{2}}e^{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Psi}_k\boldsymbol{\Sigma}_k^{-1}\right)}\right]d\boldsymbol{\mu}_k d\boldsymbol{\Sigma}_k\\
&\propto\int_{\boldsymbol{\Sigma}_k}|\boldsymbol{\Sigma}_k|^{-\frac{(N_k+\nu_k+p+1)}{2}}\\
&\qquad\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_k^{-1}\left(\begin{bmatrix}\boldsymbol{Y}_{p\times N_k}^\top & \boldsymbol{\eta}_k\end{bmatrix}\left(\boldsymbol{I}_{N_k+1}-\frac{1}{N_k+1}\boldsymbol{J}_{N_k+1}\right)\begin{bmatrix}\boldsymbol{Y}_{N_k\times p}\\ \boldsymbol{\eta}_k^\top\end{bmatrix}+\boldsymbol{\Psi}_k\right)\right)\right\}d\boldsymbol{\Sigma}_k\\
&\propto\left|\left(\boldsymbol{Y}_{N_k\times p}-\boldsymbol{1}_{N_k}\boldsymbol{\eta}_k^\top\right)^\top\left(\boldsymbol{I}_{N_k}-\frac{1}{N_k+1}\boldsymbol{J}_{N_k}\right)\left(\boldsymbol{Y}_{N_k\times p}-\boldsymbol{1}_{N_k}\boldsymbol{\eta}_k^\top\right)+\boldsymbol{\Psi}_k\right|^{-\frac{(N_k+\nu_k)}{2}}\\
&\propto\left|\boldsymbol{I}_{N_k}+\left(\boldsymbol{I}_{N_k}-\frac{1}{N_k+1}\boldsymbol{J}_{N_k}\right)\left(\boldsymbol{Y}_{N_k\times p}-\boldsymbol{1}_{N_k}\boldsymbol{\eta}_k^\top\right)\boldsymbol{\Psi}_k^{-1}\left(\boldsymbol{Y}_{N_k\times p}-\boldsymbol{1}_{N_k}\boldsymbol{\eta}_k^\top\right)^\top\right|^{-\frac{((\nu_k+1-p)+N_k+p-1)}{2}}
\end{aligned}
\tag{12}
$$

We can recognize the above as proportional to a matrix t-distribution, where $\boldsymbol{Y}_{N_k\times p}\sim T_{N_k,p}(\nu_k+1-p,\boldsymbol{1}_{N_k}(\boldsymbol{\eta}_k)^\top,\boldsymbol{I}_{N_k}+\boldsymbol{J}_{N_k},\boldsymbol{\Psi}_k)$. The choice of ν_k has an effect on the properties of this distribution and later derivations. If the degrees of freedom, ν_k+1-p, are equal to one, the distribution is matrix variate Cauchy (Gupta and Nagar, 2018). The degrees of freedom must be greater than 2 for the variance to be defined (Gupta and Nagar, 2018).

B.4 Conditional Missing Data Distributions

There is typically not a simple closed form for $p(\boldsymbol{Y}_k^-|\boldsymbol{Y}_k^+,\boldsymbol{\eta}_k,\boldsymbol{\Psi}_k,\nu_k)$, although we will cover the case in this section when a simple density is available. Details on the hyperparameters $\boldsymbol{\eta},\boldsymbol{\Psi}$, and $\boldsymbol{\nu}$ can be found in Appendices B.2 and B.3. Instead, we derive distributions of subsets of the missing data, conditional on all other data, which are available in closed form as recognizable densities. We choose each subset of $\boldsymbol{Y}_k^-$ to be the full set of missing entries found in a single column of $\boldsymbol{Y}_{N_k\times p}$. Depending on the application details ,using a single row may be advantageous, and in unlikely cases, the set can be matrix of missing entries instead of a vector. The end goal is implementation within a Gibbs sampler (Hoff, 2009; Robert and Casella, 1999) to obtain marginal samples from $p(\boldsymbol{Y}_k^-|\boldsymbol{Y}_k^+,\boldsymbol{\eta}_k,\boldsymbol{\Psi}_k,\nu_k)$. With the goal of numerical integration in mind, the derivation for the posterior distributions required for MCMC are as follows.

First, partition $\boldsymbol{Y}_{N_k\times p},\boldsymbol{\eta}_k$, and $\boldsymbol{\Psi}_k$ in a similar manner to Appendix C

$$
\begin{matrix}\boldsymbol{Y}_{N_k\times p}=\\ \\ \\ \end{matrix}\begin{matrix}\begin{bmatrix}\boldsymbol{y}^-_{N_k^m\times 1} & \boldsymbol{Y}_{N_k^m\times(p-1)}\\ \boldsymbol{y}^+_{N_k^o\times 1} & \boldsymbol{Y}_{N_k^o\times(p-1)}\end{bmatrix}\\ 1\qquad\quad (p-1)\end{matrix}\begin{matrix}N_k^m\\ N_k^o\\ \\ \end{matrix},\ \begin{matrix}\boldsymbol{\eta}_k=\\ \\ \\ \end{matrix}\begin{matrix}\begin{bmatrix}\eta_k^-\\ \boldsymbol{\eta}_k^+\end{bmatrix}\\ 1\qquad (p-1)\end{matrix},\begin{matrix}\boldsymbol{\Psi}_k=\\ \\ \\ \end{matrix}\begin{matrix}\begin{bmatrix}\Psi_{k(11)} & \boldsymbol{\Psi}_{k(12)}\\ \boldsymbol{\Psi}_{k(21)} & \boldsymbol{\Psi}_{k(22)}\end{bmatrix}\\ 1\qquad (p-1)\end{matrix}\begin{matrix}1\\ (p-1)\\ \\ \end{matrix}
\tag{13}
$$

Using this new notation, we want to derive the conditional distribution $p(\boldsymbol{y}^-_{N_k^m\times 1}|\boldsymbol{y}^+_{N_k^o\times 1}, \boldsymbol{Y}_{N_k^m\times(p-1)}, \boldsymbol{Y}_{N_k^o\times(p-1)}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$. $\boldsymbol{y}^-_{N_k^m\times 1}$ are the missing entries in a single column of $\boldsymbol{Y}_{N_k\times p}$. The number of missing entries in this column are N_k^m, and the number of observed entries in the same column are N_k^o such that $N_k^m + N_k^o = N_k$. The remaining $p-1$ columns are represented by the matrices $\boldsymbol{Y}_{N_k^m\times(p-1)}$ and $\boldsymbol{Y}_{N_k^o\times(p-1)}$. Although some of the entries in these matrices are in fact missing, we will condition on all entries represented by these matrices where the values are imputed in other steps of the MCMC. See Algorithm 1.

To obtain $p(\boldsymbol{y}^-_{N_k^m\times 1}|\boldsymbol{y}^+_{N_k^o\times 1}, \boldsymbol{Y}_{N_k^m\times(p-1)}, \boldsymbol{Y}_{N_k^o\times(p-1)}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$, first we use the conditional equations in Appendix C to obtain $p(\boldsymbol{y}^-_{N_k^m\times 1}, \boldsymbol{y}^+_{N_k^o\times 1}|\boldsymbol{Y}_{N_k^m\times(p-1)}, \boldsymbol{Y}_{N_k^o\times(p-1)}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$.

$$
\begin{aligned}
&p(\boldsymbol{y}^-_{N_k^m\times 1}, \boldsymbol{y}^+_{N_k^o\times 1}|\boldsymbol{Y}_{N_k^m\times(p-1)}, \boldsymbol{Y}_{N_k^o\times(p-1)}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k) = \\
&\quad T_{N_k,1}\left(\nu_k, \eta_k^-\mathbf{1}_{N_k} + \left(\begin{bmatrix}\boldsymbol{Y}_{N_k^m\times(p-1)}\\ \boldsymbol{Y}_{N_k^o\times(p-1)}\end{bmatrix} - \mathbf{1}_{N_k}(\boldsymbol{\eta}_k^+)^\top\right)\boldsymbol{\Psi}^{-1}_{k(22)}\boldsymbol{\Psi}_{k(21)},\right. \\
&\qquad \boldsymbol{I}_{N_k} + \boldsymbol{J}_{N_k} + \left(\begin{bmatrix}\boldsymbol{Y}_{N_k^m\times(p-1)}\\ \boldsymbol{Y}_{N_k^o\times(p-1)}\end{bmatrix} - \mathbf{1}_{N_k}(\boldsymbol{\eta}_k^+)^\top\right)\boldsymbol{\Psi}^{-1}_{k(22)}\left(\begin{bmatrix}\boldsymbol{Y}_{N_k^m\times(p-1)}\\ \boldsymbol{Y}_{N_k^o\times(p-1)}\end{bmatrix} - \mathbf{1}_{N_k}(\boldsymbol{\eta}_k^+)^\top\right)^\top, \\
&\qquad \left.\Psi_{k(11)} - \boldsymbol{\Psi}_{k(12)}\boldsymbol{\Psi}^{-1}_{k(22)}\boldsymbol{\Psi}_{k(21)}\right)
\end{aligned}
\tag{14}
$$

In the event that $N_k^m = N_k$, then no further conditioning is required. Otherwise, the conditional t-distribution equations in Appendix C are used again to find $p(\boldsymbol{y}^-_{N_k^m\times 1}|\boldsymbol{y}^+_{N_k^o\times 1}, \boldsymbol{Y}_{N_k^m\times(p-1)}, \boldsymbol{Y}_{N_k^o\times(p-1)}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$, given by

$$
\begin{aligned}
\boldsymbol{y}^-_{N_k^m\times 1}|\boldsymbol{y}^+_{N_k^o\times 1}, \boldsymbol{Y}_{N_k^m\times(p-1)}, \boldsymbol{Y}_{N_k^o\times(p-1)} \sim T_{N_k^m,1}\Big(&\nu_k + N_k^o, \boldsymbol{M}^{1r} + \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}(\boldsymbol{y}^+_{N_k^o\times 1} - \boldsymbol{M}^{2r}), \\
&\boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}, \\
&\Omega + (\boldsymbol{y}^+_{N_k^o\times 1} - \boldsymbol{M}^{2r})^\top\boldsymbol{\Lambda}_{22}^{-1}(\boldsymbol{y}^+_{N_k^o\times 1} - \boldsymbol{M}^{2r})\Big),
\end{aligned}
\tag{15}
$$

where

$$\boldsymbol{M}^{1r} = \eta_k^-\mathbf{1}_{N_k^m} + (\boldsymbol{Y}_{N_k^m\times(p-1)} - \mathbf{1}_{N_k^m}(\boldsymbol{\eta}_k^+)^\top)\boldsymbol{\Psi}^{-1}_{k(22)}\boldsymbol{\Psi}_{k(21)} \tag{16}$$

$$\boldsymbol{M}^{2r} = \eta_k^-\mathbf{1}_{N_k^o} + (\boldsymbol{Y}_{N_k^o\times(p-1)} - \mathbf{1}_{N_k^o}(\boldsymbol{\eta}_k^+)^\top)\boldsymbol{\Psi}^{-1}_{k(22)}\boldsymbol{\Psi}_{k(21)} \tag{17}$$

$$\boldsymbol{\Lambda}_{11} = (\boldsymbol{Y}_{N_k^m\times(p-1)} - \mathbf{1}_{N_k^m}(\boldsymbol{\eta}_k^+)^\top)\boldsymbol{\Psi}^{-1}_{k(22)}(\boldsymbol{Y}_{N_k^m\times(p-1)} - \mathbf{1}_{N_k^m}(\boldsymbol{\eta}_k^+)^\top)^\top + \boldsymbol{I}_{N_k^m} + \boldsymbol{J}_{N_k^m} \tag{18}$$

$$\boldsymbol{\Lambda}_{12} = (\boldsymbol{Y}_{N_k^m\times(p-1)} - \mathbf{1}_{N_k^m}(\boldsymbol{\eta}_k^+)^\top)\boldsymbol{\Psi}^{-1}_{k(22)}(\boldsymbol{Y}_{N_k^o\times(p-1)} - \mathbf{1}_{N_k^o}(\boldsymbol{\eta}_k^+)^\top)^\top + \boldsymbol{J}_{N_k^m\times N_k^o} \tag{19}$$

$$\boldsymbol{\Lambda}_{21} = \boldsymbol{\Lambda}_{12}^\top \tag{20}$$

$$\boldsymbol{\Lambda}_{22} = (\boldsymbol{Y}_{N_k^o\times(p-1)} - \mathbf{1}_{N_k^o}(\boldsymbol{\eta}_k^+)^\top)\boldsymbol{\Psi}^{-1}_{k(22)}(\boldsymbol{Y}_{N_k^o\times(p-1)} - \mathbf{1}_{N_k^o}(\boldsymbol{\eta}_k^+)^\top)^\top + \boldsymbol{I}_{N_k^o} + \boldsymbol{J}_{N_k^o} \tag{21}$$

$$\Omega = \Psi_{k(11)} - \boldsymbol{\Psi}_{k(12)}\boldsymbol{\Psi}^{-1}_{k(22)}\boldsymbol{\Psi}_{k(21)}. \tag{22}$$

B.5 Predictive Distributions

These Appendix sections detail the derivations for the densities on the right-hand side of

$$p(\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha}) = \frac{p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k\times p}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)p(\tilde{z}_k = 1|\boldsymbol{Y}_{N\times p}, \boldsymbol{\alpha})}{p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha})}, \tag{23}$$

where $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k\times p}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$ is the predictive distribution for $\tilde{\boldsymbol{y}}_{\tilde{p}}$, $p(\tilde{z}_k = 1|\boldsymbol{Y}_{N\times p}, \boldsymbol{\alpha})$ is the predictive distribution of category k, and $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha})$ is the predictive distribution marginalized over all K categories.

B.5.1 $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k\times p}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$

The predictive distribution of $\tilde{\boldsymbol{y}}_{\tilde{p}}$ can be found by solving for the integral

$$p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k\times p}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k) = \int_{\boldsymbol{\Sigma}}\int_{\mathbb{R}^p} p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{Y}_{N_k\times p}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k) d\boldsymbol{\mu}_k d\boldsymbol{\Sigma}_k, \tag{24}$$

and then by using the properties of the marginal multivariate t-distribution for $\tilde{p} < p$. However, because we already have $p(\boldsymbol{Y}_{N_k\times p}|\tilde{z}_k = 1, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$ and the properties of conditional distributions for matrix t-distributed random variables, we choose to exploit conditional probability for this illustration;
$p(\tilde{\boldsymbol{y}}_p, \boldsymbol{Y}_{N_k\times p}|\tilde{z}_k = 1, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k) = p(\tilde{\boldsymbol{y}}_p|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k\times p}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k) p(\boldsymbol{Y}_{N_k\times p}|\tilde{z}_k = 1, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$. Note that we switch to the notation $\tilde{\boldsymbol{y}}_p$ from $\tilde{\boldsymbol{y}}_{\tilde{p}}$ to indicate a full $\tilde{p} = p$ dimensional vector. Properties of the marginal distribution of a multivariate t-distribution are applied at the end for the the case where $\tilde{p} < p$.

Given the result in Appendix B.3 , we can write the joint distribution $p(\tilde{\boldsymbol{y}}_p, \boldsymbol{Y}_{N_k\times p}|\tilde{z}_k = 1, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$ as

$$p(\tilde{\boldsymbol{y}}_p, \boldsymbol{Y}_{N_k\times p}|\tilde{z}_k = 1, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k) \propto \\ \left|\boldsymbol{I}_{N_k+1} + \left(\boldsymbol{I}_{N_k+1} - \frac{1}{N_k+1+1}\boldsymbol{J}_{N_k+1}\right)\right. \\ \left.\left(\begin{bmatrix}\tilde{\boldsymbol{y}}_p^\top \\ \boldsymbol{Y}_{N_k\times p}\end{bmatrix} - \boldsymbol{1}_{N_k+1}\boldsymbol{\eta}_k^\top\right)\boldsymbol{\Psi}_k^{-1}\left(\begin{bmatrix}\tilde{\boldsymbol{y}}_p^\top \\ \boldsymbol{Y}_{N_k\times p}\end{bmatrix} - \boldsymbol{1}_{N_k+1}\boldsymbol{\eta}_k^\top\right)^\top\right|^{-\frac{((\nu_k+1-p)+N_k+1+p-1)}{2}}, \tag{25}$$

which is $T_{(N_k+1)\times p}(\nu_k + 1 - p, \boldsymbol{1}_{N_k+1}\boldsymbol{\eta}_k^\top, \boldsymbol{I}_{N_k+1} + \boldsymbol{J}_{N_k+1}, \boldsymbol{\Psi}_k)$. Then, the conditioning equations in Appendix C can be used to provide $p(\tilde{\boldsymbol{y}}_p|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k\times p}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$. Rearranging terms as well as using the relationship between the matrix t-distribution and the multivariate t-distribution in Appendix C produces a density proportional to a multivariate t-distribution, noted as $t_p(N_k + \nu_k + 1 - p, \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k)$, where $\tilde{\boldsymbol{\mu}}_k = (N_k\bar{\boldsymbol{y}}_k + \boldsymbol{\eta}_k)/(N_k + 1)$ and

$$\tilde{\boldsymbol{\Sigma}}_k = \frac{N_k+2}{(N_k+1)(N_k+\nu_k+1-p)}\left(\boldsymbol{\Psi}_k + (\boldsymbol{Y}_{N_k\times p} - 1_{N_k}\boldsymbol{\eta}_k^\top)^\top\left(\boldsymbol{I}_{N_k} - \frac{1}{N_k+1}\boldsymbol{J}_{N_k}\right)(\boldsymbol{Y}_{N_k\times p} - 1_{N_k}\boldsymbol{\eta}_k^\top)\right). \tag{26}$$

Additionally, $\bar{\boldsymbol{y}}_k = \boldsymbol{Y}_{N_k\times p}^\top\boldsymbol{1}_{N_k}/N_k$.

For the case when a new observation $\tilde{\boldsymbol{y}}_{\tilde{p}}$ occurs and $\tilde{p} < p$, the properties of a multivariate t-distribution (Kotz and Nadarajah, 2004) are used to find the marginal distribution of this shorter vector. In short, the marginal distribution of $\tilde{\boldsymbol{y}}_{\tilde{p}<p}$ has the same degrees of freedom as the distribution of $\tilde{\boldsymbol{y}}_{\tilde{p}=p}$, but the elements of the mean vector and scale matrix corresponding to the missing elements of $\tilde{\boldsymbol{y}}_p$ are removed.

B.5.2 $p(\tilde{z}_k = 1|\boldsymbol{Y}_{N\times p}, \boldsymbol{\alpha})$

This distribution can be thought of as the probability that one of the categories will occur before $\tilde{\boldsymbol{y}}_{\tilde{p}}$ is observed. In simple terms, it is the fraction of observations in the k^{th} category in the training data, modified by the prior parameter $\boldsymbol{\alpha}$. In the event that the training data is not an accurate reflection of the frequency of future events, choosing $p(\tilde{z}_k = 1|\boldsymbol{\alpha})$, uninformed by the training data, may be preferable. An option in such a scenario is $p(\tilde{z}_k = 1|\alpha_1 = \cdots = \alpha_K) = 1/K$, but any

subjective choice under the constraint $\sum_{i=1}^{K} p(\tilde{z}_k = 1|\boldsymbol{\alpha}) = 1$ is valid. Here, we focus on the result when conditioning on the training data.

The vector of weights $\boldsymbol{\pi}$ is present in the likelihood shown in Appendix B.1. However, from Equation (10), the joint posterior of the parameters can be written as

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}|\boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha}) = p(\boldsymbol{\pi}|\boldsymbol{Y}_{N\times p}, \boldsymbol{\alpha}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}), \tag{27}$$

meaning that $\boldsymbol{\pi}$ is conditionally independent of the other parameters. Subsequently, we note $p(\boldsymbol{\pi}|\boldsymbol{N}) = p(\boldsymbol{\pi}|\boldsymbol{Y}_{N\times p})$, where $\boldsymbol{N} = [N_1, \ldots, N_K]$, because the number of training observations in each event category is the only data that informs the posterior $p(\boldsymbol{\pi}|\boldsymbol{N}, \boldsymbol{\alpha})$, and therefore $p(\tilde{\boldsymbol{z}}_K|\boldsymbol{N}, \boldsymbol{\alpha})$. The posterior $p(\boldsymbol{\pi}|\boldsymbol{N}, \boldsymbol{\alpha})$ can be recognized as the Dirichlet distribution (Ng, Tian, and Tang, 2011) $\text{Dir}(N_1 + \alpha_1, \ldots, N_K + \alpha_K)$. A draw of the random variable $\boldsymbol{\pi}$ is the probability of each category, based on counts of each category from the likelihood and the choice of the vector $\boldsymbol{\alpha}$.

The variable encoding the *predictive category* is $\tilde{\boldsymbol{z}}_K$. The probability distribution $p(\tilde{\boldsymbol{z}}_K|\boldsymbol{\pi})$ is a categorical distribution with category probabilities of $\boldsymbol{\pi}$. With the form of $p(\tilde{\boldsymbol{z}}_K|\boldsymbol{\pi})$ and $p(\boldsymbol{\pi}|\boldsymbol{N}, \boldsymbol{\alpha})$ specified, $p(\tilde{\boldsymbol{z}}_K|\boldsymbol{Y}_{N\times p}, \boldsymbol{\alpha})$ is the result of integration

$$p(\tilde{\boldsymbol{z}}_K|\boldsymbol{N}, \boldsymbol{\alpha}) = \int_{[0,1]^K} p(\tilde{\boldsymbol{z}}_K|\boldsymbol{N}, \boldsymbol{\pi}, \boldsymbol{\alpha}) p(\boldsymbol{\pi}|\boldsymbol{N}, \boldsymbol{\alpha}) d\boldsymbol{\pi}. \tag{28}$$

The analytical form of the vector of probabilities $p(\tilde{\boldsymbol{z}}_K|\boldsymbol{N}, \boldsymbol{\alpha})$, and thereby a specific $p(\tilde{z}_k = 1|\boldsymbol{N}, \boldsymbol{\alpha})$, is found as

$$\begin{aligned}
p(\tilde{\boldsymbol{z}}_K|\boldsymbol{N}, \boldsymbol{\alpha}) &= \int_{[0,1]^K} \prod_{k=1}^{K} \left[\pi_k^{\tilde{z}_k}\right] \Gamma\left(\sum_{k=1}^{K} (N_k + \alpha_k)\right) \prod_{k=1}^{K} \frac{\pi_k^{N_k+\alpha_k-1}}{\Gamma(N_k + \alpha_k)} d\boldsymbol{\pi} \\
&= \frac{\Gamma\left(\sum_{k=1}^{K} (N_k + \alpha_k)\right)}{\prod_{k=1}^{K} \Gamma(N_k + \alpha_k)} \int_{[0,1]^K} \prod_{k=1}^{K} \pi_k^{N_k+\alpha_k+\tilde{z}_k-1} d\boldsymbol{\pi} \\
&= \frac{\Gamma\left(\sum_{k=1}^{K} (N_k + \alpha_k)\right)}{\prod_{k=1}^{K} \Gamma(N_k + \alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(N_k + \alpha_k + \tilde{z}_k)}{\Gamma\left(\sum_{k=1}^{K} (N_k + \alpha_k + \tilde{z}_k)\right)} \\
\tilde{z}_k = 1 \implies & \\
p(\tilde{z}_k = 1|\boldsymbol{N}, \boldsymbol{\alpha}) &= \frac{N_k + \alpha_k}{\sum_{k=1}^{K} (N_k + \alpha_k)},
\end{aligned} \tag{29}$$

where $\Gamma(\cdot)$ is the gamma function that has the property $\Gamma(x+1) = x\Gamma(x)$ utilized in the simplification. This result is a simplification of the Dirichlet-multinomial distribution (Ng, Tian, and Tang, 2011) for a single predictive draw, where a single $\tilde{z}_k = 1$ for an arbitrary k.

B.5.3 $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha})$

This density is a marginalization over all K categories to produce the total predictive density of observing $\tilde{\boldsymbol{y}}_{\tilde{p}}$ from those categories. Using the law of total probability, the result is simply

$$p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha}) = \sum_{k=1}^{K} p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k\times p}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k) p(\tilde{z}_k = 1|\boldsymbol{Y}_{N\times p}, \boldsymbol{\alpha}). \tag{30}$$

C Multivariate and Matrix t-Distributions

The properties of the multivariate t-distribution are explored in great detail in Kotz and Nadarajah (2004), with the relation to the matrix variate t-distribution summarized in Gupta and Nagar (2018). Herein, the main parameterization used in Kotz and Nadarajah (2004) is set as the standard, where a random vector $\boldsymbol{t}_{p\times 1}$ is said to be multivariate t-distributed, $t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if the probability density function is given as

$$\frac{\Gamma((\nu+p)/2)}{(\pi\nu)^{\frac{p}{2}}\Gamma(\nu/2)}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\left[1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right]^{-\frac{(\nu+p)}{2}}, \tag{31}$$

where $\boldsymbol{\mu}_{p\times 1}$ is a p dimensional scale vector, ν is the degrees of freedom parameter, $\boldsymbol{\Sigma}$ is a scale or correlation matrix, and $\Gamma(x)$ is the gamma function of x. When $\nu > 1$, $\mathbb{E}[\boldsymbol{t}] = \boldsymbol{\mu}$, and when $\nu > 2, \mathbb{C}\text{ov}(\boldsymbol{t}) = \nu\boldsymbol{\Sigma}/(\nu-2)$. Otherwise, these moments are undefined.

We utilize the form of the matrix variate t-distribution density in Gupta and Nagar (2018), with some slight changes in notation from this reference. The random variate $\boldsymbol{T}_{N\times p}$ is matrix variate t-distributed $T_{N\times p}(m, \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ with a probability density function

$$\frac{\Gamma_p[\frac{1}{2}(N+m+p-1)]}{\pi^{\frac{1}{2}Np}\Gamma_p[\frac{1}{2}(m+N-1)]}|\boldsymbol{\Sigma}|^{-\frac{p}{2}}|\boldsymbol{\Omega}|^{-\frac{N}{2}}|\boldsymbol{I}_N+\boldsymbol{\Sigma}^{-1}(\boldsymbol{T}-\boldsymbol{M})\boldsymbol{\Omega}^{-1}(\boldsymbol{T}-\boldsymbol{M})^\top|^{-\frac{(N+m+p-1)}{2}}, \tag{32}$$

where m is the degrees of freedom, $\boldsymbol{M}$ is a $N\times p$ location matrix, and $\boldsymbol{\Omega}_{p\times p}, \boldsymbol{\Sigma}_{N\times N}$ are positive definite symmetric scale matrices. Analogous to the multivariate t-distribution, $\mathbb{V}\text{ar}(\text{vec}(\boldsymbol{T})) = \boldsymbol{\Sigma}\otimes\boldsymbol{\Omega}/(m-2)$ but is undefined when $m \leq 2$ and $\text{vec}(\boldsymbol{T})$ is the vectorization operator. Additionally, note that $T_{1\times p}(\nu, \boldsymbol{\mu}^\top, \sigma, \boldsymbol{\Omega}) = t_\nu(\boldsymbol{\mu}, \sigma\boldsymbol{\Omega}/\nu)$.

The following theorem is taken from Gupta and Nagar (2018) and utilized in Algorithm 1. Let $\boldsymbol{A}_{N\times N}$ and $\boldsymbol{B}_{p\times p}$ be non-singular square matrices, and let $\boldsymbol{T}_{N\times p} \sim T_{N\times p}(m, \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$. Then, $\boldsymbol{A}_{N\times N}\boldsymbol{T}_{N\times p}\boldsymbol{B}_{p\times p} \sim T_{N\times p}(m, \boldsymbol{AMB}, \boldsymbol{A\Sigma A}^\top, \boldsymbol{B\Omega B}^\top)$.

Properties of the marginal and conditional distributions are reproduced here from Gupta and Nagar (2018). These properties are necessary for evaluating the predictive density of $\tilde{\boldsymbol{y}}_{\tilde{p}}$ for $\tilde{p} < p$ and obtaining conditional draws of $\boldsymbol{Y}^-$ given $\boldsymbol{Y}^+$ in Algorithm 1. For the random variable $\boldsymbol{T}_{N\times p} \sim T_{N\times p}(m, \boldsymbol{M}_{N\times p}, \boldsymbol{\Sigma}_{N\times N}, \boldsymbol{\Omega}_{p\times p})$, arbitrarily partition the matrices as follows:

$$\begin{gathered}
\boldsymbol{T}_{N\times p} = \begin{bmatrix}\boldsymbol{T}_{1r}\\ \boldsymbol{T}_{2r}\end{bmatrix}\begin{matrix}N_1\\ N_2\end{matrix} = \begin{bmatrix}\boldsymbol{T}_{1c} & \boldsymbol{T}_{2c}\end{bmatrix}, \boldsymbol{M}_{N\times p} = \begin{bmatrix}\boldsymbol{M}_{1r}\\ \boldsymbol{M}_{2r}\end{bmatrix}\begin{matrix}N_1\\ N_2\end{matrix} = \begin{bmatrix}\boldsymbol{M}_{1c} & \boldsymbol{M}_{2c}\end{bmatrix}\\
\begin{matrix}p_1 & p_2\end{matrix} \qquad\qquad \begin{matrix}p_1 & p_2\end{matrix}\\
\boldsymbol{\Sigma}_{N\times N} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\begin{matrix}N_1\\ N_2\end{matrix}, \boldsymbol{\Omega}_{p\times p} = \begin{bmatrix}\boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12}\\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22}\end{bmatrix}\begin{matrix}p_1\\ p_2\end{matrix}\\
\begin{matrix}N_1 & N_2\end{matrix} \qquad\qquad \begin{matrix}p_1 & p_2\end{matrix}
\end{gathered} \tag{33}$$

Then, the following marginal and conditional distributions of the partitions of $\boldsymbol{T}$ are distributed as

$$\boldsymbol{T}_{2r} \sim T_{N_2,p}(m, \boldsymbol{M}_{2r}, \boldsymbol{\Sigma}_{22}, \boldsymbol{\Omega}_{p\times p}) \tag{34}$$

$$\begin{aligned}\boldsymbol{T}_{1r}|\boldsymbol{T}_{2r} \sim T_{N_1,p}(&m + N_2, \boldsymbol{M}_{1r} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{T}_{2r} - \boldsymbol{M}_{2r}),\\ &\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Omega}_{p\times p} + (\boldsymbol{T}_{2r} - \boldsymbol{M}_{2r})^\top\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{T}_{2r} - \boldsymbol{M}_{2r}))\end{aligned} \tag{35}$$

$$\boldsymbol{T}_{2c} \sim T_{N,p_2}(m, \boldsymbol{M}_{2c}, \boldsymbol{\Sigma}_{N\times N}, \boldsymbol{\Omega}_{22}) \tag{36}$$

$$\begin{aligned}\boldsymbol{T}_{1c}|\boldsymbol{T}_{2c} \sim T_{N,p_1}(&m + p_2, \boldsymbol{M}_{1c} + (\boldsymbol{T}_{2c} - \boldsymbol{M}_{2c})\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21},\\ &\boldsymbol{\Sigma}_{N\times N} + (\boldsymbol{T}_{2c} - \boldsymbol{M}_{2c})\boldsymbol{\Omega}_{22}^{-1}(\boldsymbol{T}_{2c} - \boldsymbol{M}_{2c})^\top, \boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21})\end{aligned} \tag{37}$$

The last property of the multivariate t-distribution highly relevant to this application is a correction from Kotz and Nadarajah (2004), useful for the Bayesian typicality index detailed in Appendix D.3. Let the $\tilde{p}$ dimensional random vector $\tilde{\boldsymbol{y}}_{\tilde{p}} \sim t_{\tilde{\nu}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$; then, the quadratic form $Z = (\tilde{\boldsymbol{y}}_{\tilde{p}} - \tilde{\boldsymbol{\mu}})^\top\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\boldsymbol{y}}_{\tilde{p}} - \tilde{\boldsymbol{\mu}})/\tilde{p}$, conditioned on $\tilde{\boldsymbol{\mu}}$ is F distributed as $\mathbb{F}(\tilde{p}, \tilde{v})$.

D Decision Theory

D.1 Binary Decisions

When a binary decision needs to be made, some simplifications can be utilized to reduce computation time relative to categorization, especially for large K. One simplification is the result of

$$\begin{aligned}\sum_{k=2}^{K} p(\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha}) &= p(\tilde{z}_1 \neq 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha})\\ &= 1 - p(\tilde{z}_1 = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha}).\end{aligned} \tag{38}$$

Therefore, under the binary action space of deciding if $\tilde{\boldsymbol{y}}_{\tilde{p}}$ belongs in the first out of K categories, we only need to know $p(\tilde{z}_1 = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha})$. Then, given the loss matrix $\boldsymbol{C}_{2\times 2}$

$$\boldsymbol{C}_{2\times 2} = \begin{array}{c} \\ \tilde{z}_1 = 1 \\ \tilde{z}_1 \neq 1 \end{array} \begin{array}{c} \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{bmatrix} C_{1,1} & C_{1,2} \\ C_{2,1} & C_{2,2} \end{bmatrix} \end{array}, \tag{39}$$

the posterior expected loss becomes

$$\begin{aligned}\mathbb{E}[L(a_1, \tilde{z}_1)] &= \mathbb{E}[\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta}]C_{1,1} + (1 - \mathbb{E}[\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta}])C_{2,1}\\ \mathbb{E}[L(a_2, \tilde{z}_1)] &= \mathbb{E}[\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta}]C_{1,2} + (1 - \mathbb{E}[\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta}])C_{2,2},\end{aligned} \tag{40}$$

where $\boldsymbol{\theta} = [\boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha}]$. The posterior expected loss of a_1 can be simplified as

$$\mathbb{E}[L(a_1, \tilde{z}_1)] = (C_{1,1} - C_{2,1})\mathbb{E}[\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta}] + C_{2,1}, \tag{41}$$

and similarly, the posterior expected loss of a_2 can be simplified as

$$\mathbb{E}[L(a_2, \tilde{z}_1)] = (C_{1,2} - C_{2,2})\mathbb{E}[\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta}] + C_{2,2}. \tag{42}$$

Action a_1 is chosen when $\mathbb{E}[L(a_1, \tilde{z}_1)] < \mathbb{E}[L(a_2, \tilde{z}_1)]$. Equivalently,

$$(C_{1,1} - C_{2,1})\mathbb{E}[\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta}] + C_{2,1} < (C_{1,2} - C_{2,2})\mathbb{E}[\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta}] + C_{2,2} \tag{43}$$

Rearranging terms allows for the simplifications

$$\delta(a_1) = 1 \text{ if } \begin{cases} \mathbb{E}[\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta}] > \frac{C_{2,1}-C_{2,2}}{(C_{1,2}-C_{2,2}+C_{2,1}-C_{1,1})} & \text{if } (C_{1,2} - C_{2,2} + C_{2,1} - C_{1,1}) > 0 \\ \mathbb{E}[\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta}] < \frac{C_{2,1}-C_{2,2}}{(C_{1,2}-C_{2,2}+C_{2,1}-C_{1,1})} & \text{if } (C_{1,2} - C_{2,2} + C_{2,1} - C_{1,1}) < 0 \\ \text{undefined} & \text{if } (C_{1,2} - C_{2,2} + C_{2,1} - C_{1,1}) = 0 \end{cases} \tag{44}$$

Action a_1 has the minimum posterior expected loss, indicated by $\delta(a_1) = 1$, when the relevant inequality under $\delta(a_1)$ is satisfied. Notably, under 0-1 loss, when $C_{1,1} = C_{2,2} = 0$ and $C_{2,1} = C_{1,2} = 1$, a_1 provides minimum expected loss when $\mathbb{E}[\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\theta}] > \frac{1}{2}$. Binary decision making under 0-1 loss is therefore simple to evaluate because the only quantity required is $p(\tilde{z}_1 = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha})$.

D.2 Posterior Expected Loss with Missing Training Data

The notation $p(\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}_{N\times p}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha})$ has been used to express the general case of the probability of $\tilde{z}_k = 1$ given all the data and prior parameters. However, in application, we expect that there are some missing entries in $\boldsymbol{Y}_{N\times p}$, where with generality $\boldsymbol{Y}^+$ indicates the observed entries and $\boldsymbol{Y}^-$ indicates the missing entries.

In Algorithm 1 and Appendix B.4, details are provided for using MCMC to generate joint samples from $p(\boldsymbol{Y}_k^-|\boldsymbol{Y}_k^+, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$ for the k^{th} event category. Decisions are not made by directly using the imputed $\boldsymbol{Y}^-$ and instead made using only $\boldsymbol{Y}^+$ and the prior hyperparameters. For decisions where there are missing data entries, we therefore need each

$$\begin{aligned} &p(\tilde{z}_k = 1|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}^+, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha}) \\ &= \int \cdots \int p(\tilde{z}_k = 1, \boldsymbol{Y}_1^-, \ldots, \boldsymbol{Y}_K^-|\tilde{\boldsymbol{y}}_{\tilde{p}}, \boldsymbol{Y}^+, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha}) d\boldsymbol{Y}_1^- \ldots d\boldsymbol{Y}_K^- \\ &= \int \cdots \int \frac{p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_k^+, \boldsymbol{Y}_k^-, \boldsymbol{\theta}) p(\tilde{z}_k = 1|\boldsymbol{N}, \boldsymbol{\theta})}{p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\boldsymbol{Y}^+, \boldsymbol{\theta})} p(\boldsymbol{Y}_1^-, \ldots, \boldsymbol{Y}_K^-|\boldsymbol{Y}^+, \boldsymbol{\theta}) d\boldsymbol{Y}_1^- \ldots d\boldsymbol{Y}_K^-, \end{aligned} \tag{45}$$

where $\boldsymbol{\theta} = [\boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\nu}, \boldsymbol{\alpha}]$ and $\boldsymbol{N} = [N_1, \ldots, N_K]$, noting that

$$p(\boldsymbol{Y}_1^-, \ldots, \boldsymbol{Y}_K^-|\boldsymbol{Y}^+, \boldsymbol{\theta}) = \prod_{k=1}^{K} p(\boldsymbol{Y}_k^-|\boldsymbol{Y}_k^+, \boldsymbol{\theta}_k) \tag{46}$$

for $\boldsymbol{\theta}_k = [\boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k]$.

D.3 Bayesian Typicality Index

In calculating the typicality index (McLachlan, 2005) using B-ECM, one has to account for the fact that $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k\times p}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$ is a multivariate t-distribution instead of a multivariate normal distribution, as in classical ECM. We know that for the density $p(\tilde{\boldsymbol{y}}_{\tilde{p}}|\tilde{z}_k = 1, \boldsymbol{Y}_{N_k\times p}, \boldsymbol{\eta}_k, \boldsymbol{\Psi}_k, \nu_k)$, the quadratic form $(\tilde{\boldsymbol{y}}_{\tilde{p}} - \tilde{\boldsymbol{\mu}}_k)^\top \tilde{\boldsymbol{\Sigma}}_k^{-1} (\tilde{\boldsymbol{y}}_{\tilde{p}} - \tilde{\boldsymbol{\mu}}_k)/\tilde{p}$ conditioned on $\tilde{\boldsymbol{\mu}}_k$ is F distributed as $\mathbb{F}(\tilde{p}, \tilde{\nu}_k)$(Kotz and

Nadarajah, 2004). Let $Z = (\tilde{\boldsymbol{y}}_{\tilde{p}} - \tilde{\boldsymbol{\mu}}_k)^\top \tilde{\boldsymbol{\Sigma}}_k^{-1} (\tilde{\boldsymbol{y}}_{\tilde{p}} - \tilde{\boldsymbol{\mu}}_k)/\tilde{p}$. Then, the upper tail probability under the null F-distribution is given by

$$\begin{aligned} \int_Z^\infty f(x)dx &= F(\infty) - F(Z) \\ &= 1 - F(Z), \end{aligned} \tag{47}$$

where $f(\cdot)$ is the $\mathbb{F}(\tilde{p}, \tilde{\nu}_k)$ density function, and $F(\cdot)$ is the $\mathbb{F}(\tilde{p}, \tilde{\nu}_k)$ cumulative distribution function. $F(Z)$ is the regularized incomplete beta function in the form of

$$I_{\tilde{b}}\left(\frac{\tilde{p}}{2}, \frac{\tilde{\nu}_k}{2}\right) = \frac{\int_0^{\tilde{b}} x^{\frac{\tilde{p}}{2}-1}(1-x)^{\frac{\tilde{\nu}_k}{2}-1}dx}{\int_0^1 x^{\frac{\tilde{p}}{2}-1}(1-x)^{\frac{\tilde{\nu}_k}{2}-1}dx}, \tag{48}$$

where $\tilde{b} = (\tilde{\boldsymbol{y}}_{\tilde{p}} - \tilde{\boldsymbol{\mu}}_k)^\top \tilde{\boldsymbol{\Sigma}}_k^{-1} (\tilde{\boldsymbol{y}}_{\tilde{p}} - \tilde{\boldsymbol{\mu}}_k) / ((\tilde{\boldsymbol{y}}_{\tilde{p}} - \tilde{\boldsymbol{\mu}}_k)^\top \tilde{\boldsymbol{\Sigma}}_k^{-1} (\tilde{\boldsymbol{y}}_{\tilde{p}} - \tilde{\boldsymbol{\mu}}_k) + \tilde{\nu}_k)$.

With (48), we can calculate the p-value (47) related to $\tilde{\boldsymbol{y}}_{\tilde{p}}$ given $\boldsymbol{Y}_{N_k \times p}$ and the prior hyperparameters $\boldsymbol{\eta}_k, \boldsymbol{\Psi}_k$, and ν_k. When all training data is complete, Equation (48) evaluates to a point that can be used to judge the typicality of $\tilde{\boldsymbol{y}}_{\tilde{p}}$ in event category k. However, when $\boldsymbol{Y}_{N_k \times p}$ contains partial observations, there is uncertainty related to $\boldsymbol{Y}_k^-$, and therefore a distribution of the p-values. Samples from the distribution of p-values are readily obtained by plugging the MC samples of $\boldsymbol{Y}_k^-$ into Equation (48). Let Φ indicate a random p-value from the distribution of p-values $p(\Phi)$.

Next, we must decide to either reject or fail to reject $\tilde{\boldsymbol{y}}_{\tilde{p}}$ as being typical of event category k, given the user-specified significance level $\tilde{\alpha}$. When no data in $\boldsymbol{Y}_{N_k \times p}$ is missing, the next steps are the same as classical typicality and hypothesis testing. When there is a distribution on Φ, we have to decide what action to take for hypothesis testing, given $p(\Phi)$ and a loss function. Here, we examine the difference between choosing between the actions a_1 and a_2, rejection and failure to reject respectively, when the decision is made using minimum expected loss and simply testing if $\mathbb{E}[\Phi] < \tilde{\alpha}$.

First, observe that for 0,1 loss matrix

$$\boldsymbol{C}_{2\times 2}^{\Phi} = \begin{array}{c} \\ \Phi < \tilde{\alpha} \\ \Phi \geq \tilde{\alpha} \end{array} \begin{array}{c} \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array}, \tag{49}$$

the loss function of action a_1 is

$$\begin{aligned} L(\Phi, a_1) &= I[\Phi \geq \tilde{\alpha}] \\ &= I[1 - F(Z) \geq \tilde{\alpha}], \end{aligned} \tag{50}$$

where $I[\cdot]$ is the indicator function equal to 1 when the statement inside the brackets is true. The expected loss, equivalent to the expectation of a function of a random variable, is then

$$\begin{aligned} \mathbb{E}\left[L(\Phi, a_1)\right] &= \int_0^1 I[\Phi \geq \tilde{\alpha}] p(\Phi) d\Phi \\ &= 0 \times \int_0^{\tilde{\alpha}} p(\Phi) d\Phi + 1 \times \int_{\tilde{\alpha}}^1 p(\Phi) d\Phi \\ &= Pr(\Phi \geq \tilde{\alpha}). \end{aligned} \tag{51}$$

Similarly, the expected loss of a_2 is

$$\begin{aligned}\mathbb{E}\left[L(\Phi, a_2)\right] &= \int_0^1 I[\Phi < \tilde{\alpha}]p(\Phi)d\Phi \\ &= 1 \times \int_0^{\tilde{\alpha}} p(\Phi)d\Phi + 0 \times \int_{\tilde{\alpha}}^1 p(\Phi)d\Phi \\ &= Pr(\Phi < \tilde{\alpha}).\end{aligned} \tag{52}$$

This implies that action a_1, rejection, is the minimum loss action when $Pr(\Phi \geq \tilde{\alpha}) < Pr(\Phi < \tilde{\alpha})$. This is equivalent to the condition $\frac{1}{2} < Pr(\Phi < \tilde{\alpha})$. $Pr(\Phi < \tilde{\alpha})$ is obtained from the CDF function of $p(\Phi)$ evaluated at $\tilde{\alpha}$, subsequently noted as $F^{\Phi}(\tilde{\alpha})$.

One may be tempted to evaluate $\mathbb{E}[\Phi] < \tilde{\alpha}$ as the decision criterion for rejection, using the loss matrix

$$\boldsymbol{C}^{\Phi}_{2\times 2} = \begin{array}{c} \\ \mathbb{E}[\Phi] < \tilde{\alpha} \\ \mathbb{E}[\Phi] \geq \tilde{\alpha} \end{array} \begin{array}{c} \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array}. \tag{53}$$

However, this second set of decision criteria will not always lead to the same action as in (49).

Lemma D.1. *The loss functions stipulated by* (49) *and* (53) *do not produce equivalent minimum expected loss for the action space, where* Φ *is distributed as* $p(\Phi)$*.*

Proof. For the minimum expected loss actions to be consistent between (49) and (53), there would be no density of $p(\Phi)$ and no values of $\tilde{\alpha}$ where the simultaneous conditions $F^{\Phi}(\tilde{\alpha}) = \int_0^{\tilde{\alpha}} p(\Phi)d\Phi > 0.5$ and $\mathbb{E}[\Phi] \geq \tilde{\alpha}$ occur. Such conditions would mean a_1 produces minimum loss for (49) and a_2 produces minimum loss for (53).

Under the following conditions, we can observe this inconsistency of the loss functions. Let $p(\Phi) = \text{Beta}(2, 5)$ and $\tilde{\alpha} = 0.2687$. In this case, $F^{\Phi}(\tilde{\alpha}) \approx 0.51 > 0.5$ and $\mathbb{E}[\Phi] = 2/7 \approx 0.2857 > \tilde{\alpha}$. □

As a more general note, the definition of expectation for real-valued random variables is $\int_0^{\mathbb{E}[\Phi]} F(\Phi)d\Phi = \int_{\mathbb{E}[\Phi]}^1 (1 - F(\Phi))d\Phi$ and not necessarily where $F(\Phi) = 0.5$. That is, the mean does not necessarily equal the median.

We choose the first decision criterion because this criterion is invariant to a one-to-one bijective transformation $g(\Phi)$ of the random variable Φ. As long as the $F^{g(\Phi)}(\tilde{\alpha})$ can be evaluated, the loss function will be consistent. However, $\mathbb{E}[g(\Phi)] < g(\tilde{\alpha})$ may be inconsistent with the condition $\mathbb{E}[\Phi] < \tilde{\alpha}$.

In this case, evaluating $F^{\Phi}(\tilde{\alpha})$ is implemented using the empirical CDF of the $T - B$ Monte Carlo samples

$$F^{\Phi}(\tilde{\alpha}) \approx \frac{\sum_{t=1}^{T-B} I[\Phi^{(t)} < \tilde{\alpha}]}{T - B}, \tag{54}$$

which should work well for 0,1 loss where we are evaluating if $F^{\Phi}(\tilde{\alpha}) < 0.5$ but may require large $T - B$, or other techniques, for loss functions that require evaluation of the CDF $F^{\Phi}(\tilde{\alpha}) \gtrapprox 0$ or $F^{\Phi}(\tilde{\alpha}) \lessapprox 1$.

Algorithm 3 Data generation for a single Monte Carlo experiment using synthetic data.

Require: $p, N^{\text{train}}, N^{\text{train}^-}, N^{\text{test}}, o$

$[\boldsymbol{\mu}_1^\star, \boldsymbol{\mu}_2^\star, \boldsymbol{\mu}_3^\star] \sim \mathcal{N}_{p,3}(\mathbf{0}_{p\times 3}, 0.25 \times \mathbb{I}_p \otimes \mathbb{I}_3)$
$[N_1, N_2, N_3] \sim \text{Multinomial}(N_{\text{train}} + N_{\text{train}^-} + N_{\text{test}}, [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}])$
$k^\star \sim \text{Categorical}([\frac{1}{3}, \frac{1}{3}, \frac{1}{3}])$
for $k \in \{1, 2, 3\}$ **do**
    $x \sim \text{Bernoulli}(0.5)$
    $b \leftarrow 2 - x$
    $\boldsymbol{\Sigma}_k^\star \sim \mathcal{W}^{-1}(p + 4, \mathbb{I}_p)$
    if $b = 2$ **then**
        $\dim(\boldsymbol{b}^1) \sim \text{Categorical}(\frac{1}{p-1}\mathbf{1}_{p-1}^\top)$
        $\dim(\boldsymbol{b}^2) \leftarrow p - \dim(\boldsymbol{b}^1)$
        $\boldsymbol{b}^1 \sim$ sample $\dim(\boldsymbol{b}^1)$ elements from $\{1, \ldots, p\}$ without replacement
        $\boldsymbol{b}^2 \leftarrow \{1, \ldots, p\} \notin \boldsymbol{b}^1$
        $(\boldsymbol{\Sigma}_k^\star)_{[\boldsymbol{b}^1, \boldsymbol{b}^2]} \leftarrow 0$
        $(\boldsymbol{\Sigma}_k^\star)_{[\boldsymbol{b}^2, \boldsymbol{b}^1]} \leftarrow 0$
    end if
    $\boldsymbol{Y}_{N_k\times p} \sim \mathcal{N}_{N_k,p}(\mathbf{1}_{N_k}(\boldsymbol{\mu}_k^\star)^\top, \mathbb{I}_{N_k} \otimes \boldsymbol{\Sigma}_k^\star)$
    $\boldsymbol{Y}_{N_k\times p} \leftarrow \text{logit}^{-1}(\boldsymbol{Y}_{N_k\times p})$
end for
$\mathcal{S}^{\text{all}} \equiv \{1^{(1)}, \ldots, N_1^{(1)}, 1^{(2)}, \ldots, N_2^{(2)}, \ldots, 1^{(3)}, \ldots, N_3^{(3)}\}$
while Any $\dim(\mathcal{S}_k^{\text{train}} : k \in \{1, 2, 3\}) < 2$ **do**
    $\mathcal{S}^{\text{test}} \sim$ sample N^{test} elements from $\mathcal{S}^{\text{all}}$ without replacement
    $\mathcal{S}^{\text{train}} \leftarrow \mathcal{S}^{\text{all}} \notin \mathcal{S}^{\text{test}}$
    $\mathcal{S}^{\text{train}^-} \sim$ sample N^{train^-} elements from $\mathcal{S}^{\text{train}}$ without replacement
    $\mathcal{S}^{\text{train}} \leftarrow \mathcal{S}^{\text{all}} \notin \{\mathcal{S}^{\text{test}}, \mathcal{S}^{\text{train}^-}\}$
end while
$\boldsymbol{Y}_{N_k^{\text{train}^-}\times p} \leftarrow (\boldsymbol{Y}_{N_k\times p})_{[\mathcal{S}_k^{\text{train}^-}, \cdot]} : \forall k \in \{1, 2, 3\}$
$\boldsymbol{Y}_{N^{\text{train}^-}\times p} \leftarrow \{\boldsymbol{Y}_{N_1^{\text{train}^-}\times p}, \boldsymbol{Y}_{N_2^{\text{train}^-}\times p}, \boldsymbol{Y}_{N_3^{\text{train}^-}\times p}\}$
for $i \in \{1, \ldots, N^{\text{train}^-}\}$ **do**
    $\mathcal{P}_i^{\star+} \sim \text{Categorical}(\frac{1}{p}\mathbf{1}_p^\top)$
    $\mathcal{P}_i^- = \{1, \ldots, p\} - \mathcal{P}_i^{\star+}$
    $\mathcal{P}_i^{\star-} \sim$ sample one element from $\mathcal{P}_i^-$
    $\mathcal{P}_i^- = \{1, \ldots, p\} - \{\mathcal{P}_i^{\star+}, \mathcal{P}_i^{\star-}\}$
    $\underline{\mathcal{P}}_i^- =$ sample $\lfloor (p-2) \times o \rfloor$ elements from $\mathcal{P}_i^-$ without replacement
    $(\boldsymbol{Y}_{N^{\text{train}^-}\times p})_{[i, \underline{\mathcal{P}}_i^-]} \leftarrow$ `NA`
end for
$\boldsymbol{Y}_{N_k^{\text{test}}\times p} \leftarrow (\boldsymbol{Y}_{N_k\times p})_{[\mathcal{S}_k^{\text{test}}]} : \forall k \in \{1, 2, 3\}$
$\boldsymbol{Y}_{N^{\text{test}}\times p} \leftarrow \{\boldsymbol{Y}_{N_1^{\text{test}}\times p}, \boldsymbol{Y}_{N_2^{\text{test}}\times p}, \boldsymbol{Y}_{N_3^{\text{test}}\times p}\}$
for $i \in \{1, \ldots, N^{\text{test}}\}$ **do**
    $\mathcal{P}_i^{\star+} \sim \text{Categorical}(\frac{1}{p}\mathbf{1}_p^\top)$
    $\mathcal{P}_i^- = \{1, \ldots, p\} - \mathcal{P}_i^{\star+}$
    $\underline{\mathcal{P}}_i^- =$ sample $\lfloor (p-1) \times o \rfloor$ elements from $\mathcal{P}_i^-$ without replacement
    $(\boldsymbol{Y}_{N^{\text{test}}\times p})_{[i, \underline{\mathcal{P}}_i^-]} \leftarrow$ `NA`
end for
$\boldsymbol{Y}_{N_k^{\text{train}}\times p} \leftarrow (\boldsymbol{Y}_{N_k\times p})_{[\mathcal{S}_k^{\text{train}}, \cdot]} : \forall k \in \{1, 2, 3\}$
$\boldsymbol{Y}_{N^{\text{train}}\times p} \leftarrow \{\boldsymbol{Y}_{N_1^{\text{train}}\times p}, \boldsymbol{Y}_{N_2^{\text{train}}\times p}, \boldsymbol{Y}_{N_3^{\text{train}}\times p}\}$

E Monte Carlo Experiments

E.1 Synthetic Data

Additional notation is required to describe the algorithm used for generating synthetic data. Much of the notation is the same as the other text provided; however, in some cases, it was necessary to change the notation in order to provide a better description for this context. The fraction of data missing from $\boldsymbol{Y}_{N^{\text{train}^-}\times p}$ and $\boldsymbol{Y}_{N^{\text{test}}\times p}$ is o. When missing observations are generated, the code ensures one of the discriminants is guaranteed to remain in order to reduce computation time relative to random deletion where the possibility of deleting all data for a single observation is a possibility. The full set of training data $\{\boldsymbol{Y}_{N^{\text{train}}\times p}, \boldsymbol{Y}_{N^{\text{train}^-}\times p}\}$ has a proportion of missing data less than o.

The additional nomenclature follows. $N^{\text{train}}, N^{\text{train}^-}$, and N^{test} are the integer sizes of the total training set of full observations, training set of partial observations, and testing set, respectively, which remain constant across MC iterations. Adding the k subscript, as in N_k^{train}, indicates the integer number specific to the k^{th} event category. $\boldsymbol{\mu}_k^\star$ is the true mean for the k^{th} event category. $\mathcal{N}_{p,n}(\boldsymbol{M}, \boldsymbol{\Sigma}\otimes\boldsymbol{\Psi})$ is the Matrix Variate Normal distribution (Gupta and Nagar, 2018). N_k, within the algorithm, is the total number of training and testing observations in the k^{th} event category. Categorical($\boldsymbol{p}^\top$) defines N draws from the categorical distribution; the number of categories is defined by the length of the vector of probabilities $\boldsymbol{p}$. $k^\star$ is the randomly selected category of interest, which is used to define the binary decision problem, calculate false negatives, and calculate false positives.

$\mathcal{W}^{-1}(m, \boldsymbol{\Psi})$ is the invese Wishart distribution, also known as the inverted Wishart distribution (Gupta and Nagar, 2018). $\dim(\boldsymbol{x})$ defines the dimension of a vector $\boldsymbol{x}$. The random variable b indicates the number of independent blocks in random covariance matrix $\boldsymbol{\Sigma}_k^\star$. If $b = 2$, then the random vectors $\boldsymbol{b}^1$ and $\boldsymbol{b}^2$ index the elements of $\boldsymbol{\Sigma}_k^\star$, which correspond to each block.

The notation $(\boldsymbol{X})_{[a,b]}$ indicates a rectangular subset of the matrix $\boldsymbol{X}$, where the selected rows are indicated by a and the selected columns are indicated by b. When the subset includes all columns, then the subset is noted as $(\boldsymbol{X})_{[a,\cdot]}$, and when a subset of a vector is taken, only one dimension is given in the brackets. $\text{logit}^{-1}(x)$ denotes the inverse logit function of x, where $\text{logit}^{-1}(x) = 1/(1+\exp\{-x\})$.

The set $\mathcal{S}^{\text{all}}$ indexes all data in all event categories. Similarly, $\mathcal{S}^{\text{train}}$, $\mathcal{S}^{\text{train}^{-1}}$, and $\mathcal{S}^{\text{test}}$ index the training and testing sets. Adding the k subscript, as in $\mathcal{S}_k^{\text{train}}$, indicates indexing specific to the k^{th} event category. The notation $i^{(k)}$ used in defining $\mathcal{S}^{\text{all}}$ indicates the i^{th} of N_k observations generated for the k^{th} event category.

Algorithm steps using any variant of the notation $\mathcal{P}$ are related to a particular discriminant. The subscript i used in any variant of $\mathcal{P}_i$ indexes observation i. $\mathcal{P}_i^{\star+}$ denotes a discriminant in observation i that is guaranteed to be part of the final data set. $\mathcal{P}_i^-$ is the set of discriminants in observation i that have a chance to be missing. $\mathcal{P}_i^{\star-}$ is a random sample from $\mathcal{P}_i^-$ that is designated as guaranteed to be missing. $\underline{\mathcal{P}_i^-}$ is the set of discriminants randomly designated to be missing for observation i within the final data set.

E.1.1 Reproducible Code

In an effort to improve reproducibility, the code used to produce Fig. 3 is provided as a vignette within the R package `ezECM`. In order to access the code, first, install the development version of the package from GitHub with the command
`remotes::install_github("lanl/ezECM")`, or from CRAN using

`install.packages("ezECM")`. Then, load the package with the command `library(ezECM)`. Last, display the vignette with `vignette("syn-data-code", package = "ezECM")`, where the code used for generating the data and statistical inference can be inspected.

E.2 Seismic Discriminant Experiment

Algorithm 4 Data set generation for a single Monte Carlo simulation using **Seismic Discriminant Data**

Require: $\boldsymbol{Y}_{N^- \times p}, \boldsymbol{Y}_{N^+ \times p}$

1: $\mathcal{S}^+ \equiv \{1^{(1)}, \ldots, (N^+_{\texttt{EX}})^{(1)}, 1^{(2)}, \ldots, (N^+_{\texttt{SEQ}})^{(2)}, \ldots, 1^{(3)}, \ldots, (N^+_{\texttt{DEQ}})^{(3)}\}$

2: $\mathcal{S}^{\text{train}} \sim$ sample $\lfloor 0.8 \times (N^+_{\texttt{EX}} + N^+_{\texttt{SEQ}}) \rfloor$ elements from $\mathcal{S}^+ - \{1^{(3)}, \ldots, (N^+_{\texttt{DEQ}})^{(3)}\}$ without replacement

3: $\mathcal{S}^{\text{train}} \leftarrow \{\mathcal{S}^{\text{train}}, \{1^{(3)}, \ldots, (N^+_{\texttt{DEQ}})^{(3)}\}\}$

4: $\mathcal{S}^- \equiv \{1^{(1)}, \ldots, (N^-_{\texttt{EX}})^{(1)}, 1^{(2)}, \ldots, (N^-_{\texttt{SEQ}})^{(2)}, \ldots, 1^{(3)}, \ldots, (N^-_{\texttt{DEQ}})^{(3)}\}$

5: $\mathcal{S}^{\text{train}^-} \sim$ sample $\left\lfloor (N^-_{\texttt{EX}} + N^-_{\texttt{SEQ}} + N^-_{\texttt{DEQ}}) \times \sqrt{p}/(1+\sqrt{p}) \right\rfloor$ elements from $\mathcal{S}^-$ without replacement

6: $\mathcal{S}^{\text{test}} \leftarrow \{\mathcal{S}^- - \mathcal{S}^{\text{train}^-}, \mathcal{S}^+ - \mathcal{S}^{\text{train}}\}$

7: $\boldsymbol{Y}_{N^{\text{train}} \times p} \leftarrow (\boldsymbol{Y}_{N^+ \times p})_{[\mathcal{S}^{\text{train}}, \cdot]}$

8: $\boldsymbol{Y}_{N^{\text{train}^-} \times p} \leftarrow (\boldsymbol{Y}_{N^- \times p})_{[\mathcal{S}^{\text{train}^-}, \cdot]}$

9: $\boldsymbol{Y}_{N^{\text{test}} \times p} \leftarrow \{\boldsymbol{Y}_{N^- \times p}, \boldsymbol{Y}_{N^+ \times p}\}_{[\mathcal{S}^{\text{test}}, \cdot]}$

In this real data set, the elements with missing data do not have to be selected and are truly missing. Three usable event categories are present in the data set—explosions, shallow earthquakes, and deep earthquakes—denoted as `EX`, `SEQ`, and `DEQ`, respectively. $\mathcal{S}^+$ indexes the fully observed data $\boldsymbol{Y}_{N^+ \times p}$, and $\mathcal{S}^-$ indexes the partially observed data $\boldsymbol{Y}_{N^- \times p}$. The `DEQ` category had only enough full observations to train the C-ECM model, and therefore, all full observations are included in training in line 3. The remainder of the notation is the same as what is used in Appendix E.1.

E.2.1 Data

After installation of the `ezECM` package, a global variable with the data used in the experiment can be specified using the function `read.csv(system.file("extdata", "ECM_validation_data_scrubbed.csv", package="ezECM"))[,c("Source.Type", "pLP", "pCF_EKA", "pCF_GBA", "pCF_WRA", "pCF_YKA")]`. Then, the final data set can be obtained by removing any event rows where all discriminants are missing as well as any `Source.Type` of `"MEX"`.