Misspecified Parameter Estimation for Heavy-tailed Noise Models: Student's t-distribution or Bivariance Gaussian Mixture?

Hamish McPhee[a,c], Jean-Yves Tourneret[b,c]

[a] *Telecommunications for Space and Aeronautics (TéSA), Toulouse, France*
[b] *University of Toulouse, Toulouse, France*
[c] *EURASIP Member,*

Abstract

Heavy-tailed noise models concern data contaminated with outliers. If the presence of outliers is not considered in the assumed model, the estimation performance of important parameters such as the mean and variance deteriorates. In this work, a misspecified Cramér-Rao bound is derived to show the reduced estimation performance when assuming a Gaussian distribution, although some portion of the data is generated by a Gaussian with inflated variance. This provides insight into one heavy-tailed distribution; the assumption of a different heavy-tailed distribution, the Student's t distribution, is also investigated. The Cramér-Rao Bound for joint estimation of the location, scale, and shape parameters of the Student's t-distribution is also derived to quantify the difference in performance when the number of degrees of freedom is unknown. Analysis of the corresponding maximum likelihood estimators and practical implementations of those estimators using the Expectation Maximization algorithm reveals the misspecified estimation performance when the contaminated data is not perfectly modeled by the chosen heavy-tailed distribution. Each of the assumptions is tested on realistic data with labeled outliers to identify the more advantageous assumption between a mixture of Gaussians and a Student's t distribution when the true distribution of measurements is not necessarily a specific heavy-tailed model.

Keywords: Heavy-tailed distributions, Misspecified estimation, Cramér Rao Bounds,

This work presents results obtained with the support of the French Space Agency (CNES), Thales Alenia Space, and the Telecommunications for Space and Aeronautics (TéSA) research laboratory. The authors would like to thank TéSA, Thales Alenia Space, and CNES for their financial and technical support, as well as their productive collaboration throughout the duration of this work. They would also like to extend a thank you to Prof. Daniel P. Palomar from the Hong Kong University of Science and Technology for support in obtaining and understanding the financial data used in this article. Finally, the authors are also grateful for the valuable discussions with Prof. Olivier Besson from ISAE-Supaero to clarify the gap in the current literature for joint estimation of the parameters of a Student's t-distribution.

1. Introduction

Heavy-tailed statistical distributions are useful in modeling data contaminated with outliers as a result of anomalous measurements or unexpected changes in system dynamics. Assuming uncontaminated data follows a Gaussian distribution, it makes sense to estimate the mean and variance with the corresponding Gaussian Maximum Likelihood Estimator (MLE). As anomalies occur unexpectedly, the assumption of a Gaussian distribution could be wrongfully maintained and as a result the estimation performance deteriorates. In this case, the Gaussian MLE is not the preferred estimator, i.e., the correctly specified estimator can have better performance by appropriately modeling the outliers. This article derives the estimation limits of the misspecified MLE when the true statistical model of the data has heavier tails than the assumed Gaussian distribution.

Some distributions considered in this work belong to the family of elliptically symmetric distributions [1] for which misspecified bounds for the location parameter have already been derived [2, 3]. A recent derivation for the misspecified parameter estimation of location and scale of a Student's t distribution has shown that the estimator of the scale parameter depends on the distribution that truly models the data [4]. This derivation addressed the specific example of the Student's t-distribution as a heavy-tailed distribution that accounts for the increased likelihood of outliers [5]. In this work, the choice of the Student's t-distribution is compared to another type of heavy-tailed distribution that is based on a Gaussian mixture model. In contrast to the Student's t-distribution, which is represented by a mixture of Gaussian distributions with common mean and variance that follows an inverse-gamma distribution [6], we consider the much more basic Bi-variance Gaussian Mixture (BGM) model, i.e., a finite mixture of two Gaussian distributions with equal mean and different variances. This ensures the resulting model is symmetric and heavy-tailed, and defines a proportion of contaminated data as samples of a Gaussian distribution with increased variance. This model is another relevant method of modeling contamination in robust estimation [7, 8, 9]. The interest of comparing these models is to see if there are preferences to assume one of those heavy-tailed models over the other, even if the true distribution of the contaminated data does not match the assumed model.

The misspecified estimation performance with BGM and Student's t models is evaluated using Misspecified Cramér-Rao Bounds (MCRBs), which indicate the expected asymptotic estimation performance if users make assumptions that do not align with the true properties of their measurements. These bounds are computed using the formula available in [10, 11, 12], which is an extension of the formula for computing the Cramér-Rao Bound (CRB), being the performance limit when the noise model is correctly specified. Other existing extensions of the CRB such as the Generalized MCRB (GMCRB) [13], the Constrained CRB or Constrained MCRB [14, 15] are relevant to particular signal models that do not have full rank measurements or involve constrained estimation. However, we focus on the special case of univariate heavy-tailed data as indicated in the real data section with some motivating examples in financial investments, clock time estimation, and medical data.

By comparing the CRB of heavy-tailed models to the appropriate MCRB, the loss in estimation performance due to assuming a Gaussian distribution can be quantified. A closed-form expression of the CRB for the location parameter of the Student's t-distribution is available [16] and the Fisher Information Matrix (FIM) for the joint estimation of location, scale, and shape was already derived in [5]. This work shows that the derived FIM can be inverted providing a closed-form CRB to joint estimation of the location, scale, and shape parameters. The resulting expression provides a mathematical definition of the effect of jointly estimating the number of degrees of freedom compared to fixing it to a known value. A numeric approximation is available for the CRB of the location parameter of a complex-valued BGM [17]. In this work we demonstrate when numerical approximation is also necessary for a real-valued BGM. Similarly, the MCRB when assuming a Student's t-distribution although the true distribution is a BGM is developed to the point of needing numerical approximations and saved for future work to confirm the numerical results. Note that a closed-form for the MCRB assuming a Gaussian noise when the noise is actually produced by a real BGM is more complicated to obtain.

Alongside the derived and existing bounds, appropriate estimators are also tested to verify the convergence of their Mean Square Errors (MSEs) to the new bounds. Expectation-Maximization (EM) algorithms are useful in obtaining robust estimators that converge to the MLE for the parameters of the two different heavy-tailed distributions [18, 19, 20, 21]. By testing two different assumptions on how to model contaminated data, we can observe the difference in performance if a particular heavy-tailed distribution is assumed although the outliers are rather modeled by some other law, as can be the case in real data. Examples of real heavy-tailed data are used to demonstrate applications where univariate random variables are observed with outliers or high volatility and to highlight the benefits of robust location and scale estimation in these applications. Some other practical applications that have considered misspecified estimation do not usually consider heavy-tails in the true distribution but instead misspecification in the model parameters themselves [22, 23, 24].

To finish this introduction, we would like to summarize the contributions of this work: i) Derivation of new MCRBs for the joint estimation of location and scale, assuming a Gaussian distribution while the noise is actually generated by a BGM, ii) Derivation of the CRB for joint estimation of the location, scale, and shape parameters when the noise is generated by a Student's t-distribution, the resulting expression is presented as a function of the CRB that assumes the number of degrees of freedom is known, iii) Analysis of the MLE of the relevant parameters, showing that there is a significant overlap in performance between the Student's t-distribution and the BGM, iv) The trade-off between MLEs and EM estimates for each heavy-tailed model is presented in terms of computational complexity and estimation accuracy, v) The robustness of assuming either the Student's t-distribution or the BGM is evaluated in the case of real data with known and labeled outliers. As a result, a conclusion is made about when it is preferred to assume a Student's t-distribution or a BGM for general contaminated data.

The paper is organized as follows: beginning in Section 2, the two heavy-tailed distributions are introduced as models for measurements with outliers. The CRB for joint estimation of the location, scale, and shape parameters of the Student's t-distribution is presented in Section 3 alongside the newly derived MCRB when the contaminated model is defined by the BGM. Section 4 validates the derived CRB and MCRB with simulated performance of the associated estimators. The estimators based on either of the two heavy-tailed models are compared to confirm the consequences of assuming the incorrect heavy-tailed model. Section 5 shows that modeling a known contaminated dataset with the investigated heavy-tailed distributions provides an improved performance compared to the misspecified Gaussian model. The known contaminated data is obtained in the context of an application that experiences outliers. The article concludes with a suggestion on which of the heavy-tailed distributions is preferred based on analysis of the robustness and computational complexity.

Notations: Bold lowercase symbols represent vectors, whereas bold uppercase symbols denote matrices. Superscripts T denotes the transpose operator. A $\tilde{(\cdot)}_p$ denotes a pseudotrue parameter for the true distribution defined by the likelihood p, which is short for $p(z;\boldsymbol{\theta})$ that defines the likelihood of observing sample z, and is defined by the parameters contained in parameter vector $\boldsymbol{\theta}$. A $\bar{(\cdot)}$ denotes a true parameter, considered to be the value of some parameter actually used to generate data. The mathematical expectation with respect to true likelihood function p is denoted $E_p\left[(\cdot)\right] = \int_{-\infty}^{\infty}(\cdot)p(z;\boldsymbol{\theta})dz$. A subscript G denotes a parameter from the Gaussian distribution, subscript T a parameter from the Student's t-distribution, and subscript GM a parameter for the Bivariance Gaussian Mixture. A term that depends on (p||q) is referring to the statistical similarity between true distribution p and assumed distribution q. The location parameter is denoted μ, scale parameter σ^2, and nuisance parameters refer to the shape parameter of the Student's t-distribution ν and the parameters ε, and α referred to as the proportion and scaling of the Bivariance Gaussian Mixture.

2. Heavy-tailed noise models

The heavy-tailed distributions explored for contaminated data are the Student's t-distribution and the BGM distribution. Each can be used to define the likelihood of independent and identically distributed samples z_i, where some of those samples are outliers. Probability density functions (PDFs) are defined for each of the models such that the likelihood of outlying values is higher than the likelihood given by the standard Gaussian model. The Student's t-distribution has a PDF parameterized by the location parameter μ_T, the scale parameter σ_T^2, and the shape parameter ν, also referred to as the number of degrees of freedom. The PDF of a Student-t distribution is defined as:

$$p_T(z_n;\boldsymbol{\eta}) = \frac{1}{\sqrt{\pi\nu\sigma_T^2}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{1}{\nu}\left(\frac{z_n-\mu_T}{\sigma_T}\right)^2\right)^{-\frac{(\nu+1)}{2}}, \quad z_n \in \mathbb{R}, \tag{1}$$

where $\boldsymbol{\eta} = [\mu_T, \sigma_T^2, \nu]$ is the vector of parameters of interest. The number of degrees of freedom specifies the weight of the tails of the distribution. As $\nu \to \infty$, the Student's t-distribution approaches the Gaussian distribution. Conversely, as $\nu \to 2$, the weight of the tails of the distribution increases corresponding to more probable outlying values. The variance of the Student's t-distribution is undefined for $\nu \leq 2$, so these cases are not considered in the later analysis of the estimation performance, which is evidently linked to the variance of the data.

In an attempt to eliminate the complexity linked to estimating the number of degrees of freedom in the Student's t-distribution, the heavy-tailed nature of the noise distribution can instead be assumed to be a mixture of Gaussian distributions. The BGM is a simple case of a Gaussian mixture distribution, i.e., with two modes corresponding to either nominal data or contaminated data. The contaminated data is considered as potentially nominal data with an inflated variance, effectively representing an increased noise. Hence, the contaminating mode has the same mean as the nominal mode μ_{GM} but with an increased variance $\alpha\sigma^2_{\mathrm{GM}}$ scaled by the factor $\alpha > 1$. The proportion of data that belongs to the contaminated mode is denoted as $0 < \varepsilon < 1$, where $\varepsilon = 0$ implies no contaminated data and $\varepsilon = 1$ corresponds to all observations coming from the Gaussian model with increased variance. The parameter vector for the BGM is defined as $\boldsymbol{\beta} = [\mu_{\mathrm{GM}}, \sigma^2_{\mathrm{GM}}, \varepsilon, \alpha]$ and the corresponding likelihood is:

$$p_{\mathrm{GM}}(z_n; \boldsymbol{\beta}) = \frac{1-\varepsilon}{\sqrt{2\pi\sigma^2_{\mathrm{GM}}}} e^{-\frac{(z_n-\mu_{\mathrm{GM}})^2}{2\sigma^2_{\mathrm{GM}}}} + \frac{\varepsilon}{\sqrt{2\pi\alpha\sigma^2_{\mathrm{GM}}}} e^{-\frac{(z_n-\mu_{\mathrm{GM}})^2}{2\alpha\sigma^2_{\mathrm{GM}}}}, \quad z_n \in \mathbb{R}, \tag{2}$$

where for future notation, the above is split into the linear combination of the PDF for the nominal Gaussian distribution with PDF $g(z_n; \mu, \sigma^2)$:

$$p_{\mathrm{GM}}(z_n; \boldsymbol{\beta}) = (1-\varepsilon) g(z_n; \mu_{\mathrm{GM}}, \sigma^2_{\mathrm{GM}}) + \varepsilon g(z_n; \mu_{\mathrm{GM}}, \alpha\sigma^2_{\mathrm{GM}}). \tag{3}$$

2.1. Expectation Maximization for Heavy-tailed models

The estimation of $\boldsymbol{\eta}$ or $\boldsymbol{\beta}$ from a set of measurements $\boldsymbol{z} = (z_1, ..., z_N)^T$ using maximum likelihood theory is not straightforward. One way to bypass this problem is to use the iterative EM algorithm that has received a lot of attention in the literature [18, 25, 9]. The EM algorithm estimates latent variables in the expectation step, then finds the parameters of interest that maximize a resulting complete likelihood function in the maximization step. The derivations of the corresponding EM algorithms for each heavy-tailed distribution are shown in Appendix A and Appendix B. The pseudocode that explains how to implement these EM algorithms is summarized in Algorithms 1 and 2. The stopping rule $d_k > 10^{-5}$ decides the resolution of the estimators by continuing the iterations until consecutive estimates differ by an amount specified by the user.

Algorithm 1 EM for Student's t-distribution

```
**Input**: $z_1, \cdots, z_N$
**Output**: $\hat{\mu}_k$, $\hat{\sigma}_k^2$, $\hat{\nu}_k$, $u_{1,k}, \cdots, u_{N,k}$
**Init.**: $\hat{\mu}_0 = \frac{1}{N}\sum_{n=1}^{N} z_n$, $\hat{\sigma}_0^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(z_n - \hat{\mu}_0\right)^2$, $\hat{\nu}_0 = 3$,
**while** $d_k > 10^{-5}$ **do**
    $u_{n,k} = \frac{\hat{\nu}_{k-1}+1}{\hat{\nu}_{k-1}+\frac{(z_n-\hat{\mu}_{k-1})^2}{\hat{\sigma}_{k-1}^2}}$,
    $\hat{\mu}_k = \frac{\sum_{n=1}^{N} u_{n,k} z_n}{\sum_{n=1}^{N} u_{n,k}}$,
    $\hat{\sigma}_k^2 = \frac{\sum_{n=1}^{N} u_{n,k}(z_n-\hat{\mu}_k)^2}{N}$,
    $\omega_{n,k} = \psi\left(\frac{\hat{\nu}_{k-1}+1}{2}\right) - \log\left(\frac{1}{2}\left(\hat{\nu}_{k-1} + \frac{(z_n-\hat{\mu}_k)^2}{\hat{\sigma}_k^2}\right)\right)$
    $\nu_k$ estimated as the solution of the following equation:

    $$N\left[\phi\left(\frac{\nu_k}{2}\right) - \phi\left(\frac{\hat{\nu}_{k-1}+1}{2}\right)\right] + \sum_{n=1}^{N}\left[u_{n,k} - \omega_{n,k} - 1\right] = 0. \tag{4}$$

    $d_k = |\hat{\mu}_k - \hat{\mu}_{k-1}|$
    $k = k + 1$
**end while**
with the digamma function $\psi(x) = \Gamma'(x)/\Gamma(x)$ and $\phi(x) = \psi(x) - \log(x)$.
```

Algorithm 2 EM for Bi-variance Gaussian Mixture

```
**Input**: $z_1, \cdots, z_N$
**Output**: $\hat{\mu}_k$, $\hat{\sigma}_k^2$, $\hat{\beta}_k$, $\hat{\alpha}_k$, $u_{1,k}, \cdots, u_{N,k}$
**Init.**: $\hat{\mu}_0 = \frac{1}{N}\sum_{n=1}^{N} z_n$, $\hat{\sigma}_0^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(z_n - \hat{\mu}_0\right)^2$, $\hat{\nu}_0 = 3$,
**while** $d_k > 10^{-5}$ **do**
    $u_{n,k} = \frac{\hat{\varepsilon}_{k-1} g\left(z_n; \hat{\mu}_{k-1}, \hat{\alpha}_{k-1}\hat{\sigma}_{k-1}^2\right)}{(1-\hat{\varepsilon}_{k-1}) g\left(z_n; \hat{\mu}_{k-1}, \hat{\sigma}_{k-1}^2\right) + \hat{\varepsilon}_{k-1} g\left(z_n; \hat{\mu}_{k-1}, \hat{\alpha}_{k-1}\hat{\sigma}_{k-1}^2\right)}$
    $w_{n,k} = 1 - u_{n,k} + \frac{u_{n,k}}{\hat{\alpha}_{k-1}}$
    $\hat{\mu}_k = \frac{\sum_{n=1}^{N} w_{n,k} z_n}{\sum_{n=1}^{N} w_{n,k}}$
    $\hat{\sigma}_k^2 = \frac{\sum_{n=1}^{N} w_{n,k}(z_n-\hat{\mu}_k)^2}{N}$
    $\hat{\varepsilon}_k = \frac{\sum_{n=1}^{N} u_{n,k}}{N}$
    $\hat{\alpha}_k = \frac{\sum_{n=1}^{N} u_{n,k}(z_n-\hat{\mu}_k)^2}{\hat{\sigma}_k^2 \sum_{n=1}^{N} u_{n,k}}$
    $d_k = |\hat{\mu}_k - \hat{\mu}_{k-1}|$
    $k = k + 1$
**end while**
```

3. New and Existing Bounds

Table 1 indicates the possible combinations of true and assumed distributions from those presented above alongside the relevant bounds. A general result for the bound of the mean has already been obtained for the case of assuming the Gaussian distribution while the true distribution could be any elliptically symmetric distribution [2, 3]. This result is verified with the closed-form expression for the MCRB of the location parameter when the true distribution is either Student's t or BGM. This provides the new derivation of the MCRB for the misspecified Gaussian estimator of the scale parameter, which depends on the true distribution. The MCRB for assuming a Gaussian distribution when the data follows Student's t-distribution was already derived in [4].This paper derives the MCRB for data following a BGM. In addition to this contribution, the joint estimation of the parameters of the Student's t-distribution is considered in the derivation of the CRB for joint location and scale estimation.

True / Assumed	Gaussian (p_{G})	BGM (p_{GM})	Student's t (p_T)
Gaussian (p_{G})	$\mathbf{CRB}_{\boldsymbol{\theta}}$	$\mathbf{MCRB}_{\boldsymbol{\theta}}(p_{\mathrm{GM}}\|\|p_{\mathrm{G}})*$	$\mathbf{MCRB}_{\boldsymbol{\theta}}(p_T\|\|p_{\mathrm{G}})$
BGM (p_{GM})	$\mathbf{MCRB}(p_{\mathrm{G}}\|\|p_{\mathrm{GM}})$	$\mathbf{CRB}_{\boldsymbol{\beta}}$	$\mathbf{MCRB}(p_T\|\|p_{\mathrm{GM}})$ *
Student's t (p_T)	$\mathbf{MCRB}(p_{\mathrm{G}}\|\|p_T)$	$\mathbf{MCRB}(p_{\mathrm{GM}}\|\|p_T)$ *	$\mathbf{CRB}_{\boldsymbol{\eta}}$ *

Table 1: Combinations of assumed and true distributions used to compute the bounds. MCRBs correspond to when the assumed model is incorrect, and CRBs when the assumed model is the true model. Those with an asterisk (*) were either derived in closed-form in this work or shown to require numerical approximations with relevant derivations.

For precision in the notations, the MCRB is presented with subscripts to refer to the parameters being estimated by the misspecified estimator. For example $\mathbf{MCRB}_{\boldsymbol{\theta}}$ refers to the matrix that contains the bounds (independent and joint) for each parameter in $\boldsymbol{\theta}$. As an example, in the Gaussian case, one has:

$$\mathbf{MCRB}_{\boldsymbol{\theta}}(p||q) = \begin{bmatrix} \mathrm{MCRB}_{\mu}(p||q) & \mathrm{MCRB}_{\mu,\sigma^2}(p||q) \\ \mathrm{MCRB}_{\sigma^2,\mu}(p||q) & \mathrm{MCRB}_{\sigma^2}(p||q) \end{bmatrix}. \tag{5}$$

The MCRB also specifies the true (p) and assumed (q) distributions in parentheses because several combinations are considered in this work. Specifically, $\mathbf{MCRB}_{\boldsymbol{\theta}}(p_T||p_{\mathrm{G}})$ refers to the MCRB when outliers are present, i.e., the true model follows a Student's t-distribution with PDF p_T but it is incorrectly assumed that the model is Gaussian with PDF p_{G}. Correspondingly, we can write $\mathbf{MCRB}_{\boldsymbol{\theta}}(p_{\mathrm{GM}}||p_{\mathrm{G}})$ for the BGM distribution.

3.1. Theory

The starting point to derive the MCRB is the definition of the Kullback Leibler Divergence (KLD), a statistical similarity measure between the true and assumed models defined as follows [26]:

$$D_{\mathrm{KL}}(p(\boldsymbol{z};\boldsymbol{\eta})||q(\boldsymbol{z};\boldsymbol{\theta})) = E_p\left[\log\left(\frac{p(\boldsymbol{z};\boldsymbol{\eta})}{q(\boldsymbol{z};\boldsymbol{\theta})}\right)\right], \tag{6}$$

where the subscript of E_p indicates that the expectation is computed with respect to the true PDF p, which remains the general notation for declaring any true distribution. To derive the MCRB, the pseudo-true parameters $\tilde{\boldsymbol{\theta}}_p = [\tilde{\mu}_p, \tilde{\sigma}_p^2]^T$ must first be derived, which are defined as the parameters that minimize the KLD between the true and assumed models

$$\tilde{\boldsymbol{\theta}}_p = \arg\min_{\boldsymbol{\theta}} \{D_{\mathrm{KL}}\} = \arg\min_{\boldsymbol{\theta}} \left\{ E_p \left[\log \left(\frac{p(\boldsymbol{z};\boldsymbol{\eta})}{q(\boldsymbol{z};\boldsymbol{\theta})} \right) \right] \right\}. \tag{7}$$

The subscript p is included to indicate that the pseudo-true parameters depend on the true distribution. Conveniently, the expression for the KLD can be simplified to remove the components that do not depend on $\boldsymbol{\theta}$:

$$\tilde{\boldsymbol{\theta}}_p = \arg\min_{\boldsymbol{\theta}} \{-E_p [\log (q(\boldsymbol{z};\boldsymbol{\theta}))]\}. \tag{8}$$

We obtain the following pseudo-true values for the two true distributions that model observations with outliers:

$$\tilde{\mu}_{p_T} = \bar{\mu}_T, \qquad \tilde{\mu}_{p_{\mathrm{GM}}} = \bar{\mu}_{\mathrm{GM}}, \tag{9}$$

$$\tilde{\sigma}^2_{p_T} = \bar{\sigma}^2_T \frac{\nu}{\nu-2}, \qquad \tilde{\sigma}^2_{p_{\mathrm{GM}}} = \bar{\sigma}^2_{\mathrm{GM}} \left((\bar{k}-1)\bar{\varepsilon} + 1 \right), \tag{10}$$

where the full derivations are detailed in Appendix C. It is noteworthy that the parameters that minimize the KLD between the misspecified Gaussian distribution and the distributions that model anomalous data are the mean and variance of the corresponding true distributions. The variance of the Student's t-distribution, and as a result, the pseudo-true scale parameter is only defined for numbers of degrees of freedom $\nu > 2$. This restriction must also exist for the bounds, so we assume that this condition is satisfied throughout this paper.

The largest lower bound for misspecified estimation under ML constraints is provided by the Huber sandwich covariance [27, 28], which is equivalent to the MCRB defined by

$$\mathbf{MCRB}_{\boldsymbol{\theta}}(p||q) = \mathbf{A}(q(\mathbf{z};\boldsymbol{\theta}), \tilde{\boldsymbol{\theta}}_p)^{-1} \mathbf{B}(q(\mathbf{z};\boldsymbol{\theta}), \tilde{\boldsymbol{\theta}}_p) \mathbf{A}(q(\mathbf{z};\boldsymbol{\theta}), \tilde{\boldsymbol{\theta}}_p)^{-1}, \tag{11}$$

with the matrices $\mathbf{A}(q(\mathbf{z};\boldsymbol{\theta}), \tilde{\boldsymbol{\theta}}_p)$ and $\mathbf{B}(q(\mathbf{z};\boldsymbol{\theta}), \tilde{\boldsymbol{\theta}}_p)$ defined by the curvature of the assumed joint likelihood $q(\mathbf{z};\boldsymbol{\theta})$ of the N observations $\mathbf{z} = [z_1, \cdots, z_N]$ [12, 29],

$$\mathbf{A}(q(\mathbf{z};\boldsymbol{\theta}), \tilde{\boldsymbol{\theta}}_p) = E_p \left[\left(\frac{\partial^2 \log(q(\mathbf{z};\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \right) \right]_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_p}, \tag{12}$$

$$\mathbf{B}(q(\mathbf{z};\boldsymbol{\theta}), \tilde{\boldsymbol{\theta}}_p) = E_p \left[\left(\left(\frac{\partial \log(q(\mathbf{z};\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}} \right) \left(\frac{\partial \log(q(\mathbf{z};\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}^T} \right) \right) \right]_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_p}, \tag{13}$$

where the subscript p refers to the true likelihood function. The above matrices are equivalent to the Fisher Information when the true distribution is equal to the assumed distribution and the Huber sandwich covariance provides the CRB. The parameters being estimated are substituted for the corresponding pseudotrue parameters because the misspecified estimator converges to the values that minimize the KLD from the true distribution [28]. The derivatives are evaluated at these values since we use the curvature at the asymptotic convergence of the misspecified estimator to define the asymptotic estimation performance of the misspecified estimator. In the case that each sample z_i is i.i.d., the above matrices are only concerned with the marginal PDF of a single observation $p(z_i;\boldsymbol{\theta})$. Under the assumption that of i.i.d. samples the MCRB is then defined as:

$$\mathbf{MCRB}_{\boldsymbol{\theta}}(p||p_{\mathrm{G}}) = \frac{1}{N}\mathbf{A}(q(z_i;\boldsymbol{\theta}),\tilde{\boldsymbol{\theta}}_p)^{-1}\mathbf{B}(q(z_i;\boldsymbol{\theta}),\tilde{\boldsymbol{\theta}}_p)\mathbf{A}(q(z_i;\boldsymbol{\theta}),\tilde{\boldsymbol{\theta}}_p)^{-1}. \tag{14}$$

3.2. Misspecified Cramér-Rao Bounds

The derivations of (12) and (13) are provided in Appendix D for a true distribution from each of the heavy-tailed models and assuming a Gaussian distribution. The results are then used to compute the MCRB for each of the two true distributions in Appendix E. The derivations for the Student's t-distribution have already been published in previous work [4] but are repeated here for the sake of comparison:

$$\mathbf{MCRB}_{\boldsymbol{\theta}}(p_T||p_{\mathrm{G}}) = \begin{bmatrix} \frac{\tilde{\sigma}^2_{p_T}}{N} & 0 \\ 0 & \left(\frac{\nu-1}{\nu-4}\right)\frac{2\tilde{\sigma}^4_{p_T}}{N} \end{bmatrix}. \tag{15}$$

The pseudo-true scale parameter is given by the second order moment of the true distribution, $\tilde{\sigma}^2_{p_T} = \frac{\nu}{\nu-2}\sigma^2_T$. Therefore, the bound for the misspecified estimation of the location parameter is equivalent to the CRB defined for a Gaussian distribution with equivalent variance. As expected, the misspecified bound for σ^2 also simplifies to the Gaussian CRB for $\nu \to \infty$ because the true distribution then simplifies to a Gaussian. The bound for the location parameter is undefined for $\nu \leq 2$ and the bound associated with σ^2 is undefined for $\nu \leq 4$. This coincides with the singularities for the second and fourth order moments of the Student's t-distribution, which are undefined for $\nu \leq 2$ and $\nu \leq 4$, respectively. These moments are necessary in the calculation of the MCRB, hence, the bounds are mathematically undefined when the moments are undefined. These singularities are a type that cannot be overcome by the GMCRB [13] as it involves diverging values instead of rank deficiency. Defining constraints on the number of degrees of freedom could be useful for an estimator targeting arbitrary contaminated data, and as a result the Constrained CRB or Constrained MCRB [14, 15] could be interesting to analyse in future work. Intuitively, the lower bound on the error of the estimators cannot be infinite or negative, so the final closed-form equation clearly indicates the restriction on ν. Similar results are visible when the true model is represented by the BGM:

$$\mathbf{MCRB}_{\boldsymbol{\theta}}(p_{\mathrm{GM}}||p_{\mathrm{G}}) = \begin{bmatrix} \frac{\tilde{\sigma}^2_{p_{\mathrm{GM}}}}{N} & 0 \\ 0 & \frac{Q(\phi)}{2(\phi+1)^2}\frac{2\tilde{\sigma}^4_{p_{\mathrm{GM}}}}{N} \end{bmatrix}, \tag{16}$$

where $\tilde{\sigma}^2_{p_{\mathrm{GM}}} = (\phi+1)\sigma^2_{\mathrm{GM}}$, $\phi = \varepsilon(\alpha-1)$ and $Q(\phi) = -\phi^2 + (3\alpha+1)\phi + 2$. The predefined domains of ε and α suggest that $0 \leq \phi \leq \alpha - 1$, which means that there are no values of ε

or α that result in an undefined bound. The values $\frac{\nu}{\nu-2}$, $(\phi+1)$, $\frac{\nu-1}{\nu-4}$, and $\frac{Q(\phi)}{2(\phi+1)^2}$ are always greater than one for the defined range of values for ν and ϕ so the bounds are always larger than the equivalent Gaussian CRB for non-contaminated data with the same signal power. Indeed, the ratio of these scaling factors with the scaling factors of the corresponding fully specified CRBs describe the losses in estimation performance when assuming a Gaussian distribution while the true noise model follows one of these two heavy-tailed distributions, or equivalently the gains when correctly specifying the model.

3.3. Cramér-Rao Bounds

In contrast to the newly derived MCRBs that define the performance limit for misspecified estimation, the CRB shows the asymptotic estimation performance for the correctly specified estimator. In the case of heavy-tailed distributions, this means the nuisance parameters and type of noise statistic are correctly modeled, and the corresponding estimator is used to estimate the parameters of interest. Knowledge of the CRB is convenient for assessing the performance of an MLE because it is known that the MSE of the correctly specified MLE converges to the CRB in the asymptotic regime under mild conditions [30]. The asymptotic regime refers to the case where the observed information has a sufficient number of samples or a reasonably high signal-noise ratio. A closed-form CRB allows a simple analysis on the expected estimation performance of a correctly-specified estimator, which can then be compared to a closed form MCRB to directly indicate the gain in estimation accuracy when correctly specifying the model. The derivations are of interest because once they have been validated with simulations, those simulations do not necessarily need to be repeated to understand the performance of the appropriate estimators.

The CRB for the estimation of the mean and variance of a Gaussian distribution is well-defined [30]:

$$\mathrm{CRB}_{\mu_{\mathrm{G}}} = \frac{\sigma_{\mathrm{G}}^2}{N}, \ \mathrm{CRB}_{\sigma_{\mathrm{G}}^2} = \frac{2\sigma_{\mathrm{G}}^4}{N}. \tag{17}$$

The CRB for the location parameter of the multivariate Student's t-distribution was derived in [16]. In this work, we consider the simpler univariate case to match applications of interest surrounding observations of a common scalar value, as demonstrated in Section 5. This bound is compared impartially to the CRB of the Gaussian distribution by including an additional scaling factor that normalizes the variance of the true distribution. The normalized variance is then considered the same and equal to $\frac{\nu}{\nu-2}\sigma_T^2 = \sigma_{\mathrm{G}}^2 = 1$, leading to the scaling of the bound defined in [16]:

$$\mathrm{CRB}_{\mu_T} = \left(\frac{\nu+3}{\nu+1}\right)\left(\frac{\nu-2}{\nu}\right)\frac{\sigma_T^2}{N}. \tag{18}$$

The estimation of the location parameter is known to be decoupled from the estimation of the scale and shape parameters [5]. This means the bound is applicable whether the other parameters are known or not. In practice, the joint estimation of the location, scale, and shape parameters corresponds to a particular bound for the scale parameter that is coupled with the bound for the number of degrees of freedom. However, in some practical applications, the number of degrees of freedom is treated as a tunable parameter to adjust the level of robustness with the value of ν set at a certain value instead of estimating it, resulting in a trade-off in performance under nominal conditions [5, 8]. Although the FIM is already known, the exact form of the CRB under joint estimation of the parameters of the Student's t-distribution is not known. Hence, we have derived a closed-form in Appendix F that highlights the modification to the bound when performing joint estimation of ν with the scale parameter. The derivation of the CRB is completed for a case with a k-variate Student's t-distribution with covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. Our contribution of a new, closed-form joint CRB is:

$$\mathrm{CRB}_{\sigma^2} = \frac{2\sigma_T^4}{N}\left(\frac{\nu+k+2}{\nu+k-1}\right)\left(1 - \frac{4\nu}{f(\nu,k)+4\nu}\right), \tag{19}$$

where $f(\nu,k) = 2\nu(\nu+k)^2(\nu+k-1)(\nu+k+2)\xi(\nu) + k(\nu+k+4)(\nu+k)(\nu+k-1)$, and $\xi(\nu) = \frac{1}{4}\left(\psi'\left(\frac{\nu+1}{2}\right) - \psi'\left(\frac{\nu}{2}\right)\right)$. The above closed-form expression is easily comparable to the bound that considers a known number of degrees of freedom, which is simply:

$$\mathrm{CRB}_{\sigma^2|\nu} = F_{\sigma^2}^{-1} = \frac{2\sigma_T^4}{N}\left(\frac{\nu+k+2}{\nu+k-1}\right), \tag{20}$$

where the subscript specifies that ν is known. We see the factor $C = 1 - \frac{4\nu}{f(\nu,k)+4\nu}$ scales the joint estimation bound with respect to the bound for the scale when ν is known. This factor converges to one when the number of degrees of freedom goes to infinity, and the magnitude of the bound is reduced when the number of degrees of freedom is small, hence improving estimation performance of the optimal estimator in the asymptotic regime. For the simple univariate case ($k = 1$), the function $f(\nu,k)$ is at a minimum, hence resulting in the maximum reduction in the lower bound when performing joint estimation. This result is relevant to the applications that consider univariate random variables in Section 5.

Some comments are appropriate to explain how the above results can be used in practice. When the true number of degrees of freedom is $\nu = 5$, the scaling factor has a magnitude of $C \approx 0.97$, suggesting a 3% reduction in estimation error for the scale parameter if we jointly estimate the number of degrees of freedom. Reducing the number of degrees of freedom to $\nu = 3$ only achieves a further reduction of 1%, as we approach the domain of undefined ν. Indeed, the improvement in the lower bound due to joint estimation reaches a maximum of around 4.2% when the number of degrees of freedom is $\nu = 2.37$ and using univariate samples. The minor difference in the theoretical lower bound when jointly estimating the scale and shape parameters is shown to be negligible. This result provides some evidence that it is usually acceptable to assume a fixed number of degrees of freedom.

The CRB for the joint parameter estimation of the BGM is not derived in this work, Appendix G indicates that the corresponding bounds require numerical approximations. However, the MSE of estimators that assume a BGM are still comparable to the bounds derived for the Student's t distribution to allow judgment of relative performance between the two heavy-tailed distributions.

4. Estimator Performance

This section is dedicated to the asymptotic performance of the fully specified and misspecified estimators in cases where the true distribution of the data is known. This analysis allows verification of the derived bounds as well as inference of the bounds for estimators without an associated closed form. The following is broken into three subsections, each presenting a toy example where the noise is generated with a known statistical model and the assumptions of the Gaussian, Student's t, and BGM are tested under each case.

The MCRB provides the minimum MSE that can be attained in the asymptotic regime (number of samples or SNR tend to infinity, denoted as →) by the MLE that assumes a misspecified model for the observed data [11, 10], i.e. for the location parameter:

$$E_p\left[(\hat{\mu}_q - \bar{\mu})^2\right] = \mathrm{MSE}_{\hat{\mu}_q}(p||q) \to \mathrm{MCRB}_{\mu}(p||q), \tag{21}$$

given the pseudo-true location parameter is equal to the true location parameter. The misspecified estimator of the location parameter is denoted as $\hat{\mu}_q$ and is made assuming that the distribution of the data follows a statistical model with PDF q while the true distribution has true PDF p and true location parameter $\bar{\mu}$. The asymptotic convergence is different in the case of a biased estimator, e.g., the misspecified Gaussian MLE for the scale parameter when the true distribution is heavy-tailed satisfies the following relation:

$$E_p\left[(\hat{\sigma}^2_{p_\mathrm{G}} - \bar{\sigma}^2)^2\right] = \mathrm{MSE}_{\hat{\sigma}^2_{p_\mathrm{G}}}(p||p_\mathrm{G}) \to \mathrm{MCRB}_{\sigma^2}(p||p_\mathrm{G}) + \left(\Delta\sigma^2\right)^2, \tag{22}$$

where the bias $\Delta\sigma^2 = \tilde{\sigma}^2_p - \bar{\sigma}^2_p$ is the difference between the pseudo-true and true scale parameters, which is known in closed form using the results of the derivations of pseudo-true parameters.

4.1. Experimental Setup

The MCRBs can be compared to the CRBs for the true distribution to show the expected loss in MSE when the presence of outliers is neglected. In Section 3, it is demonstrated that the misspecified estimation performance depends on the nuisance parameters that describe the intensity of the contaminating outliers. That is, the loss in estimation accuracy compared to the correctly specified model depends on the number of degrees of freedom for the Student's t-distribution or the proportion and scaling factors in the BGM. In the following simulations, the heavy-tailed distributions are parameterized such that the presence of anomalies is still significant, i.e., $\nu = 3$, $\varepsilon = 0.1$, $\alpha = 13.034$ when estimating the location parameter, and $\nu = 5$, $\varepsilon = 0.1$, and $\alpha = 7.670$ when estimating the scale parameter. Additional analysis of the estimation performance is provided with respect to the nuisance parameters, where ν and α are varied for t-distributed data and Gaussian mixture data, respectively, and ε is fixed. The different settings for estimating different parameters ensures the associated bounds are not undefined in the context of the Student's t-distribution, e.g., the MCRB for the location parameter is undefined for $\nu \leq 2$ and the MCRB for the scale parameter is undefined for $\nu \leq 4$. To select the corresponding values of parameters to generate the BGM noise, ε is fixed and the value of α is appropriately adjusted to ensure the second order central moment is equal in the two heavy-tailed models. This is necessary to ensure a fair comparison for the misspecified estimators that assume a different heavy-tailed model while the noise power is the same in each case. For the same contaminated dataset, either the Student's t-distribution or the BGM can be assumed to model the presence of outliers, without knowing the true distribution. As a result, appropriate estimators for ν, ε, and α are obtained that fit the actual dispersion of the data. Let σ^2 describe the fixed dispersion of an arbitrary contaminated dataset with the relation to the variance $\mathrm{var}(z) = f(\sigma^2)$, where f is a function depending on the true distribution. The equivalent nuisance parameters for the Student's t-distribution and BGM model can be fixed such that the variance of the data generated by either distribution is the same, i.e.

$$\frac{\nu}{\nu-2}\sigma^2 = (\varepsilon(\alpha-1)+1)\sigma^2 = f(\sigma^2). \tag{23}$$

The Gaussian MLEs of the location and scale parameters are computed using the sample mean and the sample variance, respectively. The correctly specified estimators for the two heavy-tailed distributions are obtained with an empirical MLE of all parameters defined in each model that uses a grid search with 2000 Monte Carlo iterations. The empirical MLE provides the optimum estimator but may not be feasible in practice, so the MLE is also approximated using the EM algorithms defined in Appendix A and Appendix B to show achievable performance in real time applications.

4.2. True Distribution: Student's t

The test data is generated using a Student's t-distribution with a low number of degrees of freedom to model the presence of outliers in the measurements. The correctly specified estimator is the MLE for the Student's t-distribution, which is known to converge to the associated CRB. The misspecified estimators are the Gaussian MLE and the BGM MLE whose MSEs will each converge to a different MCRB. The MCRB for the Gaussian assumption has already been derived in [4] where it was shown that the misspecified Gaussian estimator for the location parameter always converges to a higher MSE than the correctly specified estimator.

In the case that the contaminated data truly follows a Student's t-distribution, the assumption of the BGM can provide another heavy-tailed model with equivalent variance by appropriately adjusting the parameters ε and α, as defined in (23). Due to this equivalence of nuisance parameters for the two heavy-tailed distributions, it is hypothesized that assuming the BGM although the true model follows a Student's t-distribution (and vice versa) does not necessarily cause the error of the misspecified heavy-tailed estimator to deviate from the CRB. Therefore, the associated MCRBs are assumed to not deviate much from the CRB. This is tested by observing the misspecified estimation performance when the wrong heavy-tailed distribution is assumed. The result provides insight into an empirical definition of $\mathrm{MCRB}(p_T||p_{\mathrm{GM}})$ and $\mathrm{MCRB}(p_{\mathrm{GM}}||p_T)$, which are demonstrated in Appendix H as unobtainable in closed form. However, an interesting result on the derivation of the pseudotrue location parameter in these two cases shows that there is no bias in the misspecified estimate of the location parameter.

Since the MCRBs for the misspecified heavy-tailed distributions are not obtainable in closed form, we may as well observe the misspecified estimation performance by simulations. Making the assumption that the observations are generated by a BGM results in joint estimation of the parameters μ_{GM}, σ^2_{GM}, ε, and α, where the combination of the estimated values $\hat{\varepsilon}$ and $\hat{\alpha}$ provide an approximation of the true variance of the Student's t-distribution, meaning $\hat{\varepsilon}$ and $\hat{\alpha}$ should satisfy (23) within some margin of error. Since the condition of $\hat{\varepsilon}$ and $\hat{\alpha}$ satisfying (23) is not guaranteed, an error in those estimates could contribute to a misspecification between the two heavy-tailed distributions. As a result, using an EM algorithm instead of the true MLE is likely to result in a notable misspecification as well as a bias.

Fig. 1 shows the estimation performance for the estimators of the location parameter given data that is generated by a Student's t-distribution with $\nu = 3$. Firstly, we observe that the estimator of the location parameter assuming the Gaussian distribution (blue circles) has an MSE converging to the upper left element in (15), $\mathrm{MCRB}_{\mu_{\mathrm{G}}}(p_T||p_{\mathrm{G}})$ (blue triangles), verifying that the misspecified Gaussian MLE is limited by the bound derived in [4]. As expected, the MSE of the correctly specified MLE (black crosses) for data generated by a Student's t-distribution is lower than the MCRB, converging to (18), the appropriate CRB (yellow squares) derived in [16]. This highlights that there is an improvement in estimation accuracy by correctly specifying the model. The misspecified MLE assuming a BGM (magenta diamonds) obtains a similar MSE to the correctly specified MLE (black crosses) and is unbiased, in agreement with the derived pseudotrue location parameter. This supports the assumption of either of the two heavy-tailed distributions when estimating the location. Since the EM algorithms (green crosses and red diamonds) provide estimates close to the MLE of the location parameter, the corresponding MSEs are also similar for the MLE and the EM-based estimators. That being said, the error is slightly higher when assuming the opposing heavy-tailed model, although the difference is likely negligible in practice.

Next, Fig. 2 shows the MSE of the misspecified and correctly specified estimators for the scale parameter. The number of degrees of freedom is increased to $\nu = 5$ because the misspecified bound is undefined for $\nu \leq 4$. Fig. 2 confirms that the MSE of the misspecified Gaussian MLE (blue circles) converges to the MCRB plus the square of the bias (light blue triangles), which was not previously published in the original derivation of the MCRB [4]. The MSE of the misspecified estimator converges from below to the MCRB because the likelihood of outliers occurring and the magnitude of outliers in a Student's t-distribution with $\nu = 5$ are low enough that the difference between the true distribution and a Gaussian assumption is not significant at small sample sizes. The CRB for the scale parameter of the Student's t-distribution derived in this work (see (19)) is attained by the correctly specified MLE (black crosses) whose MSE converges to the CRB (yellow squares).

Fig. 2 also shows that the EM algorithm based on the Student's t-distribution (green crosses) performs significantly better than that defined by the BGM (red diamonds), although its MSE does not converge to the CRB (yellow squares). Since the EM algorithm considers knowledge of a latent variable in the maximization of the complete likelihood function, the MSE of the resulting estimator converges to a different bound, such as in other applications that use latent variables [31]. The misspecified MLE when assuming a

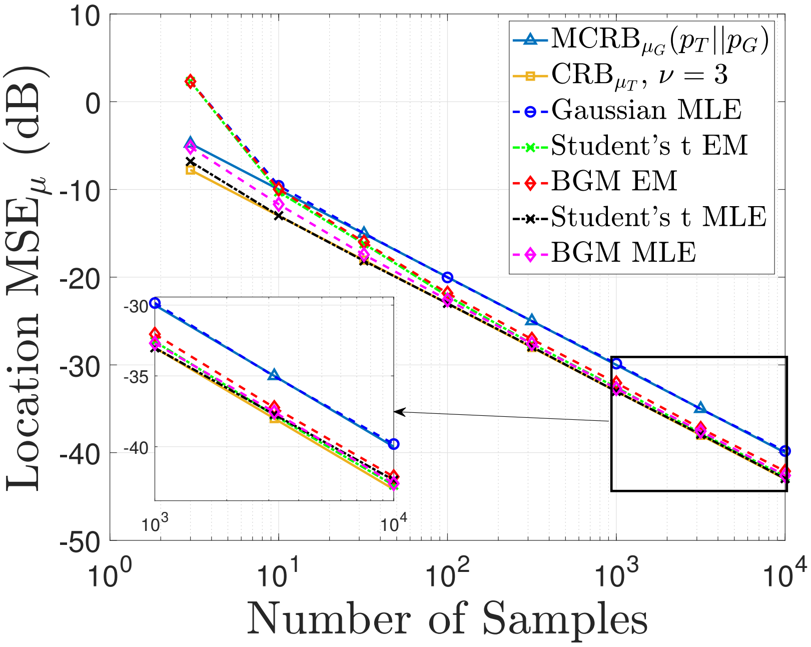

Figure 1: MSE performance of the correctly specified and misspecified estimators of the location parameter, when the data is generated by a Student's t-distribution.

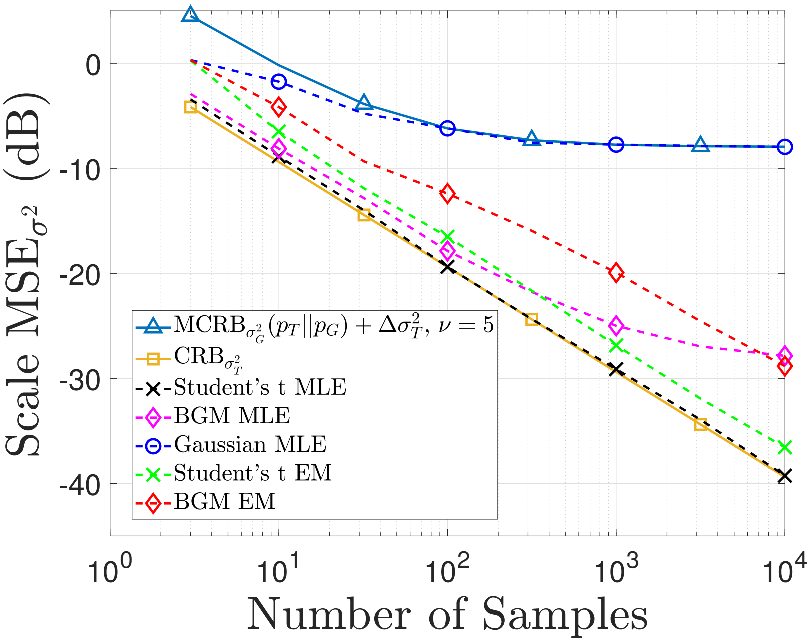

Figure 2: Convergence of the correctly specified and misspecified estimators of the scale parameter when the true distribution is a Student's t-distribution.

BGM (magenta diamonds) seems to converge asymptotically to a biased value, showing the loss when the incorrect heavy-tailed model is chosen, compared to the optimal MLE (black crosses) whose MSE converges to the CRB. As the pseudotrue scale parameter could not be found in closed form, we cannot compare this bias to a theoretical value. The extent to this misspecification between the heavy-tailed distributions is further explored in the following by making the same experiments with noise generated by a BGM to see if assuming the Student's t-distribution is also disadvantageous.

An analysis of how the bounds and estimation performance change according to the number of degrees of freedom is provided in Figs. 3 and 4. The most extreme case investigated is $\nu = 2.02$ to remain within the domain of finite variance. For each value of ν, the scale parameter is varied to ensure that the second order central moment of the distribution is equal to one and the number of samples is fixed to 1000. For this reason, we see the misspecified MLE that assumes a Gaussian distribution has an MSE (blue circles) that converges to the MCRB plus the bias squared (light blue triangles). The bias is zero for the location parameter so the associated bound (blue triangles Fig. 3) is constant for any number of degrees of freedom. The MSE of the Gaussian scale estimate (blue circles Fig. 4) decreases as the bias in the scale parameter decreases. We also observe that the MSE of the correctly specified estimator (black crosses) and the misspecified BGM estimator (magenta diamonds) converge to the CRB.

For a more extreme contamination (left of the figure) the correctly specified CRB (yellow squares) is reduced due to the scaling of $\sigma^2 = \frac{\nu-2}{\nu}$ making it directly proportional to the number of degrees of freedom. Comparing this to the Gaussian MLE represents the gain in performance achieved by basing the estimator on a heavy-tailed model instead of assuming a Gaussian model. When the tails are heavier, the gain is improved. In the most extreme case $\nu = 2.02$ there is an instability in the misspecified estimation of the scale parameter (blue circles in Fig. 4) because the variance of the Student' t distribution is theoretically approaching infinity and the kurtosis is undefined. However, this also introduces a difficulty in generating data that truly matches the Student's t -distribution specified by $\nu = 2.02$. As the variance has exploded so much, we need an increasingly high number of samples to realistically generate data that really matches the theoretical model. This is why we see a reduction in the MSE of the Gaussian location estimator (blue circles Fig. 3) below the MCRB at the most extreme case $\nu = 2.02$. The data being generated does not have a sufficient number of samples to represent the desired distribution so the effective variance is reduced and hence the estimators have a better performance than theoretically expected. This is also visible with the correctly specified estimator performing slightly better than the CRB at $\nu = 2.02$.

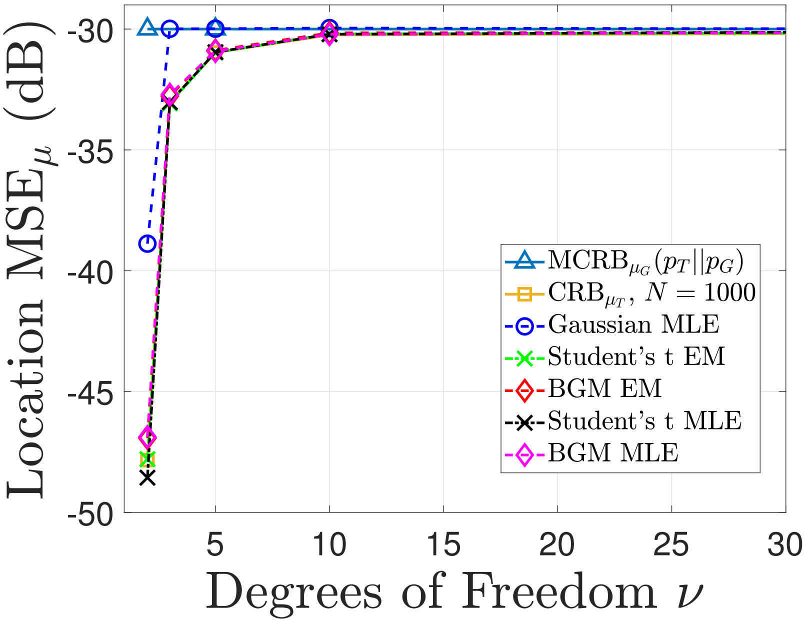

Figure 3: Performance of the correctly specified and misspecified estimators of the location parameter when the true distribution is a Student's t-distribution and we vary the heaviness of the tails (left heavier), with $\sigma_T^2 = \frac{\nu-2}{2}$ to keep the second order central moment equal.

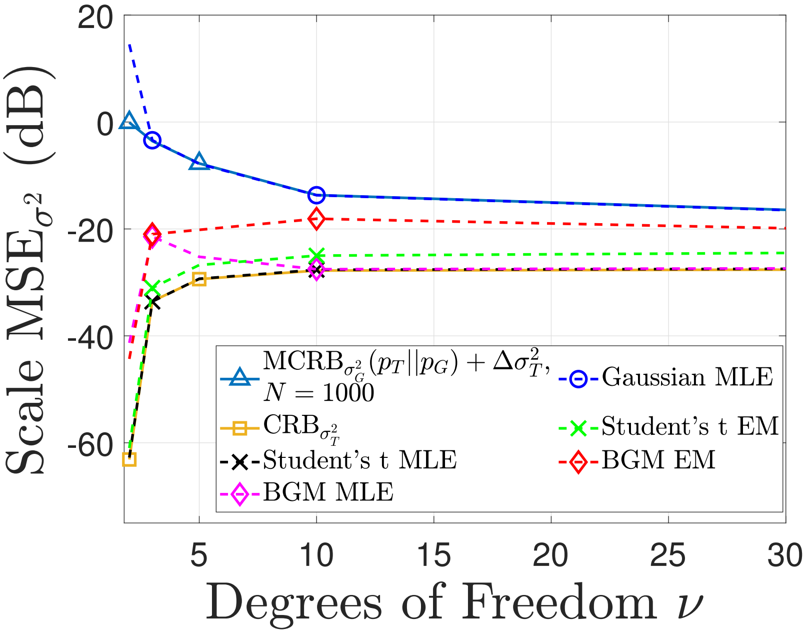

Figure 4: Performance of the correctly specified and misspecified estimators of the scale parameter when the true distribution is a Student's t-distribution and we vary the heaviness of the tails (left heavier), with $\sigma_T^2 = \frac{\nu-2}{2}$ to keep the second order central moment equal.

4.3. True Distribution: Bi-variance Gaussian Mixture

Fig. 5 shows the estimation performance of the estimators of the location parameter when the data is generated by a BGM model with $\varepsilon = 0.1$ and $\alpha = 13.034$. The MSE of the misspecified Gaussian MLE of the location parameter (blue circles) converges to the first element of (16), the new MCRB (red triangles), confirming the performance of the misspecified estimator. The correctly specified MLE (magenta diamonds) has an MSE close to (18), the CRB defined for a Student's t-distribution (yellow squares) despite using the estimator defined for a different heavy-tailed model. On the other hand, the misspecified MLE assuming the Student's t-distribution (black crosses) is able to obtain an MSE marginally lower than the MSE of the correctly specified MLE. This error is slightly below the CRB for $\nu = 3$, meaning the BGM generated noise was well approximated by a Student's t-distribution with $\nu \approx 3$. The EM-based estimators (green crosses and red diamonds) obtain effectively the same MSE as the correctly specified MLE, meaning the slight advantage of assuming the Student's t-distribution is not present when using the estimator based on the EM algorithm as an approximation of the true MLE for a BGM.

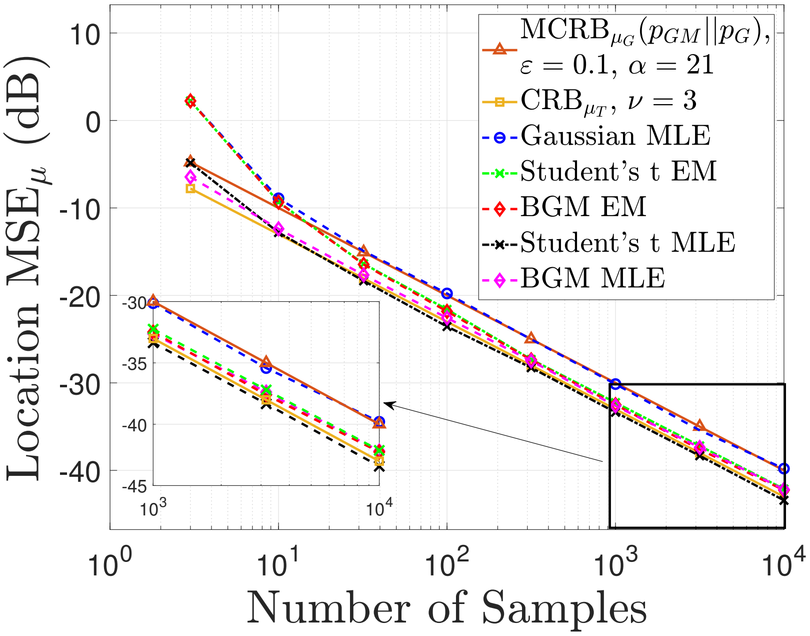

Figure 5: MSE performance of the correctly specified and misspecified estimators of the location parameter, when the data is generated by a BGM distribution.

Next, the results obtained regarding the estimation of the scale of the data are displayed in Fig. 6. The MSE of the misspecified Gaussian estimator (blue circles) once again converges to the derived MCRB (see lower right element of (16)) plus the bias squared (orange triangles). The convergence from below the bound is also repeated due to the similarity to the Gaussian distribution when the number of samples is small. Another equivalence between the heavy-tailed models is shown with the correctly specified MLE (magenta diamonds) whose MSE converges to the CRB for the Student's t distribution (yellow squares). The MSE of the misspecified MLE assuming a Student's t-distribution (black crosses) converges to the same bound at low sample sizes before diverging from the bound at higher sample sizes. Besides showing that the two estimators obtain similar performance at low sample sizes, this result implies that the derived bound for the Student's t-distribution has some relevance in showing the best possible performance for other heavy-tailed models with equivalent nuisance parameters. As was the case in the previous section, we observe a larger difference between the EM-based estimators and the MLEs when estimating the scale parameter. Although the noise is actually produced by a BGM, the assumption of the Student's t-distribution (green crosses) is still the preferred choice for smaller sample sizes before the estimator becomes biased, after which the BGM EM algorithm (red diamonds) performs better.

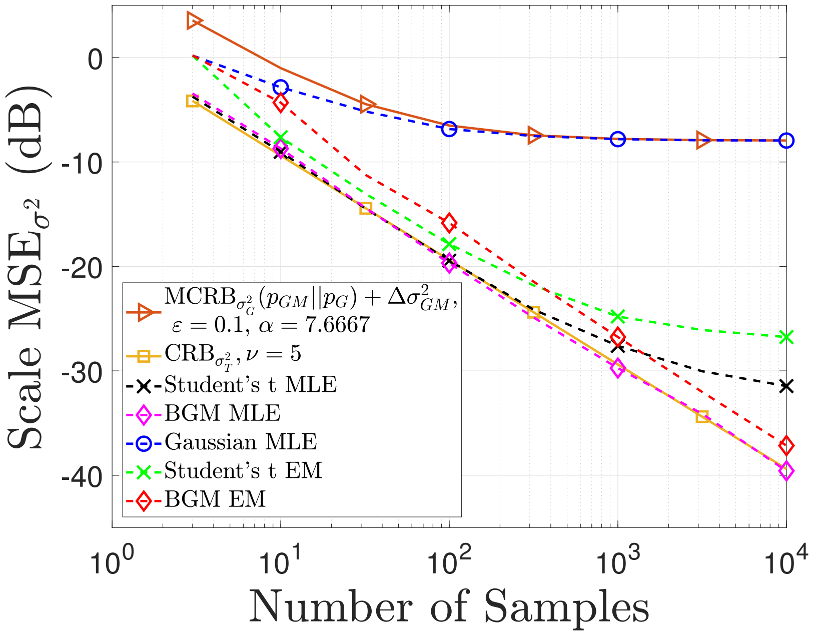

Figure 6: Convergence of the correctly specified and misspecified estimators of the scale parameter when the true distribution is a BGM distribution.

We continue with an analysis of how the bounds and estimation performance change according to the scaling of the contaminated variance α in the BGM with fixed contamination proportion $\varepsilon = 0.1$. Figs. 7 and 8 indicate the estimation performance for the location and scale parameters, respectively. The most extreme case investigated is $\alpha = 1001$ to match the variance of the most extreme Student's t-distribution for the same value of σ^2. For each value of α, the scale parameter is varied to ensure that the second order central moment of the distribution is equal to one and the number of samples is fixed to 1000.

The same convergences as discussed in the section with Student's t-distributed data are

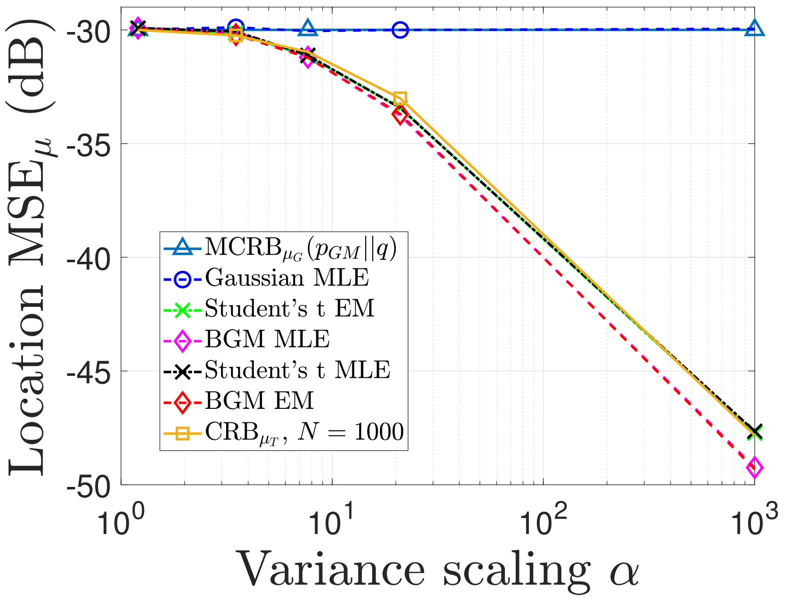

Figure 7: Performance of the correctly specified and misspecified estimators of the location parameter when the true distribution is a Bi-variance Gaussian Mixture and we vary the heaviness of the tails (right heavier), with $\sigma^2 = \frac{1}{(\varepsilon(\alpha-1)+1)}$ and $\varepsilon = 0.1$ to keep the second order central moment equal.

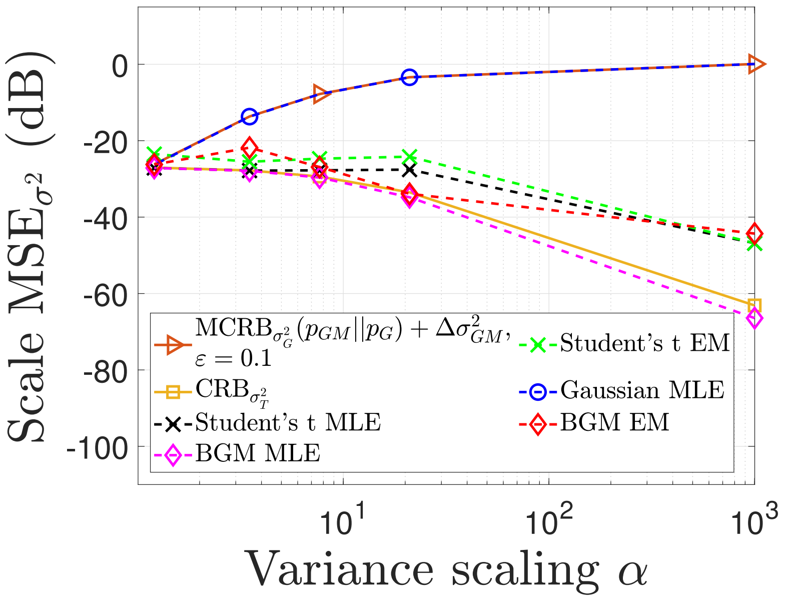

Figure 8: Performance of the correctly specified and misspecified estimators of the scale parameter when the true distribution is a Bi-variance Gaussian Mixture and we vary the heaviness of the tails (right heavier), with $\sigma^2_{\mathrm{GM}} = \frac{1}{(\varepsilon(\alpha-1)+1)}$ and $\varepsilon = 0.1$ to keep the second order central moment equal.

observed in Figs. 7 and 8 for the MSE of the Gaussian MLE (blue circles) to the MCRB (blue triangles) plus the bias squared when relevant. The heavy-tailed estimators perform better and better than the Gaussian MLE for a data of the same variance as the contamination increases. The CRB for the Student's t-distribution was earlier used as an approximation of the asymptotic lower bound on the MSE of the BGM estimator. Figs. 7 and 8 show that this is no longer a good approximation as α increases, the BGM estimator performs better than the Student's t estimator. This is not a surprise as the Student's t CRB is not based on the true distribution at all but is at least available in closed-form for comparisons.

4.4. Computational Complexity

Assessing the performance of the EM algorithms for the heavy-tailed distributions shows the best achievable error in a real-time application. It is also interesting to observe the time it takes to converge to an acceptable estimate. We consider that grid search MLE is always several orders of magnitude larger in computation time so we do not compare it to the EM algorithms. Fig. 9 shows the time it takes for the different EM algorithms to converge to their corresponding estimates for the same stopping rule. The average convergence time for the Student EM algorithm is shown by the crosses and the average time for the BGM EM algorithm is shown with diamonds. The colors and line styles specify the type of noise present for the resulting convergence times, i.e., green solid lines for Student's t-noise, blue dashed lines for Gaussian noise, and red dotted lines for BGM noise. The results indicate that the assumption of the BGM distribution provides a less complex algorithm when fewer samples are available, whatever type of noise is present (see diamonds with lowest computation time). Indeed, the BGM EM algorithm is faster than the Student EM algorithm even when the noise is generated by a Student's t-distribution (see the solid green diamonds lower than the solid green crosses). The Student EM algorithm is only fastest when the number of samples is greater than around 150 and the noise is Gaussian (dashed blue crosses), meaning that the algorithm can be executed faster when the data is nominal. This is an advantage in the long term as the data should most often be nominal.

The combination of these performances implies the BGM EM is the generally preferred algorithm for fast and robust estimation when the number of samples is small. The shorter time until convergence for the BGM EM is likely due to the additional iterative method required in the Student's t EM for estimation of ν. Combining this conclusion with the estimation accuracy of the BGM EM in Figs. 1 to 6 explains that the improved computation speed comes with a trade-off of slightly higher MSE when choosing to assume a BGM instead of a Student's t-distribution. This is true even when the noise is generated by a BGM and the number of samples is small. Hence, it is initially suggested to assume the Student's t-distribution for heavy-tailed noise to maximize robustness with an added cost in computational complexity. Nevertheless, the BGM model still provides a robust estimate compared to the Gaussian MLE so could be considered in applications with fewer samples and a strong restriction on computational power. To further refine this conclusion, the two heavy-tailed models are evaluated in situations with real data in the next section.

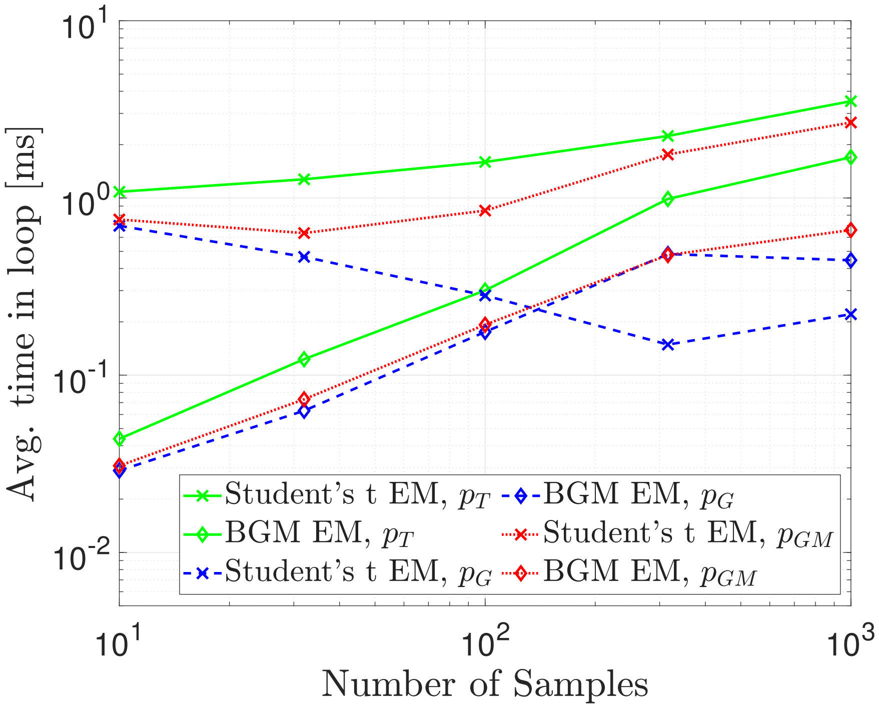

Figure 9: Average time in the EM loop for the algorithms used to estimate the parameters of the two heavy-tailed models. The convergence time is displayed for data with noise generated by each of the heavy-tailed models and the Gaussian model.

5. Real Examples with Contaminated Data

Assuming the Student's t-distribution for data contaminated by outliers is shown in the previous section to perform well even when outliers are modeled by the BGM. However, the assumption of the BGM was also able to achieve competitive estimation accuracy under both heavy-tailed models, while maintaining a faster computation time using the associated EM algorithm. The next step is to see how well these assumptions perform with a set of realistic heavy-tailed data, where observations are not necessarily generated by a specific model but can be assumed to come from a heavy-tailed model.

This section first considers financial data which is known to follow a heavy-tailed model [32]. The linear returns for crypto currencies have heavy-tails due to volatility in the prices, and the estimation of the location of the distribution allows investors to determine the most probable return. The two heavy-tailed models are tested as estimators of location for the daily returns of a collection of crypto currencies. This analysis does not have access to the ground truth for the true location parameter. Additionally, there are no labeled outliers to define an uncontaminated subset. This is because this data is a case of truly heavy-tailed data as was explored in the toy examples, whereas the other applications investigated are more like Gaussian data with outliers.

The next example considers data obtained by simulating the clock bias measurements made within a swarm of satellites as explained in [33]. Realistic anomalies that can occur in the timing measurements due to issues in the clocks themselves or contaminated noise in the inter-satellite links are introduced at known time epochs and on labeled data. This data demonstrates a real application with small sample sizes that can benefit from the robust estimation obtained when assuming a heavy-tailed model.

Finally, real breast cancer data from [34] is used to test the two heavy-tailed models. The obtained data is used for benchmarking anomaly detection methods, allowing a clean Gaussian subset to be defined by removing the labeled outliers. As a result, the "true" location and scale parameters can be defined and we can observe the MSE with respect to these true values to determine which of the heavy-tailed models is more efficient at mitigating the presence of the outliers in the classification of the data.

5.1. Crypto currency returns

Crypto currency price data was obtained from [32], with 930 to 1326 days worth of data, depending on when the corresponding currency came into the market. The daily linear returns were then calculated in percentages, with distribution displayed in Fig. 10. Each of the heavy-tailed models as well as the Gaussian model are fit to the histogram with PDFs defined by the estimated location, scale, and nuisance parameters.

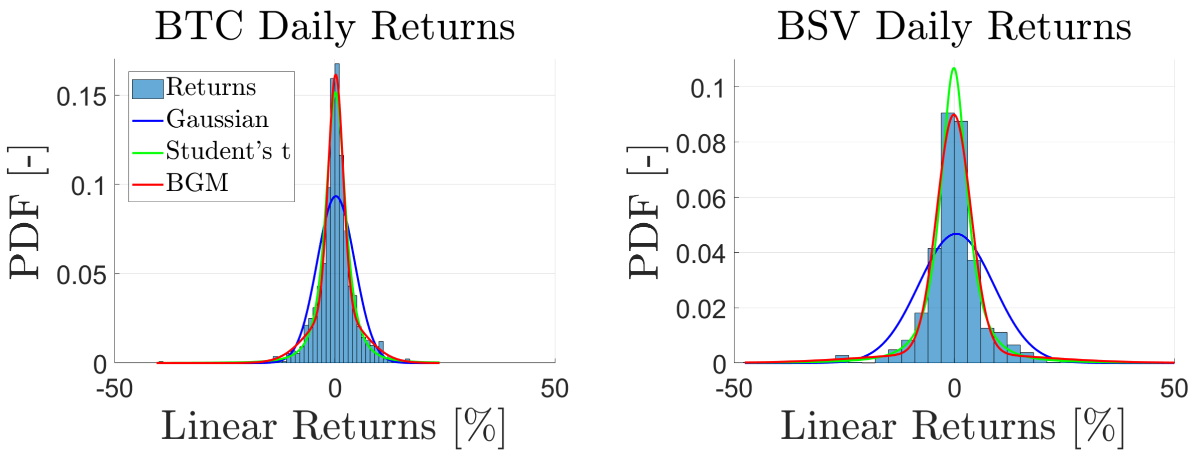

Figure 10: Root mean square error for the location estimator and MSE for the scale estimator in an ensemble of contaminated clock residuals.

The estimation accuracy of the location and scale parameters is not simple to define since there is no ground truth available for the financial data. We compare the MSE between the estimated location parameter and the theoretical mode of the distribution. This is equivalent to identifying the most probable return over one day when investing in the corresponding crypto currency, a practical value that can also be applied over other time intervals with enough data. In the case of symmetric distributions, the median is a good, robust approximation of the mode of the distribution. The daily returns of crypto currency investments are shown to be reasonably symmetric in Fig. 10 and in [32]. The resulting estimation accuracy of the true location parameter is shown in Fig. 11, where the MSE over the history of all data is shown for each crypto currency in the dataset and the evolution of the estimation accuracy for increasing sample sizes is shown by bootstrapping the existing data into smaller and larger datasets. The estimation of the location of the distribution for two different crypto currencies is improved by both heavy-tailed models, but we observe that the Student's t-distribution performs better than the BGM. The Student's t-distribution is therefore the better choice for this data that is known to have heavy-tails but the exact distribution is unknown.

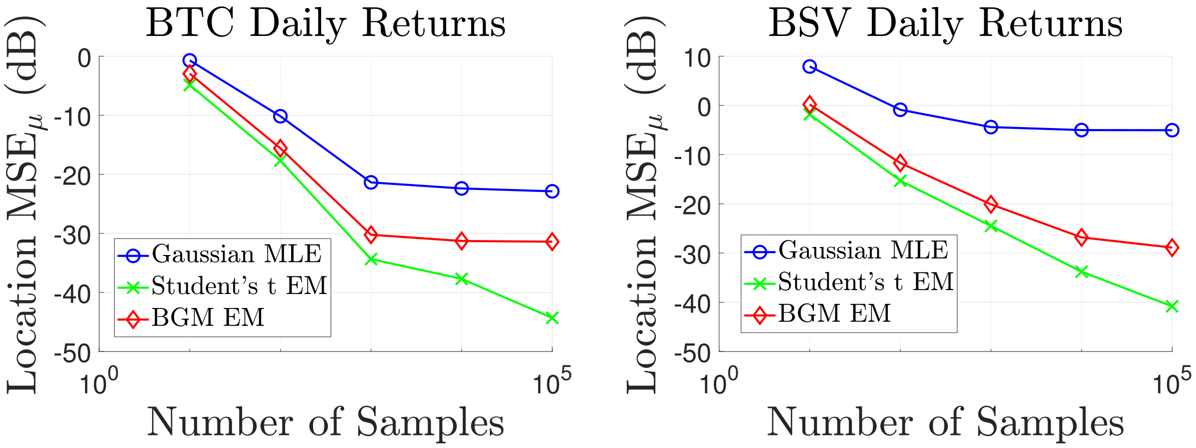

Figure 11: Mean square error for the location estimate for the daily returns of two different crypto currencies.

5.2. *Time scale data*

For an ensemble of distributed sensors that perform radio interferometry, the time of observation of different measurements is important for correctly combining or comparing the data at each sensor [35, 36, 37]. As a result, no individual timing device should be relied upon in the ensemble. Instead, a common reference time or "time scale" is computed using the measured biases between each clock in the ensemble. A common application that makes use of this reference time is navigation using Global Navigation Satellite Systems, where an outlier in one of the clocks can be propagated into the system time and cause a bias in the final positioning solution.

The use of a robust time scale algorithm based on the Student's t-distribution has already been shown to mitigate the expected types of anomalies for satellite clocks, that appear in the form of instantaneous jumps in the clock time [33], which motivated this study. Ensembles of different numbers of Oven Controlled Crystal Oscillators (OCXO) were simulated with a 10 ns jump injected on one of the clock bias measurements at a single time instant. The time scale then corresponds to a weighted mean of the instantaneous prediction errors. The lower the error is between this weighted mean and the true mean, the more stable the resulting time scale will be in presence of the anomaly. In [33], the weights assigned to the prediction errors come from the weights defined in the EM algorithm for estimating the parameters of the Student's t-distribution, defined in (A.16). Equivalently, the weights can also be defined using the EM algorithm that estimates the parameters of the BGM distribution, see (B.25).

Fig. 12 shows the MSE calculated based on the difference between the location and scale estimates with and without the contaminated measurement. The lower the MSE, the less effect the injected outlier has on the system time. As expected, assuming either of the heavy-tailed distributions results in lower MSE than the Gaussian assumption when outliers are present. The location estimate using the Student's t-distribution results in an average reduction of around 20 dB compared to the non-robust Gaussian MLE, and the BGM model reduces the error by a further 4dB on average. This equates to propagated jumps in the time scale being approximately 99% smaller when using the robust estimators, and the BGM based time scale having a jump at least half the size of the Student's t based time scale. Similarly, the average reduction in estimation of the scale parameter compared to the Gaussian MLE is 66 dB and 82 dB for the Student's t and BGM models, respectively. These results refer to the error in the estimated dispersion of the data, which can have negative consequences in defining confidence intervals and integrity measurements for the underlying ensemble of clocks.

The superior performance of the BGM model is actually the opposite of what was found in the previous examples. This suggests the distribution of the clock data with an anomaly is better modeled by the BGM than the Student's t-distribution, at least in the scenarios with a realistic number of clocks. Whereas, in the toy examples and the financial data, the distributions were strictly heavy-tailed with more significant levels of contamination. The benefits of assuming the Student's t-distribution may be more evident if a larger portion of the data is contaminated or the intensity of the contamination becomes stronger. Further analysis should be conducted with a range of anomaly magnitudes and proportions that are relevant for the application in question before concluding which heavy-tailed model works better.

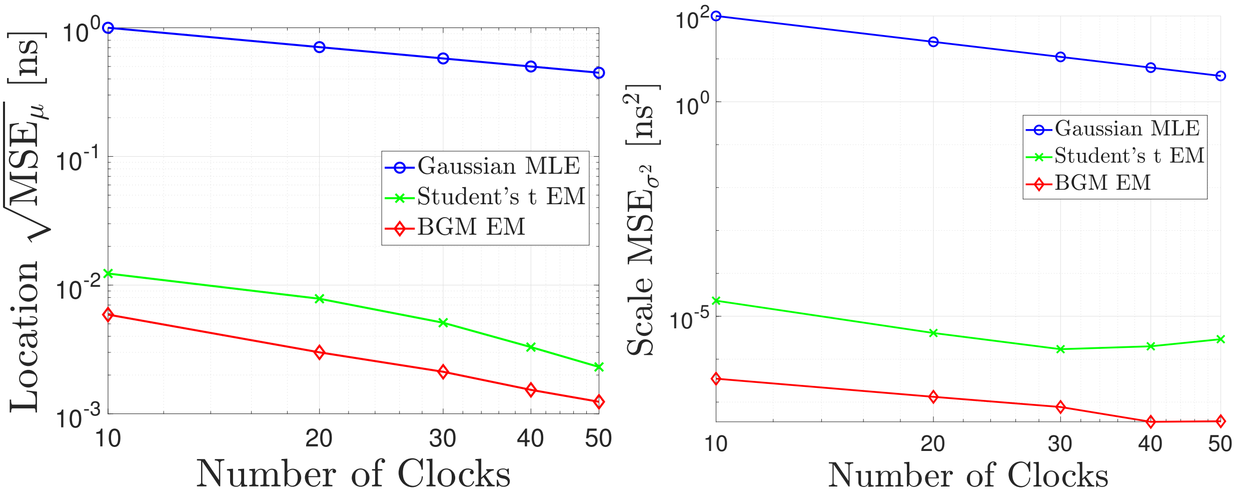

Figure 12: Root mean square error for the location estimator and MSE for the scale estimator in an ensemble of contaminated clock residuals.

5.3. Breast Cancer Measurements

The breast cancer data available from [34] includes a range of different features that together contribute to the classification of each sample being normal or anomalous. The meaning of each feature is lost in the stored database but the context of breast cancer implies a reliance more on detection than mitigation of anomalies. Robust estimation of the location and scale of each feature could then provide information on the average healthy patient and the deviation from that average.

We have observed that the distributions of certain features in the breast cancer dataset are approximately Gaussian after the outliers are removed (with 99% confidence according to the Lilliefors test). The nine features that passed the Gaussianity test are now considered as independent experiments of real data with contaminated measurements. The number of observations in the breast cancer data is $N_s = 367$ per feature. Ten outlying samples were labeled in the data ($N_o = 10$), corresponding to a contamination fraction of $\varepsilon = N_o/N_s = 0.0272$. With the outliers labeled, they can be removed and the effective "true" location and scale parameters can be determined. The error under each assumed model is then tested by estimating the location and scale using the contaminated data $(\hat{\mu}, \hat{\sigma}^2)$ and comparing the errors with respect to the true parameters computed with the clean data (μ_0, σ_0^2).

The effective number of degrees of freedom and the scaling factor of the contaminating variance can only be obtained through estimation and are not necessarily true parameters describing a true distribution. Since the derived bounds depend on those parameters, the estimation performance is not constrained by the derived bounds but instead by another limit based on the true distribution of the real data, which is unknown. However, the presence of outliers allows us to classify the distribution as heavy-tailed so the assumption of the Student's t-distribution and the BGM are expected to provide better estimation performance than the Gaussian assumption. The number of samples is synthetically increased to $N_s = 1000$ by bootstrapping the data such that samples are selected randomly from the dataset for a fixed number of Monte Carlo iterations for each feature. This bootstrapping was conducted to increase the sample size and ensure that we are in the asymptotic regime for the comparison of the estimation performance for each feature.

Fig. 13 shows the MSE between the estimated location and scale parameters and the corresponding location and scale of the data with anomalies removed. The data was also normalized before estimating the parameters to plot the error for each independent feature on the same figure. As expected, assuming the heavy-tailed distributions results in lower MSE than the Gaussian assumption when outliers are present. For every feature in the breast cancer data, and for all ensemble sizes, the MSE achieved by the BGM is lower than or the same as the estimator assuming the Student's t-distribution. This means the realistic cases of outliers appearing are better modeled by the BGM while being different enough from the Student's t-distribution that it could not provide as good of an approximation.

5.4. Real data discussion

For the heavy-tailed financial data, the volatility causing heavier tails is a characteristic of the data. As a result, the estimation performance of the Student's t-distribution is shown to be better than the BGM, just as we predicted in the toy examples that used data truly

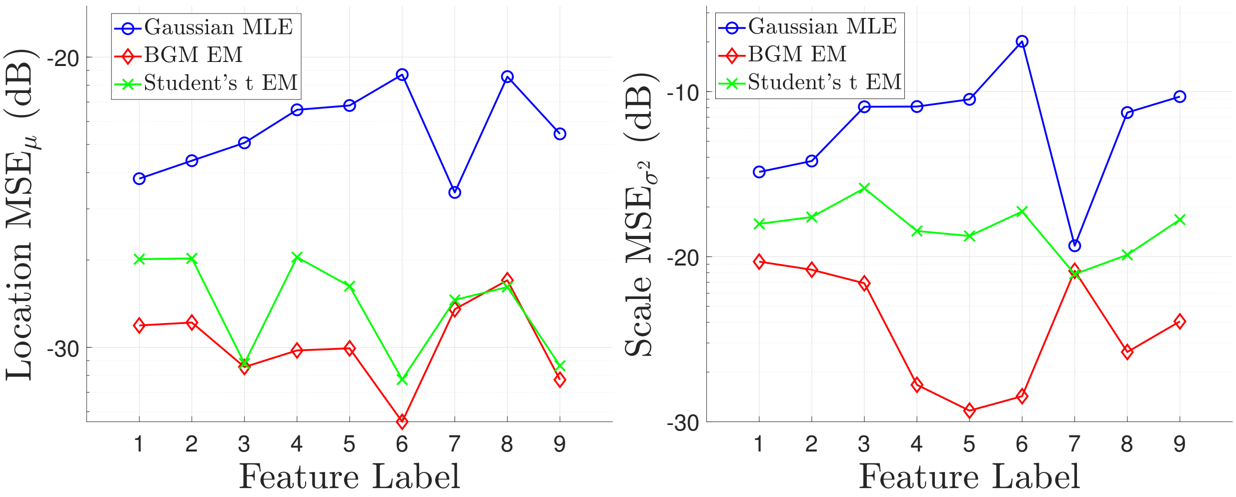

Figure 13: Mean Square Error for the estimate of the location and scale parameters for each of the different features of the breast cancer dataset [34].

generated by the two heavy-tailed models. This puts the results of the other datasets into perspective.

Contrary to the initial recommendation of a Student's t-distribution using toy datasets and the heavy-tailed financial data, the clock and breast cancer datasets combined have revealed that the BGM outshines the Student's t-distribution as an assumed heavy-tailed model when the data is Gaussian plus outliers. Since the other datasets are simply Gaussian data with anomalies occurring that cause outliers, the BGM ends up being a better model than the Student's t-distribution for estimation of the location and scale. These results make sense intuitively as the BGM models exactly a proportion of Gaussian data plus a proportion of outlying values, effectively the same as the other models although not exactly generated by a BGM.

It is important to disclaim that the above examples all consider univariate measurements that are assumed to be independent and identically distributed. Although the real data in this section provides justifications of the practicality of univariate, uncorrelated, and symmetric distributions, there are other applications that do not precisely fit these assumptions. Future work is necessary to expand the derivations and test other estimators that take into account more dimensions in the data, possible correlation with the contaminating data, and the skewness of the distributions. Each of these simplifying assumptions could potentially be the source of a another misspecified estimator that is constrained by some other MCRB.

6. Conclusion

Heavy-tailed statistical distributions are useful to model contaminated noise with higher than usual occurrence of outliers. The derivation of a new Misspecified Cramér-Rao bound has confirmed the advantage of correctly specifying the estimator over assuming a Gaussian distribution when the data is produced by a bi-variance Gaussian mixture. This new result coincides with the existing results for the Student's t-distribution and elliptically symmetric distributions in general.

The Cramér-Rao Bound for joint estimation of the location, scale, and shape of the Student's t-distribution was also derived in this work, confirming the gain in performance due to estimating the number of degrees of freedom is almost negligible compared to assuming it is known. Indeed, a maximum reduction of the CRB by 4.2% was demonstrated. The maximum gain achieved by including the joint estimation of the number of degrees of freedom is also shown to occur when we work with univariate data. Nevertheless, the examples with univariate data are still not seriously affected in the domains where the nuisance parameters are defined. This work also revealed that the Cramér-Rao bound derived for the estimation of the parameters of the Student's t-distribution is approximately valid for the correctly specified maximum likelihood estimator of the parameters of the bi-variance Gaussian mixture model with equivalent nuisance parameters.

The maximum likelihood estimator under a misspecified heavy-tailed model achieves comparable asymptotic performance as the correctly specified maximum likelihood estimator when estimating the location parameter, but not when estimating the scale parameter. This result coincides with the new derivation of the pseudotrue parameters for the misspecified heavy-tailed models. The expectation maximization algorithms used to estimate the parameters of each heavy-tailed model are not optimal because they have a bias that should be accounted for in a corresponding bound. Nevertheless, this bias is still an improvement over the Gaussian assumption for realistic sample sizes even when assuming the incorrect heavy-tailed distribution. The expectation maximization algorithm assuming a bi-variance Gaussian mixture provides an estimator with lower convergence time when compared to assuming a Student's t-distribution but with reduced accuracy when the data actually follows a t-distribution, introducing the trade-off between models.

Finally, both models were confirmed to be robust in the case of real datasets with naturally occurring outliers. The estimator based on the bi-variance Gaussian mixture achieved the best mean square error when the type of data was rather Gaussian plus outliers and the estimator based on the Student's t-distribution performed better when the data was really heavy-tailed due to the nature of the data. With these results, there cannot be a generally preferred heavy-tailed model but a simple recommendation can be made. When you expect heavy-tails in the data, assuming a Student's t-distribution is preferred over the BGM. When you do not expect heavy-tails in the data it is better to use a BGM as an adaptive robustness to outliers randomly occurring in the tails.

Future work should consist of exploring other realistic cases where anomalies are present and cause the tails of the distribution of the data to be heavier, potentially introducing skewness and correlation. These other applications will likely also benefit from multivariate models, so the corresponding derivations should be appropriately developed to apply to these cases.

Appendix A. Expectation Maximization - Student's t-distribution

An MLE is defined based on knowledge of a likelihood function that fits the observations being made. In this work, the Student's t-distribution is used to model the distribution of data contaminated with anomalies. The likelihood of N i.i.d. samples from the univariate Student's t-distribution is defined as follows [8]:

$$\begin{aligned} L(\boldsymbol{z};\mu,\sigma^2,\nu) &= \prod_{i=1}^{N} p(z_i;\mu,\sigma^2,\nu) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{\nu\pi\sigma^2}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left[1+\frac{1}{\nu}\left(\frac{z_i-\mu}{\sigma}\right)^2\right]^{-\frac{(\nu+1)}{2}}, \end{aligned} \tag{A.1}$$

with $\boldsymbol{z} = (z_1, ..., z_N)^T$. The parameters μ, σ^2, and ν are the mean, scale, and shape parameters, respectively. The shape parameter is also referred to as the number of degrees of freedom and is related to the quantity of outliers in the data. A random variable z_i that follows the Student's t-distribution is denoted as $z_i \sim T(\mu,\sigma^2,\nu)$. The MLE for each of the parameters μ, σ^2, and ν aims to identify the values of those parameters that maximize the likelihood (A.1) for a given sample $\boldsymbol{z}$:

$$\left[\hat{\mu},\hat{\sigma}^2,\hat{\nu}\right]^T = \arg\max_{\mu,\sigma^2,\nu} \left\{L(\boldsymbol{z};\mu,\sigma^2,\nu)\right\}. \tag{A.2}$$

To simplify the derivations, it is usual to derive the expression of the MLE by minimizing the negative log-likelihood [30]

$$\left[\hat{\mu},\hat{\sigma}^2,\hat{\nu}\right]^T = \arg\min_{\mu,\sigma^2,\nu} \left\{-\log L(\boldsymbol{z};\mu,\sigma^2,\nu)\right\}. \tag{A.3}$$

In the case of the univariate Student's t-distribution, the MLEs of the unknown parameters are the solutions to the following equations

$$\frac{\partial \log(L)}{\partial \mu} = \frac{\nu+1}{\sigma^2} \sum_{i=1}^{N} \frac{z_i-\mu}{\nu+\left(\frac{z_i-\mu}{\sigma}\right)^2} = 0, \tag{A.4}$$

$$\frac{\partial \log(L)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{\nu+1}{2\sigma^4} \sum_{i=1}^{N} \frac{(z_i-\mu)^2}{\nu+\left(\frac{z_i-\mu}{\sigma}\right)^2} = 0, \tag{A.5}$$

$$\frac{\partial \log(L)}{\partial \nu} = \frac{N}{2}\phi\left(\frac{\nu+1}{2}\right) - \frac{N}{2}\phi\left(\frac{\nu}{2}\right) - \frac{1}{2}\sum_{i=1}^{N}\left[\frac{\nu+1}{\nu+\left(\frac{z_i-\mu}{\sigma}\right)^2} - \log\left(\frac{\nu+1}{\nu+\left(\frac{z_i-\mu}{\sigma}\right)^2}\right) - 1\right] = 0, \tag{A.6}$$

where $\phi(x) = \psi(x) - \log(x), \quad x > 0$, and the digamma function $\psi(x)$ is:

$$\psi(x) = \frac{d}{dx}\log\left[\Gamma(x)\right] = \frac{\Gamma^{'}(x)}{\Gamma(x)}. \tag{A.7}$$

As each parameter depends on the other two, the MLEs of μ, σ^2, ν cannot be computed directly. However, it is well known that the Student distribution can be represented by an infinite mixture of Gaussian distributions:

$$z_i | v_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{v_i}\right), \; v_i \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{2}{\nu}\right), \tag{A.8}$$

where $\mathcal{G}(a, b)$ denotes a Gamma distribution with shape and scale parameters a and b, respectively. For the random variable v_i that follows this distribution, the marginal PDF is:

$$f(v_i) = \frac{1}{\Gamma(a) b^a} v_i^{a-1} e^{-v_i/b}, \quad v_i > 0. \tag{A.9}$$

The joint PDF of $\boldsymbol{z} = [z_1, \cdots, z_N]^T$ and $\mathbf{v} = [v_1, \cdots, v_N]^T$, referred to as the complete likelihood, is expressed as:

$$L_c(\boldsymbol{z}, \mathbf{v}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \frac{\left(\frac{\nu}{2}\right)^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)} v_i^{\frac{\nu+1}{2}-1} \exp\left(-\frac{v_i}{2}\left(\nu + \left(\frac{z_i - \mu}{\sigma}\right)^2\right)\right). \tag{A.10}$$

Marginalizing the complete likelihood with respect to $\mathbf{v}$ yields the likelihood as described in (A.1). This representation allows an EM algorithm to be derived [19]. The EM algorithm is based on the so-called complete log-likelihood, which is the logarithm of the joint distribution of $(\boldsymbol{z}, \mathbf{v})$:

$$l_c(\boldsymbol{z}, \mathbf{v}) = \frac{N\nu}{2} \log\left(\frac{N\nu}{2}\right) - N \log\Gamma\left(\frac{\nu}{2}\right) + \left(\frac{\nu+1}{2} - 1\right) \sum_{i=1}^{N} \log v_i - \frac{N}{2} \log(2\pi) \tag{A.11}$$

$$-\frac{N}{2} \log\left(\sigma^2\right) - \frac{1}{2} \sum_{i=1}^{N} v_i \left[\nu + \frac{(z_i - \mu)^2}{\sigma^2}\right]. \tag{A.12}$$

After an initialization of the unknown parameters, the EM alternates between Expectation (E) and Maximization (M) steps:

- **Initialization**: The location and scale parameters are initialized with the Gaussian MLEs and the number of degrees of freedom is chosen to be small because that will help minimize the number of iterations in case there is an anomaly:

$$\hat{\mu}_0 = \frac{1}{N} \sum_{i=1}^{N} z_i, \tag{A.13}$$

$$\hat{\sigma}_0^2 = \frac{1}{N-1} \sum_{i=1}^{N} (z_i - \hat{\mu}_0)^2, \tag{A.14}$$

$$\hat{\nu}_0 = 3. \tag{A.15}$$

- **E Step**: At iteration k, given $\hat{\boldsymbol{\eta}}_{k-1} = \left(\hat{\mu}_{k-1}, \hat{\sigma}^2_{k-1}, \hat{\nu}_{k-1}\right)^T$, the E step computes the expectation of $l_c(\boldsymbol{z}, \mathbf{v})$ with respect to the variables v_i, which requires the following computations

$$u_{i,k} = E[v_i|z_i, \hat{\eta}_{k-1}] = \frac{\hat{\nu}_{k-1}+1}{\hat{\nu}_{k-1} + \frac{(z_i-\hat{\mu}_{k-1})^2}{\hat{\sigma}^2_{k-1}}}, \tag{A.16}$$

$$w_{i,k} = E[\log(v_i)|z_i, \hat{\eta}_{k-1}] \tag{A.17}$$

$$= \psi\left(\frac{\hat{\nu}_{k-1}+1}{2}\right) - \log\left(\frac{1}{2}\left(\hat{\nu}_{k-1} + \frac{(z_i-\hat{\mu}_{k-1})^2}{\hat{\sigma}^2_{k-1}}\right)\right), \tag{A.18}$$

and leads to the objective function Q

$$Q(\boldsymbol{\eta}; \hat{\boldsymbol{\eta}}_k) = \frac{N\nu}{2}\log\left(\frac{N\nu}{2}\right) - N\log\Gamma\left(\frac{\nu}{2}\right) + \left(\frac{\nu+1}{2} - 1\right)\sum_{i=1}^{N} w_{i,k}$$
$$- \frac{N}{2}\log(2\pi) - \frac{N}{2}\log\left(\sigma^2\right) - \frac{1}{2}\sum_{i=1}^{N} u_{i,k}\left[\nu + \frac{(z_i-\mu)^2}{\sigma^2}\right]. \tag{A.19}$$

- **M Step**: At iteration k, the M Step maximizes the Q function with respect to the parameters μ, σ^2, ν, which yields

$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} u_{i,k} z_i}{\sum_{i=1}^{N} u_{i,k}}, \tag{A.20}$$

$$\hat{\sigma}^2_k = \frac{\sum_{i=1}^{N} u_{i,k}(z_i - \hat{\mu}_k)^2}{N}, \tag{A.21}$$

$$N\phi\left(\frac{\hat{\nu}_k}{2}\right) + \sum_{i=1}^{N}\left[u_{i,k} - w_{i,k} - 1\right] = 0. \tag{A.22}$$

The terms $u_{i,k}$ act as weights in the estimate of the location and scale parameters. These weights take small values for samples that are far from the estimated mean and are approximately equal if the degrees of freedom are large, reducing the estimator to the classical sample mean and variance or MLEs for the parameters of a Gaussian distribution. By assigning lower values to outliers the weighted average is able to mitigate the effects of anomalies.

The solution to (A.22) is obtained through the application of Newton's method, which converges to a solution after few iterations. Consider the solution to (A.22) is the root of the function $f(x)$, then Newton's method allows us to iteratively approximate that root by the following

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \tag{A.23}$$

where the initial guess can be the previous estimate of the number of degrees of freedom, $x_0 = \hat{\nu}_{k-1}(t)$. The above is repeated until reaching a maximum number of iterations or a minimum difference between consecutive estimates is reached. Once the error between consecutive estimates has reached the chosen minimum, the estimators have converged.

Appendix B. EM - Bimodal Gaussian Mixture

Assuming the Bimodal Gaussian Mixture distribution is used to model contaminated data with a proportion of anomalous data ε and variance scaling factor k, the likelihood function is:

$$\begin{aligned} L(\boldsymbol{z};\boldsymbol{\beta}) &= \prod_{i=1}^{N} p(z_i;\boldsymbol{\epsilon}) \\ &= \prod_{i=1}^{N} \frac{1-\varepsilon}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{z_i-\mu}{\sigma}\right)^2\right) + \frac{\varepsilon}{\sqrt{2\pi\alpha\sigma^2}} \exp\left(-\frac{1}{2\alpha}\left(\frac{z_i-\mu}{\sigma}\right)^2\right), \end{aligned} \tag{B.1}$$

with $\boldsymbol{z} = (z_1, ..., z_N)^T$ and $\boldsymbol{\beta} = [\mu, \sigma^2, \varepsilon, \alpha]^T$. To simplify the expression of the PDF and furthermore simplify the following derivations, the PDF is rewritten such that the exponentials are presented by the gaussian PDF $g(z_i; \mu, \sigma^2)$ evaluated with different parameter vectors $\boldsymbol{\theta_0} = [\mu, \sigma^2]^T$ and $\boldsymbol{\theta_1} = [\mu, \alpha\sigma^2]^T$

$$L(\boldsymbol{z};\boldsymbol{\beta}) = \prod_{i=1}^{N} (1-\varepsilon) g(z_i;\boldsymbol{\theta}_0) + \varepsilon g(z_i;\boldsymbol{\theta}_1), \tag{B.2}$$

The MLE for the parameters defining the BGM are the respective values that maximize the likelihood (A.1) for a given sample $\boldsymbol{z}$:

$$\hat{\boldsymbol{\beta}} = \left[\hat{\mu}, \hat{\sigma}^2, \hat{\varepsilon}, \hat{\alpha}\right]^T = \arg\max_{\boldsymbol{\beta}} \{L(\boldsymbol{z};\boldsymbol{\beta})\}. \tag{B.3}$$

The log-likelihood is maximized instead of the likelihood function, which still results in the appropriate MLE for each parameter

$$\log\left(L(\boldsymbol{z};\boldsymbol{\beta})\right) = \sum_{i=1}^{N} \log\left[(1-\varepsilon) g(z_i;\boldsymbol{\theta}_0) + \varepsilon g(z_i;\boldsymbol{\theta}_1)\right]. \tag{B.4}$$

In the case of the univariate BGM, the MLE of the location parameter is the solution to the following equation

$$\frac{\partial \log\left(L(\boldsymbol{z};\boldsymbol{\beta})\right)}{\partial \mu} = \sum_{i=1}^{N} \frac{1}{L(z_i;\boldsymbol{\beta})} \left[(1-\varepsilon)\frac{\partial g(z_i;\boldsymbol{\theta}_0)}{\partial \mu} + \varepsilon \frac{\partial g(z_i;\boldsymbol{\theta}_1)}{\partial \mu}\right] = 0, \tag{B.5}$$

with the derivatives of the sub-functions

$$\frac{\partial g(z_i;\boldsymbol{\theta}_0)}{\partial \mu} = \left(\frac{z_i-\mu}{\sigma^2}\right) g(z_i;\boldsymbol{\theta}_0), \tag{B.6}$$

$$\frac{\partial g(z_i;\boldsymbol{\theta}_1)}{\partial \mu} = \left(\frac{z_i-\mu}{\sigma^2}\right) \frac{g(z_i;\boldsymbol{\theta}_1)}{\alpha}, \tag{B.7}$$

so the equation for the MLE of the mean becomes

$$\frac{\partial \log\left(L(\boldsymbol{z};\boldsymbol{\beta})\right)}{\partial \mu} = \sum_{i=1}^{N} \frac{\left(\frac{z_i-\mu}{\sigma^2}\right)}{L(z_i;\boldsymbol{\beta})} \left[(1-\varepsilon)g(z_i;\boldsymbol{\theta}_0) + \frac{\varepsilon}{\alpha}g(z_i;\boldsymbol{\theta}_1)\right] = 0. \tag{B.8}$$

The above expression does not have a closed form expression for μ that does not depend on the other parameters. It can similarly be shown that the estimators for the other parameters will also depend on the mean. Due to this interdependence of the estimators, the method of Expectation Maximization (EM) is required to iteratively estimate each parameter until converging to the MLE.

To define the EM algorithm, latent variables must be introduced that allow a complete likelihood function to be defined. Since there are two modes of the BGM, one linked to normal data and the other linked to anomalous data, a latent variable γ_i is introduced as an outlier label for each sample so can take two possible values with known probabilities. The latent variable is then a Bernoulli random variable

$$\gamma_i \sim \mathcal{B}(1,\varepsilon), \tag{B.9}$$

where the two possible labels are $\gamma_i = 0$ for normal data and $\gamma_i = 1$ for anomalous data. The probability that either label is assigned to the latent variable is linked to the contamination proportion $L(\gamma_i = 0) = 1-\varepsilon$ and $L(\gamma_i = 1) = \varepsilon$. The conditional probability for the samples z_i given a certain label is linked to the sub-functions expressed above

$$L(z_i|\gamma_i = l) = g(z_i;\boldsymbol{\theta}_l), \tag{B.10}$$

i.e., the samples labeled as normal data will effectively be samples of the nominal Gaussian mode $g(z_i;\boldsymbol{\theta}_0)$ and the anomalies come from the contaminating Gaussian mode $g(z_i;\boldsymbol{\theta}_1)$. The joint PDF of $\boldsymbol{z} = [z_1,\cdots,z_N]^T$ and $\gamma = [\gamma_1,\cdots,\gamma_N]^T$, referred to as complete likelihood, is expressed as:

$$L_c(\boldsymbol{z},\boldsymbol{\gamma}) = \varepsilon^{N_a}(1-\varepsilon)^{N-N_a}\prod_{i=1}^{N} g(z_i,\boldsymbol{\theta}_0)^{1-\gamma_i} g(z_i,\boldsymbol{\theta}_1)^{\gamma_i}, \tag{B.11}$$

where $N_a = \sum_{i=1}^{N}\gamma_i$ is the number of samples that have been labeled as anomalous. The EM algorithm is based on the so-called complete log-likelihood, which is the logarithm of the joint distribution of $(\boldsymbol{z},\boldsymbol{\gamma})$:

$$l_c(\boldsymbol{z},\boldsymbol{\gamma}) = N_a\log(\varepsilon) + (N-N_a)\log(1-\varepsilon) \tag{B.12}$$

$$+\sum_{i=1}^{N}(1-\gamma_i)\log\left(g(z_i,\boldsymbol{\theta}_0)\right) + \sum_{i=1}^{N}\gamma_i\log\left(g(z_i,\boldsymbol{\theta}_1)\right). \tag{B.13}$$

After an initialization of the unknown parameters, the EM alternates between Expectation (E) and Maximization (M) steps:

- **Initialization**: The location and scale parameters are initialized with the Gaussian MLEs and the number of degrees of freedom is chosen to be small because that will help minimize the number of iterations in case there is an anomaly:

$$\hat{\mu}_0 = \frac{1}{N}\sum_{i=1}^{N} z_i, \tag{B.14}$$

$$\hat{\sigma}_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \hat{\mu}_0)^2, \tag{B.15}$$

$$\hat{\varepsilon}_0 = 0.01, \tag{B.16}$$

$$\hat{\alpha}_0 = 1.5. \tag{B.17}$$

- **E Step**: At iteration j, given $\hat{\boldsymbol{\beta}}_{j-1} = \left(\hat{\mu}_{j-1}, \hat{\sigma}_{j-1}^2, \hat{\varepsilon}_{k-1}, \hat{\alpha}_{j-1}\right)^T$, the E step computes the expectation of $l_c(\boldsymbol{z}, \boldsymbol{\gamma})$ with respect to the variables γ_i, and leads to the objective function Q

$$\begin{aligned} Q(\boldsymbol{\epsilon}; \hat{\boldsymbol{\beta}}_j) = E[l_c(\boldsymbol{z}, \boldsymbol{\gamma})|\boldsymbol{z}, \hat{\boldsymbol{\beta}}_{j-1}] = \log(\varepsilon)\sum_{i=1}^{N} E\left[\gamma_i|\boldsymbol{z}, \hat{\boldsymbol{\beta}}_{j-1}\right] + N\log(1-\varepsilon) \\ - \log(1-\varepsilon)\sum_{i=1}^{N} E\left[\gamma_i|\boldsymbol{z}, \hat{\boldsymbol{\beta}}_{j-1}\right] + \sum_{i=1}^{N}\log\left(g(z_i; \boldsymbol{\theta}_0)\right) \\ + \sum_{i=1}^{N} E\left[\gamma_i|\boldsymbol{z}, \hat{\boldsymbol{\beta}}_{j-1}\right]\left(\log\left(\frac{g(z_i; \boldsymbol{\theta}_1)}{g(z_i; \boldsymbol{\theta}_0)}\right)\right) \end{aligned} \tag{B.18}$$

The objective function requires some estimation of the latent variable in the form of it's expected value given the measurements z_i and previous estimates of the parameters of interest $\hat{\boldsymbol{\epsilon}}_{j-1}$. For a Bernoulli random variable it is known that the expectation is equivalent to the probability of of success, in this case, the likelihood of an outlier

$$\begin{aligned} u_{i,j} &= E[\gamma_i|z_i, \hat{\boldsymbol{\beta}}_{j-1}], \\ &= L(\gamma_i = 1|z_i, \hat{\boldsymbol{\beta}}_{j-1}), \\ &= \frac{L(\gamma_i = 1)L(z_i, \hat{\boldsymbol{\beta}}_{j-1}|\gamma_i = 1)}{L(z_i; \hat{\boldsymbol{\beta}}_{j-1})}, \\ &= \frac{\hat{\varepsilon}_{j-1} g(z_i; \hat{\boldsymbol{\theta}}_{1,j-1})}{(1-\hat{\varepsilon}_{j-1})g(z_i; \hat{\boldsymbol{\theta}}_{0,j-1}) + \hat{\varepsilon}_{j-1} g(z_i; \hat{\boldsymbol{\theta}}_{1,j-1})}, \end{aligned} \tag{B.19}$$

where the estimated reduced parameter vectors are $\hat{\boldsymbol{\theta}}_{0,j-1} = [\hat{\mu}_{j-1}, \hat{\sigma}_{j-1}^2]^T$ and $\hat{\boldsymbol{\theta}}_{1,j-1} = [\hat{\mu}_{j-1}, \hat{\sigma}_{j-1}^2, \hat{\alpha}_{j-1}]^T$. The term computed above provides a sort of normalized probability of an outlier occurring, i.e., $u_{i,j}$ is large if the likelihood of sample i being an outlier is close to the total likelihood of that observation according to the BGM. The objective function is updated to include the expectation of the latent variable and written in terms of the parameters to be estimated

$$\begin{aligned} Q(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}_j) &= \log(\varepsilon) \sum_{i=1}^{N} u_{i,j} + N \log(1-\varepsilon) - \log(1-\varepsilon) \sum_{i=1}^{N} u_{i,j} \\ &- \frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_{i=1}^{N} \frac{(z_i - \mu)^2}{\sigma^2} - \frac{\log(k)}{2} \sum_{i=1}^{N} u_{i,j} \\ &+ \frac{1}{2} \sum_{i=1}^{N} u_{i,j} \frac{(z_i - \mu)^2}{\sigma^2} - \frac{1}{2\alpha} \sum_{i=1}^{N} u_{i,j} \frac{(z_i - \mu)^2}{\sigma^2} \end{aligned} \tag{B.20}$$

- **M Step**: At iteration k, the M Step maximizes the Q function with respect to the parameters μ, σ^2, ε, α, which yields

$$\hat{\mu}_j = \frac{\sum_{i=1}^{N} w_{i,j} z_i}{\sum_{i=1}^{N} w_{i,j}}, \tag{B.21}$$

$$\hat{\sigma}_j^2 = \frac{1}{N} \sum_{i=1}^{N} w_{i,j} (z_i - \hat{\mu}_k)^2, \tag{B.22}$$

$$\hat{\varepsilon}_j = \frac{\sum_{i=1}^{N} u_{i,j}}{N}, \tag{B.23}$$

$$\hat{\alpha}_j = \frac{\sum_{i=1}^{N} u_{i,j} (z_i - \hat{\mu}_j)^2}{\hat{\sigma}_j^2 \sum_{i=1}^{N} u_{i,j}}. \tag{B.24}$$

The weights used to estimate the location and scale parameters are

$$w_{i,j} = 1 - u_{i,j} + \frac{u_{i,j}}{\hat{\alpha}_{j-1}}, \tag{B.25}$$

which can be understood as the normalized probability that a sample is nominal $(1 - u_{i,j})$ summed with the normalized probability of the sample being an anomaly divided by the scaling factor of the anomalous variance $(\frac{u_{i,j}}{\alpha})$. For anomalous data, $u_{i,j}$ is large as well as α. Therefore the weights $w_{i,j}$ decrease proportionally with the scaling factor of the contaminating variance and are small for anomalous samples. This results in a robust estimate of the mean and nominal variance of the BGM.

The iterative EM steps are repeated until the error between consecutive estimates reduces below a chosen threshold or a maximum number of iterations has occurred. By appropriately initializing the estimates, the number of iterations should remain low. Nevertheless the presence of anomalies complicates the difference between the initial estimates and true parameters so computational limitations should be considered when applying EM algorithms defined for different distributions.

Appendix C. Derivation of pseudo-true parameters.

The pseudo-true parameters $\tilde{\boldsymbol{\theta}} = [\tilde{\mu}, \tilde{\sigma}^2]^T$ are the parameters of the assumed distribution that minimize the KLD from the true distribution.

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{D_{\mathrm{KL}}\} = \arg\min_{\boldsymbol{\theta}} \left\{ E_p \left[\log \left(\frac{p(\boldsymbol{z}; \bar{\boldsymbol{\eta}})}{q(\boldsymbol{z}; \boldsymbol{\theta})} \right) \right] \right\}, \tag{C.1}$$

where $p(z; \bar{\boldsymbol{\eta}})$ represents the true distribution of the data with parameter vector of true values for a given realization of contaminated noise denoted as $\bar{\boldsymbol{\eta}}$. The bars are included to differentiate between the fixed true values and the variable estimates of these true values. The true PDF can be replaced by either $p_T(z; \bar{\mu}_T, \bar{\sigma}_T^2, \bar{\nu})$ or $p_{\mathrm{GM}}(z; \bar{\mu}_{GM}, \bar{\sigma}_{GM}^2, \bar{\alpha}, \bar{\varepsilon})$ for the two investigated distributions. The choice of the true distribution is based on how anomalies can appear in the observations. The cost function for finding the pseudo-true parameters is simplified to only include the parameters of the Gaussian distribution:

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{ -E_p [\log (q(\boldsymbol{z}; \boldsymbol{\theta}))] \}. \tag{C.2}$$

Substituting the log-likelihood function for the Gaussian distribution leads to:

$$-E_p [\log (q(\boldsymbol{z}; \boldsymbol{\theta}))] = \frac{N}{2} E_p \left[\log \left(2\pi\sigma_{\mathrm{G}}^2 \right) \right] \tag{C.3}$$

$$+ \frac{1}{2} \sum_{i=1}^{N} E_p \left[\left(\frac{z_i - \mu_{\mathrm{G}}}{\sigma_{\mathrm{G}}} \right)^2 \right]. \tag{C.4}$$

We then expand the quadratic function of z to separate the misspecified parameters μ_{G} and σ_{G}^2 from the expectation

$$-E_p [\log (q(\boldsymbol{z}; \boldsymbol{\theta}))] = \frac{N}{2} \log \left(2\pi\sigma_{\mathrm{G}}^2 \right)$$
$$+ \frac{1}{2\sigma_{\mathrm{G}}^2} \left(\sum_{i=1}^{N} E_p \left[z_i^2 \right] - 2\mu_{\mathrm{G}} \sum_{i=1}^{N} E_p[z_i] + N\mu_{\mathrm{G}}^2 \right). \tag{C.5}$$

The mean and variance of a random variable that follows the Student's t-distribution are known to be:

$$E_{p_T}[z] = \bar{\mu}_T, \mathrm{var}(z) = E_{p_T}[z_i^2] - E_{p_T}[z_i]^2 = \bar{\sigma}_T^2 \frac{\bar{\nu}}{\bar{\nu} - 2}.$$

Similarly, the moments of the BGM model are known. The expectation of the BGM can be written as a linear combination of the expectations with respect to each of the modes. Using the BGM PDF

$$p_{\mathrm{GM}}(z; \bar{\mu}_{GM}, \bar{\sigma}_{GM}^2, \bar{\alpha}, \bar{\varepsilon}) = (1 - \bar{\varepsilon}) g(z; \bar{\mu}_{GM}, \bar{\sigma}_{GM}^2) + \varepsilon g(z; \bar{\mu}_{GM}, \bar{\alpha}\bar{\sigma}_{GM}^2), \tag{C.6}$$

where $g(z; \mu, \sigma^2)$ is the PDF of the Gaussian distributed data z with mean μ and variance σ^2, the expected values with respect to the two possible Gaussian PDFs $g_0 = g(z_i; \bar{\mu}_{GM}, \bar{\sigma}_{GM}^2)$

and $g_1 = g(z_i; \bar{\mu}_{GM}, \bar{\alpha}\bar{\sigma}^2_{GM})$ for the nominal and contaminated data, respectively are evaluated as follows:

$$E_{p_{\mathrm{GM}}}\left[z_i\right] = (1-\bar{\varepsilon})E_{g_0}\left[z_i\right] + \bar{\varepsilon}E_{g_1}\left[z_i\right], \tag{C.7}$$

$$E_{p_{\mathrm{GM}}}\left[z_i\right] = \bar{\mu}_{GM}, \tag{C.8}$$

$$E_{p_{\mathrm{GM}}}\left[z_i^2\right] = (1-\bar{\varepsilon})E_{g_0}\left[z_i^2\right] + \bar{\varepsilon}E_{g_1}\left[z_i^2\right], \tag{C.9}$$

$$E_{p_{\mathrm{GM}}}\left[z_i^2\right] = (1-\bar{\varepsilon}+\bar{\varepsilon}\bar{\alpha})\bar{\sigma}^2_{GM} + \bar{\mu}^2_{GM} = \mathrm{var}(z) + E^2_{p_{\mathrm{GM}}}[z]. \tag{C.10}$$

The objective function to be optimized becomes:

$$\begin{aligned} - E_p\left[\log\left(q(\boldsymbol{z};\boldsymbol{\theta})\right)\right] &= \frac{N}{2}\log\left(2\pi\sigma^2_{\mathrm{G}}\right) \\ + \frac{1}{2\sigma^2_{\mathrm{G}}}&\left(\sum_{i=1}^{N}\mathrm{var}(z_i) + \sum_{i=1}^{N}E_p\left[z_i\right]^2 - 2\mu_{\mathrm{G}}\sum_{i=1}^{N}E_p[z_i] + N\mu^2_{\mathrm{G}}\right), \end{aligned} \tag{C.11}$$

which can be further simplified to make it easier to obtain a generalized result for the pseudo-true parameter μ_{pt}:

$$\begin{aligned} -E_p\left[\log\left(q(\boldsymbol{z};\boldsymbol{\theta})\right)\right] &= \frac{N}{2}\log\left(2\pi\sigma^2_{\mathrm{G}}\right) \\ &+ \frac{N}{2\sigma^2_{\mathrm{G}}}\left(\mathrm{var}(z_i) + (E_p[z_i]-\mu_{\mathrm{G}})^2\right). \end{aligned} \tag{C.12}$$

The value of μ_{G} that minimizes the above cost function is obtained when $\mu_{\mathrm{G}} = E_p\left[z\right]$. Therefore,

$$\tilde{\mu}_p = E_p\left[z\right], \tag{C.13}$$

where the true distribution determines whether the pseudo-true location parameter coincides with the true parameter. As was shown above, random variables following the Student's t-distribution and the BGM both have a simple expression for the expected value. For the pseudo-true scale parameter, the following result is obtained:

$$-\frac{\partial}{\partial\sigma^2_{\mathrm{G}}}E_p\left[\log\left(q(\boldsymbol{z};\boldsymbol{\theta})\right)\right] = \frac{N}{2\sigma^2_{\mathrm{G}}} \tag{C.14}$$

$$- \frac{N}{2(\sigma^2_{\mathrm{G}})^2}\left(\mathrm{var}(z_i) + (E_p[z_i]-\mu_{\mathrm{G}})^2\right). \tag{C.15}$$

Substituting the pseudo-true parameter for μ_{G}, one obtains:

$$\frac{N\sigma^2_{\mathrm{G}}}{2} - \frac{N}{2}\mathrm{var}(z) = 0. \tag{C.16}$$

The resulting pseudo-true parameter is the sample variance of the Student's t-distribution:

$$\tilde{\sigma}^2_p = \mathrm{var}_{p_T}(z), \tag{C.17}$$

where the subscript p_T specifies the variance of the true distribution of the data. With the above result, we can conclude that the mean and variance of the Student's t-distribution are the pseudo-true parameters that minimize the KLD between a Gaussian distribution and a Student's t-distribution

$$\tilde{\mu}_T = \bar{\mu}_T, \tag{C.18}$$

$$\tilde{\sigma}_T^2 = \bar{\sigma}_T^2 \frac{\bar{\nu}}{\bar{\nu} - 2}. \tag{C.19}$$

The same is true for the BGM with the variance being a linear combination of the variances of two Gaussian modes because each mode has the same mean.

$$\tilde{\mu}_{\mathrm{GM}} = \bar{\mu}_{GM}, \tag{C.20}$$

$$\tilde{\sigma}_{\mathrm{GM}}^2 = \bar{\sigma}_{GM}^2 \left((\bar{\alpha} - 1)\bar{\varepsilon} + 1 \right). \tag{C.21}$$

Appendix D. Deriving Matrices A and B

The log-likelihood function of N i.i.d. Gaussian random variables in $\boldsymbol{z} = (z_1, \cdots, z_N)^T$ is:

$$\log(q(\boldsymbol{z}; \boldsymbol{\theta})) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma_{\mathrm{G}}^2) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{z_i - \mu_{\mathrm{G}}}{\sigma_{\mathrm{G}}}\right)^2. \tag{D.1}$$

Instead of using the joint PDF, we take advantage of the linearity of derivatives and expectations to compute $\mathbf{A}$ and $\mathbf{B}$ using the marginal PDF for a single sample:

$$\log(q(z_i; \boldsymbol{\theta})) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_{\mathrm{G}}^2) - \frac{1}{2}\left(\frac{z_i - \mu_{\mathrm{G}}}{\sigma_{\mathrm{G}}}\right)^2. \tag{D.2}$$

The equation for the MCRB then takes into account the N i.i.d. random variables. Computing the Hessian with respect to the parameter vector $\boldsymbol{\theta} = [\mu_{\mathrm{G}}, \sigma_{\mathrm{G}}^2]$ provides all the terms required to compute the required matrices:

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} = \left(E_p \left[\frac{\partial^2 \log(q(z_i; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right] \right)_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_p}, \tag{D.3}$$

$$= \left(E_p \begin{bmatrix} \frac{\partial^2 \log(q(z_i;\boldsymbol{\theta}))}{\partial \mu_{\mathrm{G}}^2} & \frac{\partial^2 \log(q(z_i;\boldsymbol{\theta}))}{\partial \mu_{\mathrm{G}} \partial \sigma_{\mathrm{G}}^2} \\ \frac{\partial^2 \log(q(z_i;\boldsymbol{\theta}))}{\partial \sigma_{\mathrm{G}}^2 \partial \mu_{\mathrm{G}}} & \frac{\partial^2 \log(q(z_i;\boldsymbol{\theta}))}{\partial (\sigma_{\mathrm{G}}^2)^2} \end{bmatrix} \right)_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_p}. \tag{D.4}$$

We follow a similar process to Appendix C to find the expectations of the above expressions. Substituting the pseudo-true values $\mu_{\mathrm{G}} = \tilde{\mu}_p$, $\sigma_{\mathrm{G}}^2 = \tilde{\sigma}_p^2$ yields:

$$A_{1,1} = \left(E_p \left[\frac{\partial^2 \log(q(z_i;\boldsymbol{\theta}))}{\partial \mu_{\mathrm{G}}^2} \right] \right)_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_p} = -\left(\mathrm{var}_p(z_i)\right)^{-1}, \tag{D.5}$$

$$A_{1,2} = \left(E_p \left[\frac{\partial^2 \log(q(z_i;\boldsymbol{\theta}))}{\partial \mu_{\mathrm{G}} \partial \sigma_{\mathrm{G}}^2} \right] \right)_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_p} = 0, \tag{D.6}$$

$$A_{2,2} = \left(E_p \left[\frac{\partial^2 \log(q(z_i;\boldsymbol{\theta}))}{\partial (\sigma_{\mathrm{G}}^2)^2} \right] \right)_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_p} = -\frac{1}{2}\left(\mathrm{var}_p(z_i)\right)^{-2}, \tag{D.7}$$

$$A_{2,1} = \left(E_p \left[\frac{\partial^2 \log(q(z_i;\boldsymbol{\theta}))}{\partial \sigma_{\mathrm{G}}^2 \partial \mu_{\mathrm{G}}} \right] \right)_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_p} = 0. \tag{D.8}$$

where the term $A_{2,2}$ has been computed using $E_p \left[\left(\frac{z_i - \mu_{\mathrm{G}}}{\sigma_{\mathrm{G}}} \right)^2 \right]$ from Appendix C:

$$\left(E_p \left[\left(\frac{z_i - \mu_{\mathrm{G}}}{\sigma_{\mathrm{G}}} \right)^2 \right] \right)_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_p} = \frac{(\mathrm{var}_p(z_i) + (E_p[z_i] - E_p[z_i])^2)}{\mathrm{var}_p(z_i)} = 1. \tag{D.9}$$

This provides all expressions necessary for $\mathbf{A}$:

$$\mathbf{A}(\tilde{\boldsymbol{\theta}}_p) = \begin{bmatrix} -\left(\mathrm{var}_p(z_i)\right)^{-1} & 0 \\ 0 & -\frac{1}{2}\left(\mathrm{var}_p(z_i)\right)^{-2} \end{bmatrix}. \tag{D.10}$$

The elements of $\mathbf{B}$ can be expressed as

$$\mathbf{B} = \left(E_p \left[\left(\frac{\partial \log(q(z_i;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \log(q(z_i;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^T} \right) \right] \right)_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_p}, \tag{D.11}$$

$$= \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix}. \tag{D.12}$$

The expectation of $B_{1,1}$ is computed using (D.9):

$$E_p \left[\frac{\partial \log(q(z_i;\boldsymbol{\theta}))}{\partial \mu_{\mathrm{G}}} \frac{\partial \log(q(z_i;\boldsymbol{\theta}))}{\partial \mu_{\mathrm{G}}} \right] = \frac{1}{\sigma_{\mathrm{G}}^2} E_p \left[\left(\frac{z_i - \mu_{\mathrm{G}}}{\sigma_{\mathrm{G}}} \right)^2 \right]. \tag{D.13}$$

Substituting the pseudo-true values and using (D.9) leads to:

$$B_{1,1} = \left(\mathrm{var}_p(z_i)\right)^{-1}. \tag{D.14}$$

The terms $B_{1,2}$ and $B_{2,1}$ are zero for a symmetric distribution

$$B_{1,2} = B_{2,1} = 0. \tag{D.15}$$

Next, we expand the expression for $B_{2,2}$ and substitute the pseudo-true parameters

$$B_{2,2} = E_p \left[\frac{1}{4(\tilde{\sigma}_p^2)^2} - \frac{1}{2\tilde{\sigma}_p^2} \frac{(z_i - \tilde{\mu}_p)^2}{(\tilde{\sigma}_p^2)^2} + \frac{1}{4} \left(\frac{(z_i - \tilde{\mu}_p)^2}{(\tilde{\sigma}_p^2)^2} \right)^2 \right], \tag{D.16}$$

The expression then becomes:

$$B_{2,2} = \left(-\frac{1}{4(\sigma_{\mathrm{G}}^2)^2} + \frac{1}{4(\sigma_{\mathrm{G}}^2)^4} E_p \left[(z_i - E_p[z_i])^4 \right] \right)_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_p}. \tag{D.17}$$

For the final element of the **B** matrix, the fourth-order central moment of the true distribution should be evaluated. The expression will differ depending on the specific distribution that truly describes the data so the **B** matrix is split based on the true distribution. The fourth-order central moment is known for the Student's t-distribution [38]:

$$\left(E_{p_T} \left[(z_i - \mu_T)^4 \right] \right)_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_p} = \frac{3\bar{\nu}^2}{(\bar{\nu} - 2)(\bar{\nu} - 4)} (\bar{\sigma}_T^2)^2, \tag{D.18}$$

We substitute the above results into the expression for $B_{2,2}$ to give the result for the misspecified scenario with the two possible true models. The first result being for the Student's t-distribution

$$B_{2,2}(p_T || q) = -\frac{1}{4} \left(\bar{\sigma}_T^2 \frac{\bar{\nu}}{\bar{\nu} - 2} \right)^{-2} + \frac{3(\bar{\nu} - 2)}{4(\bar{\nu} - 4)} \left(\bar{\sigma}_T^2 \frac{\bar{\nu}}{\bar{\nu} - 2} \right)^{-2}. \tag{D.19}$$

Therefore, we obtain the following matrix:

$$\mathbf{B}(p_T || q) = \begin{bmatrix} \left(\bar{\sigma}_T^2 \frac{\bar{\nu}}{\bar{\nu} - 2} \right)^{-1} & 0 \\ 0 & \left(\frac{3(\bar{\nu} - 2)}{4(\bar{\nu} - 4)} - \frac{1}{4} \right) \left(\sigma_T^2 \frac{\bar{\nu}}{\bar{\nu} - 2} \right)^{-2} \end{bmatrix}. \tag{D.20}$$

For simplicity, we can write both **A** and **B** in terms of the pseudo-true scale parameter, which we know to be the variance of the true distribution, in this case being the Student's t-distribution

$$\mathbf{A} = \begin{bmatrix} -\left(\tilde{\sigma}_p^2 \right)^{-1} & 0 \\ 0 & -\frac{1}{2} \left(\tilde{\sigma}_p^2 \right)^{-2} \end{bmatrix}, \tag{D.21}$$

$$\mathbf{B}(p_T || q) = \begin{bmatrix} \left(\tilde{\sigma}_p^2 \right)^{-1} & 0 \\ 0 & \left(\frac{\bar{\nu} - 1}{2(\bar{\nu} - 4)} \right) \left(\tilde{\sigma}_p^2 \right)^{-2} \end{bmatrix}. \tag{D.22}$$

All terms are the same in both **A** and **B** for any other type of true distribution when considering the Gaussian model. The only element with a particular expression is $B_{2,2}$, which depends on the true distribution. For the BGM, the last term should be derived separately. The expression for the fourth-order central moment can be simplified as a linear combination of the fourth-order moments for each of the modes

$$\begin{aligned} &\left(E_{p_{\mathrm{GM}}} \left[(z_i - E_{p_{\mathrm{GM}}} [z_i])^4 \right] \right)_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_p} \\ &= (1 - \bar{\varepsilon}) \left(E_{g_0} \left[(z_i - \mu)^4 \right] \right)_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}} + \bar{\varepsilon} \left(E_{g_1} \left[(z_i - \mu)^4 \right] \right)_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}}. \end{aligned} \tag{D.23}$$

The fourth-order moment for a Gaussian distribution is also already established, so for the distributions g_0 and g_1 are:

$$E_{g_0}\left[(z_i-\mu)^4\right] = 3\bar{\sigma}_{GM}^4, \tag{D.24}$$

$$E_{g_1}\left[(z_i-\mu)^4\right] = 3\bar{\alpha}^2\bar{\sigma}_{GM}^4. \tag{D.25}$$

Leading to the resulting central moment for the BGM

$$\begin{aligned}&\left(E_{p_{\mathrm{GM}}}\left[(z_i - E_{p_{\mathrm{GM}}}[z_i])^4\right]\right)_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_p}\\ &= 3(1-\bar{\varepsilon})\bar{\sigma}_{GM}^4 + 3\bar{\varepsilon}\bar{\alpha}^2\bar{\sigma}_{GM}^4 = 3\bar{\sigma}_{GM}^4((\bar{\alpha}^2-1)\bar{\varepsilon}+1).\end{aligned} \tag{D.26}$$

The above is substituted into (D.17) and allows us to evaluate the last term of the corresponding matrix

$$B_{2,2}(p_{\mathrm{GM}}||q) = -\frac{1}{4(\tilde{\sigma}_{p_{\mathrm{GM}}}^2)^2} + \frac{1}{4(\tilde{\sigma}_{p_{\mathrm{GM}}}^2)^4}3\bar{\sigma}_{GM}^4((\bar{\alpha}^2-1)\bar{\varepsilon}+1), \tag{D.27}$$

with significant simplifications after substituting the pseudo-true parameter for dispersion:

$$B_{2,2}(p_{\mathrm{GM}}||q) = \frac{-\bar{\varepsilon}^2(\bar{\alpha}-1)^2 + \bar{\varepsilon}(\bar{\alpha}-1)(3\bar{\alpha}+1)+2}{4\bar{\sigma}_{GM}^4(\bar{\varepsilon}(\bar{\alpha}-1)+1)^4}. \tag{D.28}$$

To reduce the size of the following expressions, the term $\phi(\bar{\varepsilon},\bar{\alpha}) = \bar{\varepsilon}(\bar{\alpha}-1)$ is defined and the numerator of $B_{2,2}(p_{\mathrm{GM}}||q)$ is denoted as the quadratic function of that term, i.e.,

$$Q(\phi(\bar{\varepsilon},\bar{\alpha})) = -\bar{\varepsilon}^2(\bar{\alpha}-1)^2 + \bar{\varepsilon}(\bar{\alpha}-1)(3\bar{\alpha}+1)+2. \tag{D.29}$$

The matrix $\mathbf{B}(p_{\mathrm{GM}}||q)$ is then written as:

$$\mathbf{B}(p_{\mathrm{GM}}||q) = \begin{bmatrix}\left(\tilde{\sigma}_p^2\right)^{-1} & 0\\ 0 & \frac{Q(\phi)}{4(\phi+1)^2}\left(\tilde{\sigma}_{p_{\mathrm{GM}}}^2\right)^{-2}\end{bmatrix}. \tag{D.30}$$

Appendix E. MCRB computation

The final computation of the MCRB is a simple matrix multiplication, where the inverse of $\mathbf{A}$ is trivial

$$\mathbf{MCRB}_{\boldsymbol{\theta}}(p||q) = \frac{1}{N}\mathbf{A}^{-1}\mathbf{B}(p||q)\mathbf{A}^{-1}. \tag{E.1}$$

For the same misspecified assumption of a Gaussian distribution for the data, only the term $\mathbf{B}$ varies according to the true distribution. First the misspecified bound where the true data follows a Student's t-distribution is shown in (E.3) and the MCRB for the case where the true distribution is a BGM is then shown in (E.5).

$\mathbf{MCRB}_{\boldsymbol{\theta}}(p_T||q)$

$$= \frac{1}{N}\begin{bmatrix} -\tilde{\sigma}^2_{p_T} & 0 \\ 0 & -2\left(\tilde{\sigma}^2_{p_T}\right)^2 \end{bmatrix} \begin{bmatrix} \left(\tilde{\sigma}^2_{p_T}\right)^{-1} & 0 \\ 0 & \left(\frac{\bar{\nu}-1}{2(\bar{\nu}-4)}\right)\left(\tilde{\sigma}^2_{p_T}\right)^{-2} \end{bmatrix} \begin{bmatrix} -\tilde{\sigma}^2_{p_T} & 0 \\ 0 & -2\left(\tilde{\sigma}^2_{p_T}\right)^2 \end{bmatrix}, \tag{E.2}$$

$$= \frac{1}{N}\begin{bmatrix} \tilde{\sigma}^2_{p_T} & 0 \\ 0 & \left(\frac{2(\bar{\nu}-1)}{(\bar{\nu}-4)}\right)\left(\tilde{\sigma}^2_{p_T}\right)^2 \end{bmatrix}, \tag{E.3}$$

$\mathbf{MCRB}_{\boldsymbol{\theta}}(p_{\mathrm{GM}}||q)$

$$= \frac{1}{N}\begin{bmatrix} -\tilde{\sigma}^2_{p_{\mathrm{GM}}} & 0 \\ 0 & -2\left(\tilde{\sigma}^2_{p_{\mathrm{GM}}}\right)^2 \end{bmatrix} \begin{bmatrix} \left(\tilde{\sigma}^2_{p_{\mathrm{GM}}}\right)^{-1} & 0 \\ 0 & \frac{Q(\phi)}{4(\phi+1)^2}\left(\tilde{\sigma}^2_{p_{\mathrm{GM}}}\right)^{-2} \end{bmatrix} \begin{bmatrix} -\tilde{\sigma}^2_{p_{\mathrm{GM}}} & 0 \\ 0 & -2\left(\tilde{\sigma}^2_{p_{\mathrm{GM}}}\right)^2 \end{bmatrix}, \tag{E.4}$$

$$= \frac{1}{N}\begin{bmatrix} \tilde{\sigma}^2_{p_{\mathrm{GM}}} & 0 \\ 0 & \frac{Q(\phi)}{(\phi+1)^2}\left(\tilde{\sigma}^2_{p_{\mathrm{GM}}}\right)^2 \end{bmatrix}. \tag{E.5}$$

Appendix F. Derivation of CRB for Joint Estimation of Location, Scale, and Shape of a Student's t-distribution

The FIM for joint estimation of μ, σ^2, and ν is known in general for a k-variate Student's t-distribution [5]:

$$\mathbf{F}_{\boldsymbol{\eta}} = \begin{bmatrix} F_{\mu} & F_{\mu,\sigma^2} & F_{\mu,\nu} \\ F_{\sigma^2,\mu} & F_{\sigma^2} & F_{\sigma^2,\nu} \\ F_{\nu,\mu} & F_{\nu,\sigma^2} & F_{\nu} \end{bmatrix} = \begin{bmatrix} \frac{\nu+k}{\nu+k+2}\sigma^{-2} & 0 & 0 \\ 0 & \frac{\nu+k-1}{2(\nu+k+2)}\sigma^{-4} & -\frac{\sigma^{-2}}{(\nu+k)(\nu+k+2)} \\ 0 & -\frac{\sigma^{-2}}{(\nu+k)(\nu+k+2)} & -\xi(\nu) - \frac{k(\nu+k+4)}{2\nu(\nu+k)(\nu+k+2)} \end{bmatrix}, \tag{F.1}$$

where we have simplified the form for a covariance matrix given by $\sigma^2\mathbf{I}$, $\xi(\nu) = \frac{1}{4}\left(\psi'\left(\frac{\nu+1}{2}\right) - \psi'\left(\frac{\nu}{2}\right)\right)$, and $\psi'(x) = \frac{\partial^2}{\partial x^2}\log\left(\Gamma(x)\right)$ is the trigamma function. The inverse of the above FIM provides the CRB, the exact form of the CRB was not revealed in [5] nor analysed in works since

$$\mathbf{CRB}_{\boldsymbol{\eta}} = \mathbf{F}_{\boldsymbol{\eta}}^{-1} = \begin{bmatrix} F_{\mu} & 0 & 0 \\ 0 & F_{\sigma^2} & F_{\sigma^2,\nu} \\ 0 & F_{\nu,\sigma^2} & F_{\nu} \end{bmatrix}^{-1} = \begin{bmatrix} F_{\mu}^{-1} & \mathbf{0} \\ 0 & \mathbf{M}^{-1} \end{bmatrix}, \tag{F.2}$$

where $\mathbf{M}$ is the 2×2 block containing the information on the scale and shape parameters. The inverse of $\mathbf{M}$ provides the CRB for the scale and shape parameters of the Student's t-distribution when jointly estimating both parameters

$$\mathbf{M}^{-1} = \frac{1}{|\mathbf{M}|}\begin{bmatrix} -\xi(\nu) - \frac{k(\nu+k+4)}{2\nu(\nu+k)(\nu+k+2)} & \frac{\sigma^{-2}}{(\nu+k)(\nu+k+2)} \\ \frac{\sigma^{-2}}{(\nu+k)(\nu+k+2)} & \frac{\nu+k-1}{2(\nu+k+2)}\sigma^{-4} \end{bmatrix}. \tag{F.3}$$

The simplification of this inverse requires evaluation of the determinant

$$|\mathbf{M}| = \left(\frac{(\nu+k-1)}{2(\nu+k+2)}\sigma^{-4}\right)\left(-\xi(\nu) - \frac{k(\nu+k+4)}{2\nu(\nu+k)(\nu+k+2)}\right) - \frac{\sigma^{-4}}{(\nu+k)^2(\nu+k+2)^2}, \tag{F.4}$$

$$= \frac{-2\nu(\nu+k)^2(\nu+k-1)(\nu+k+2)\xi(\nu) - k(\nu+k+4)(\nu+k)(\nu+k-1) - 4\nu}{4\nu(\nu+k)^2(\nu+k+2)^2}\sigma^{-4}, \tag{F.5}$$

$$= \frac{-f(\nu,k) - 4\nu}{4\nu(\nu+k)^2(\nu+k+2)^2}\sigma^{-4}, \tag{F.6}$$

where $f(\nu,k) = 2\nu(\nu+k)^2(\nu+k-1)(\nu+k+2)\xi(\nu) + k(\nu+k+4)(\nu+k)(\nu+k-1)$ is defined for brevity and later use in simplification. The CRB for the estimation of the scale parameter (while jointly estimating ν) is then computed as

$$\mathrm{CRB}_{\sigma^2} = \left(2\sigma^4\right)\frac{2\nu(\nu+k)^2(\nu+k+2)^2}{-f(\nu,k)-4\nu}\left(-\xi(\nu) - \frac{k(\nu+k+4)}{2\nu(\nu+k)(\nu+k+2)}\right), \tag{F.7}$$

$$= 2\sigma^4\left(\frac{\nu+k+2}{\nu+k-1}\right)\frac{2\nu(\nu+k)^2(\nu+k+2)(\nu+k-1)}{f(\nu,k)+4\nu}\left(\xi(\nu) + \frac{k(\nu+k+4)}{2\nu(\nu+k)(\nu+k+2)}\right), \tag{F.8}$$

$$= 2\sigma^4\left(\frac{\nu+k+2}{\nu+k-1}\right)\frac{2\nu(\nu+k)^2(\nu+k-1)(\nu+k+2)\xi(\nu) + k(\nu+k+4)(\nu+k)(\nu+k-1)}{f(\nu,k)+4\nu}, \tag{F.9}$$

$$= 2\sigma^4\left(\frac{\nu+k+2}{\nu+k-1}\right)\frac{f(\nu,k)}{f(\nu,k)+4\nu}, \tag{F.10}$$

$$= 2\sigma^4\left(\frac{\nu+k+2}{\nu+k-1}\right)\left(1 - \frac{4\nu}{f(\nu,k)+4\nu}\right), \tag{F.11}$$

Appendix G. Derivation of CRB for Joint Estimation of Location, Scale, Proportion, and Contamination of Bivariance Gaussian Mixture distribution

For a real Bivariance Gaussian Mixture with PDF, parameterized by the parameter vector $\boldsymbol{\beta} = [\mu, \sigma^2, \varepsilon, \alpha]^T$

$$p(z;\boldsymbol{\beta}) = \frac{(1-\varepsilon)}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2} + \frac{\varepsilon}{\sqrt{2\pi\alpha\sigma^2}}e^{-\frac{1}{2\alpha}\left(\frac{z-\mu}{\sigma}\right)^2}, \tag{G.1}$$

the FIM is not as straight-forward as other symmetric distributions because the log-likelihood function does not simplify with the sum of two Gaussian distributions. The FIM can be computed as

$$F_{\beta_i,\beta_j} = -E\left[\frac{\partial^2 \log\left((p(z;\boldsymbol{\beta})\right)}{\partial\beta_i\partial\beta_j}\right], \tag{G.2}$$

where the logarithm is not evaluated and the chain rule provides a different formulation of the FIM

$$F_{\beta_i,\beta_j} = E\left[\frac{1}{p(z;\boldsymbol{\beta})^2}\frac{\partial p(z;\boldsymbol{\beta})}{\partial\beta_i}\frac{\partial p(z;\boldsymbol{\beta})}{\partial\beta_j}\right] - E\left[\frac{1}{p(z;\boldsymbol{\beta})}\frac{\partial^2 p(z;\boldsymbol{\beta})}{\partial\beta_i\partial\beta_j}\right]. \tag{G.3}$$

The second term can have the order of second derivative and expectation changed, resulting in that term being equal to zero. The first term is equivalent to Gaussian functions divided by the true likelihood, being a Gaussian mixture. This type of integral is not known in closed-form and must be numerically approximated. The approximation is not considered in the scope of this work, which aims at deriving closed form expressions to provide a definitive bound for the estimation performance under heavy-tailed models. Consult [17] for an example using a Lerch transcendental to approximate the bound in the case of a complex Gaussian mixture.

Appendix H. Derivation of MCRB for Estimation of Location and Scale using a Misspecified Heavy-tailed Distribution

The derivation of the MCRB begins with the derivation of the pseudotrue parameters defined for some combination of true and assumed distributions. If we want to derive the lower bound for assuming a Student's t-distribution while the data is truly generated by a BGM, then we must find the values in $\boldsymbol{\eta} = [\mu_T, \sigma_T^2, \nu]^T$ that minimize the KLD from the true distribution. That is,

$$\tilde{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}}\left\{E_{p_{\mathrm{GM}}}\left[-\log\left(p_T(z;\boldsymbol{\eta})\right)\right]\right\}, \tag{H.1}$$

where p_{GM} is the PDF of the BGM model with parameters $\boldsymbol{\beta} = [\mu_{\mathrm{GM}}, \sigma_{\mathrm{GM}}^2, \varepsilon, \alpha]^T$ and $p_T(z;\boldsymbol{\eta})$ is the PDF of the Student's t-distribution. Each of these two probability densities depends on the central, normalized univariate Mahalanobis distance, but with different parameters. For example, we can make a substitution $u_i = \frac{z_i-\mu_T}{\sigma_T}$, then the equivalent Mahalanobis distance in p_{GM} is $\frac{z_i-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}} = u_i\frac{\sigma_T}{\sigma_{\mathrm{GM}}} + \frac{\mu_T-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}$. With this change of variables substituted into the mathematical expectation, we obtain the following integral that the pseudotrue values must minimize:

$$\int_{-\infty}^{\infty}\left(\frac{1}{2}\log\left(\pi\nu\sigma_T^2\right) + \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) - \log\left(\Gamma\left(\frac{\nu+1}{2}\right)\right) + \left(\frac{\nu+1}{2}\right)\log\left(1+\frac{u_i^2}{\nu}\right)\right)\cdot$$
$$\left(\frac{(1-\varepsilon)}{\sqrt{2\pi\sigma_{\mathrm{GM}}^2}}e^{-\frac{1}{2}\left(u_i\frac{\sigma_T}{\sigma_{\mathrm{GM}}}+\frac{\mu_T-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}\right)^2} + \frac{\varepsilon}{\sqrt{2\pi\alpha\sigma_{\mathrm{GM}}^2}}e^{-\frac{1}{2\alpha}\left(u_i\frac{\sigma_T}{\sigma_{\mathrm{GM}}}+\frac{\mu_T-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}\right)^2}\right)\sigma_T du. \tag{H.2}$$

To define the pseudotrue location parameter, we only need to minimize the terms in the above equation that include μ_T

$$\tilde{\mu} = \arg\min_{\mu_T}\left\{\int_{-\infty}^{\infty}\log\left(1+\frac{u_i^2}{\nu}\right)e^{-\frac{1}{2}\left(u_i\frac{\sigma_T}{\sigma_{\mathrm{GM}}}+\frac{\mu_T-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}\right)^2}du\right.$$
$$\left.+\int_{-\infty}^{\infty}\log\left(1+\frac{u_i^2}{\nu}\right)e^{-\frac{1}{2\alpha}\left(u_i\frac{\sigma_T}{\sigma_{\mathrm{GM}}}+\frac{\mu_T-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}\right)^2}du\right\}. \tag{H.3}$$

The above integrals are minimum when their derivatives are equal to zero,

$$\int_{-\infty}^{\infty} \frac{\frac{u_i}{\nu\sigma_T}}{\left(1+\frac{u_i^2}{\nu}\right)} e^{-\frac{1}{2}\left(u_i\frac{\sigma_T}{\sigma_{\mathrm{GM}}}+\frac{\mu_T-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}\right)^2} du$$
$$+\int_{-\infty}^{\infty} \frac{\frac{u_i}{\nu\sigma_T}}{\left(1+\frac{u_i^2}{\nu}\right)} e^{-\frac{1}{2\alpha}\left(u_i\frac{\sigma_T}{\sigma_{\mathrm{GM}}}+\frac{\mu_T-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}\right)^2} du = 0. \tag{H.4}$$

and these derivatives are zero when the function being integrated is odd in u. This leads to the simple conclusion that the pseudotrue location parameter is equal to the true location parameter

$$\tilde{\mu} = \mu_{\mathrm{GM}}. \tag{H.5}$$

A similar derivation provides the same result if the assumed and true distributions are switched, so we have a mathematical proof that the misspecified estimation of the location parameter of one heavy-tailed distribution is not biased with respect to the fully specified estimator. However, the bound could still be different because it also depends on the pseudotrue scale parameter and the derivatives required to obtain the MCRB.

The pseudotrue scale parameter is not as straightforward to derive, substituting the pseudotrue location parameter, the function to minimize is

$$\int_{-\infty}^{\infty} \left(\frac{1}{2}\log\left(\pi\nu\sigma_T^2\right) + \left(\frac{\nu+1}{2}\right)\log\left(1+\frac{1}{\nu}\left(\frac{z_i-\mu_{\mathrm{GM}}}{\sigma_T}\right)^2\right)\right)\cdot$$
$$\left(\frac{(1-\varepsilon)}{\sqrt{2\pi\sigma_{\mathrm{GM}}^2}} e^{-\frac{1}{2}\left(\frac{z_i-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}\right)^2} + \frac{\varepsilon}{\sqrt{2\pi\alpha\sigma_{\mathrm{GM}}^2}} e^{-\frac{1}{2\alpha}\left(\frac{z_i-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}\right)^2}\right) dz. \tag{H.6}$$

Next, the equation to solve for the pseudotrue scale parameter is obtained after taking the derivative w.r.t. σ_T^2

$$\frac{1}{2\sigma_T^2} = \left(\frac{\nu+1}{\nu}\right)\int_0^{\infty} \frac{\left(\frac{z_i-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}\right)^2 \frac{\sigma_{\mathrm{GM}}^2}{(\sigma_T^2)^2}}{1+\frac{\sigma_{\mathrm{GM}}^2}{\nu\sigma_T^2}\left(\frac{z_i-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}\right)^2} \left(c_1 e^{-\frac{1}{2}\left(\frac{z_i-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}\right)^2} + c_2 e^{-\frac{1}{2\alpha}\left(\frac{z_i-\mu_{\mathrm{GM}}}{\sigma_{\mathrm{GM}}}\right)^2}\right) dz, \tag{H.7}$$

where $c_1 = \frac{(1-\varepsilon)}{\sqrt{2\pi\sigma_{\mathrm{GM}}^2}}$ and $c_2 = \frac{\varepsilon}{\sqrt{2\pi\alpha\sigma_{\mathrm{GM}}^2}}$ are used for brevity. We can now simplify the equation that must be satisfied to find the pseudotrue scale parameter

$$\frac{1}{2} = \sigma_{\mathrm{GM}}(\nu+1)\int_0^{\infty} \frac{u^2}{u^2+\nu\frac{\sigma_T^2}{\sigma_{\mathrm{GM}}^2}} \left(c_1 e^{-\frac{1}{2}u^2} + c_2 e^{-\frac{1}{2\alpha}u^2}\right) du, \tag{H.8}$$

The above integral has the same form as the linear combination of two integrals in the form of equation 3.466.2 in [39]. As a result, we can evaluate the integral to see if we can obtain a closed form for the pseudotrue scale parameter.

$$\begin{aligned}\frac{1}{2\sigma_{\mathrm{GM}}(\nu+1)} &= c_1\left(\sqrt{\frac{\pi}{2}} - \frac{\pi}{2}\frac{\sigma_T}{\sigma_{\mathrm{GM}}}\sqrt{\nu}e^{\frac{\nu\sigma_T^2}{2\sigma_{\mathrm{GM}}^2}}\left[1-\Phi\left(\sqrt{\frac{\nu}{2}}\frac{\sigma_T}{\sigma_{\mathrm{GM}}}\right)\right]\right)\\ &\quad + c_2\left(\sqrt{\frac{\pi}{2\alpha}} - \frac{\pi}{2}\frac{\sigma_T}{\sigma_{\mathrm{GM}}}\sqrt{\nu}e^{\frac{\nu\sigma_T^2}{2\alpha\sigma_{\mathrm{GM}}^2}}\left[1-\Phi\left(\sqrt{\frac{\nu}{2\alpha}}\frac{\sigma_T}{\sigma_{\mathrm{GM}}}\right)\right]\right). \qquad \text{(H.9)}\end{aligned}$$

Solving the above equation for σ_T is required to find the pseudotrue scale parameter, however, this will require some numerical methods to iteratively converge to a solution. At this point, it is clear that a closed-form cannot be derived for the MCRB assuming a Student's t-distribution when data follows a BGM. The numerical convergence to the solution is saved for future work along with the following steps of derivations for the MCRB for this combination of true and assumed distributions and the vice-versa.

References

[1] M. A. Chmielewski, "Elliptically symmetric distributions: A review and bibliography," *International Statistical Review/Revue Internationale de Statistique*, pp. 67–74, 1981.

[2] L. Ortega and S. Fortunati, "Misspecified time-delay and Doppler estimation over non-Gaussian scenarios," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Seoul, Korea), pp. 9346–9350, 2024.

[3] S. Fortunati and L. Ortega, "On the efficiency of misspecified Gaussian inference in nonlinear regression: Application to time-delay and Doppler estimation," *Signal Processing*, vol. 225, p. 109614, 2024.

[4] H. McPhee, J.-Y. Tourneret, D. Valat, J. Delporte, Y. Grégoire, and P. Paimblanc, "Misspecified Cramér-Rao bounds for anomalous clock data in satellite constellations," in *Proc. European Signal Processing Conference (EUSIPCO)*, (Lyon, France), pp. 1222–1226, 2024.

[5] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989.

[6] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis.* Chapman & Hall/CRC texts in statistical science, Boca Raton: CRC Press, third edition. ed., 2014.

[7] J. W. Tukey, "A survey of sampling from contaminated distributions," *Contributions to probability and statistics*, pp. 448–485, 1960.

[8] R. A. Maronna, R. D. Martin, and V. J. Yohai, *Robust Statistics: Theory and Methods.* Wiley Series in Probability and Statistics, Wiley, 1st ed., Mar. 2006.

[9] K. R. Koch, “Robust estimation by expectation maximization algorithm,” *Journal of Geodesy*, vol. 87, pp. 107–116, 2013.

[10] Q. H. Vuong, *Cramér-Rao bounds for misspecified models.* working paper 652, Div. of the Humanities and Social Sci., Caltech, Pasadena, USA, 1986.

[11] C. D. Richmond and L. L. Horowitz, “Parameter bounds on estimation accuracy under model misspecification,” *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2263–2278, 2015.

[12] A. Mennad, S. Fortunati, M. N. El Korso, A. Younsi, A. M. Zoubir, and A. Renaux, “Slepian-Bangs-type formulas and the related misspecified Cramér-Rao bounds for complex elliptically symmetric distributions,” *Signal Processing*, vol. 142, pp. 320–329, 2018.

[13] L. T. Thanh, K. Abed-Meraim, and N. L. Trung, “Misspecified Cramér–Rao bounds for blind channel estimation under channel order misspecification,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 5372–5385, 2021.

[14] J. Gorman and A. Hero, “Lower bounds for parametric estimation with constraints,” *IEEE Transactions on Information Theory*, vol. 36, no. 6, pp. 1285–1301, 1990.

[15] S. Fortunati, F. Gini, and M. S. Greco, “The constrained misspecified cramér–rao bound,” *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 718–721, 2016.

[16] R. Piché, “Cramér-Rao lower bound for linear filtering with t-distributed measurement noise,” in *Proc. International Conference on Information Fusion*, (Heidelberg, Germany), 2016.

[17] S. Kalyani, “On CRB for parameter estimation in two component Gaussian mixtures and the impact of misspecification,” *IEEE Transactions on Communications*, vol. 60, no. 12, pp. 3734–3744, 2012.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, pp. 1–22, 12 2018.

[19] F. Z. Doğru, Y. M. Bulut, and O. Arslan, “Doubly reweighted estimators for the parameters of the multivariate t-distribution,” *Communications in Statistics-Theory and Methods*, vol. 47, no. 19, pp. 4751–4771, 2018.

[20] M. Hasannasab, J. Hertrich, F. Laus, and G. Steidl, “Alternatives to the EM algorithm for ML estimation of location, scatter matrix, and degree of freedom of the Student t distribution,” *Numerical Algorithms*, vol. 87, pp. 77–118, Sept. 2020.

[21] T. Bonald, "Expectation-maximization for the gaussian mixture model," 2019. Telecom ParisTech. Available: https://perso.telecom-paristech.fr/bonald/documents/gmm.pdf.

[22] C. Ren, M. N. El Korso, J. Galy, E. Chaumette, P. Larzabal, and A. Renaux, "Performance bounds under misspecification model for mimo radar application," in *Proc. 23rd European Signal Processing Conference (EUSIPCO)*, pp. 514–518, 2015.

[23] H. McPhee, L. Ortega, J. Vilà-Valls, and E. Chaumette, "On the accuracy limits of misspecified delay-doppler estimation," *Signal Processing*, vol. 205, p. 108872, 2023.

[24] H. McPhee, L. Ortega, J. Vilà-Valls, and E. Chaumette, "Accounting for acceleration — signal parameters estimation performance limits in high dynamics applications," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 1, pp. 610–622, 2023.

[25] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[26] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79 – 86, 1951.

[27] P. J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proc. of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, (Berkeley, USA), pp. 221–233, 1965.

[28] C. D. Richmond and L. L. Horowitz, "Parameter bounds under misspecified models," in *Proc. Asilomar Conference on Signals, Systems and Computers*, pp. 176–180, 2013.

[29] S. Fortunati, F. Gini, and M. S. Greco, "The misspecified Cramér-Rao bound and its application to scatter matrix estimation in complex elliptically symmetric distributions," *IEEE Transactions on Signal Processing*, vol. 64, no. 9, pp. 2387–2399, 2016.

[30] S. M. Kay, *Fundamentals of statistical signal processing.* Prentice Hall signal processing series, Englewood Cliffs, N.J: Prentice-Hall PTR, 1993.

[31] A. Farina, B. Ristic, and L. Timmoneri, "Cramér-rao bound for nonlinear filtering with $Pd < 1$ and its application to target tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 8, pp. 1916–1924, 2002.

[32] D. P. Palomar, *Portfolio Optimization: Theory and Application.* Cambridge University Press, 2025.

[33] H. McPhee, J.-Y. Tourneret, D. Valat, J. Delporte, Y. Grégoire, and P. Paimblanc, "A robust time scale for space applications using the Student's t-distribution," *Metrologia*, vol. 61, p. 055010, sep 2024.

[34] M. Goldstein, “Unsupervised anomaly detection benchmark,” 2015. Harvard Dataverse, Version V1. Available: https://doi.org/10.7910/DVN/OPQMVF.

[35] P. J. Napier, D. S. Bagri, B. G. Clark, A. E. Rogers, J. D. Romney, A. R. Thompson, and R. C. Walker, “The very long baseline array,” *Proc. of the IEEE*, vol. 82, no. 5, pp. 658–672, 1994.

[36] R. T. Rajan and A.-J. van der Veen, “Joint ranging and synchronization for an anchorless network of mobile nodes,” *IEEE Transactions on Signal Processing*, vol. 63, no. 8, pp. 1925–1940, 2015.

[37] B. Cecconi, M. Dekkali, C. Briand, B. Segret, J. N. Girard, A. Laurens, A. Lamy, D. Valat, M. Delpech, M. Bruno, P. Gelard, M. Bucher, Q. Nenon, J.-M. Griesmeier, A.-J. Boonstra, and M. Bentum, “NOIRE study report: Towards a low frequency radio interferometer in space,” in *Proc. IEEE Aerospace Conference*, (Big Sky, MT, USA), pp. 1–19, Mar. 2018.

[38] J. Kirkby, N. Dang, and D. Nguyen, “Moments of Student’s t-distribution: A unified approach,” Dec. 2019.

[39] I. S. Gradshteĭn, I. M. Ryzhik, A. Jeffrey, and D. Zwillinger, *Table of integrals, series and products.* Oxford: Academic, Seventh edition ed., 2007.