Bayesian Lower Bounds for Dense or Sparse (Outlier) Noise in the RMT Framework

Virginie Ollier, Rémy Boyer, Mohammed Nabil El Korso, Pascal Larzabal

Bayesian Lower Bounds for Dense or Sparse (Outlier) Noise in the RMT Framework

Virginie Ollier*†, Rémy Boyer†, Mohammed Nabil El Korso‡ and Pascal Larzabal*
*SATIE, UMR 8029, Université Paris-Saclay, ENS Cachan, Cachan, France
†L2S, UMR 8506, Université Paris-Saclay, Université Paris-Sud, Gif-sur-Yvette, France
‡LEME, EA 4416, Université Paris-Ouest, Ville d'Avray, France

***Abstract*—Robust estimation is an important and timely research subject. In this paper, we investigate performance lower bounds on the mean-square-error (MSE) of any estimator for the Bayesian linear model, corrupted by a noise distributed according to an i.i.d. Student's t-distribution. This class of prior parametrized by its degree of freedom is relevant to modelize either dense or sparse (accounting for outliers) noise. Using the hierarchical Normal-Gamma representation of the Student's t-distribution, the Van Trees' Bayesian Cramér-Rao bounds (BCRBs) on the amplitude parameters and the noise hyper-parameter are derived. Furthermore, the Random Matrix Theory (RMT) framework is assumed, *i.e.*, the number of measurements and the number of unknown parameters grow jointly to infinity with an asymptotic finite ratio. Using some powerful results from the RMT, closed-form expressions of the BCRB are derived and studied. Finally, we propose a framework to fairly compare two models corrupted by noises with different degrees of freedom for a fixed common target signal-to-noise ratio (SNR). In particular, we focus our effort on the comparison of the BCRBs associated with two models corrupted by a sparse noise promoting outliers and a dense (Gaussian) noise, respectively.**

***Index Terms*—Bayesian Hierarchical Linear Model, Bayesian Cramér-Rao bound, sparse outlier noise, dense noise, Random Matrix Theory**

I. Introduction

In the context of robust data modeling [1], the measurement vector may be corrupted by noise outliers. This class of noise is sometimes referred to as sparse noise and is described by a distribution with heavy-tails [2]–[7]. Conversely, we usually call dense a noise that does not share this property and the most popular prior is probably Gaussian noise. Depending on the application context, outliers may be identified, *e.g.*, as incomplete data [8] or corrupted information.

A robust and relevant noise prior which is able to take into account of outliers is the Student's t-distribution for low degrees of freedom [9]–[12]. In addition, dense noise can also be encompassed thanks to the Student's t-distribution prior for an infinite degree of freedom. A convenient framework to deal with a wide class of distributions is well known under the name of hierarchical Bayesian modeling. The Bayesian Hierarchical Linear Model (BHLM) with hierarchical noise prior is used in a wide range of applications, including fusion [13], anomaly detection [5] of hyperspectral images, channel estimation [14], blind deconvolution [15], segmentation of astronomical times series [16], *etc.*

In this work, we adopt this hierarchical prior framework due to its flexibility and ability to modelize a wide class of priors. More precisely, the noise vector is assumed to follow a circular i.i.d. centered Gaussian prior with a variance defined by the inverse of an unknown random hyper-parameter. In addition, if this hyper-parameter is Gamma distributed [17,18], then the marginalized joint pdf over the hyper-parameter is the Student's t-distribution.

The Van Trees' Bayesian Cramér-Rao bound (BCRB) [19] is a standard and fundamental lower bound on the mean-square-error (MSE) of any estimator. The aim of this work is to derive and analyze the BCRB of the amplitude parameters and the noise hyper-parameter (i) for the considered noise prior and (ii) using some powerful results from the Random Matrix Theory (RMT) framework [20]–[22]. Regarding reference [23], the proposed work is original in the sense that the noise prior is different and the asymptotic regime is assumed. Finally, note that reference [24] tackles a similar problem but does not assume the asymptotic context.

We use the following notation. Scalars, vectors and matrices are denoted by italic lower-case, boldface lower-case and boldface upper-case symbols, respectively. The symbol $\mathrm{Tr}[\cdot]$ stands for the trace operator. The $K \times K$ identity matrix is denoted by $\mathbf{I}_K$ and $\mathbf{0}_{K\times 1}$ is the $K \times 1$ vector, filled with zeros. The probability density function (pdf) of a given random variable u is given by $p(u)$. The symbol $\mathcal{N}(\cdot,\cdot)$ refers to the Gaussian distribution, parametrized by mean and covariance matrix, $\mathcal{G}(\cdot,\cdot)$ is the Gamma distribution, described by shape and rate (inverse scale) parameters, while $\mathcal{IG}(\cdot,\cdot)$ is the inverse-Gamma distribution. If we have $u \sim \mathcal{G}(a,b)$ then $p(u|a,b) = \frac{b^a u^{a-1} e^{-bu}}{\Gamma(a)}$, where $\Gamma(\cdot)$ is the Gamma function. And if $u \sim \mathcal{IG}(a,b)$, then $p(u|a,b) = \frac{b^a u^{-a-1} e^{-\frac{b}{u}}}{\Gamma(a)}$. The non-standardized Student's t-distribution is defined by three parameters, through the pdf $p(u|\mu,\sigma^2,\nu,) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\sigma^2}}(1+\frac{1}{\nu}\frac{(u-\mu)^2}{\sigma^2})^{-\frac{\nu+1}{2}}$ such that $u \sim \mathcal{S}(\mu,\sigma^2,\nu)$. As regards the bivariate Normal-Gamma distribution, if we have $(u,w) \sim \mathrm{NormalGamma}(\mu,\lambda,\mathrm{a},\mathrm{b})$, then $p(u,w|\mu,\lambda,a,b) = \frac{b^a\sqrt{\lambda}}{\Gamma(a)\sqrt{2\pi}} w^{a-\frac{1}{2}} e^{-bw} e^{-\frac{\lambda w(u-\mu)^2}{2}}$. Finally, the symbol $\xrightarrow{a.s.}$ denotes almost sure convergence, $O(\cdot)$ is the big O notation, $\lambda_i(\cdot)$ is the i-th eigenvalue of the con-

This work was supported by the following projects: MAGELLAN (ANR-14-CE23-0004-01), MI-CNRS TITAN and ICode blanc.

sidered matrix and the symbol $\mathbb{E}_{\mathbf{u}|\mathbf{w}}$ refers to the expectation with respect to $p(\mathbf{u}|\mathbf{w})$.

II. Bayesian linear model corrupted by noise outliers

A. Definition of the random model

Let $\mathbf{y}$ be the $N \times 1$ vector of measurements. The BHLM is defined by

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{1}$$

where each element $[\mathbf{A}]_{i,j}$ of the $N \times K$ matrix $\mathbf{A}$, with $K < N$, is drawn i.i.d. as a single realization of a sub-Gaussian distribution with zero-mean and variance $1/N$ [22,25]. The unknown amplitude vector is given by

$$\mathbf{x} = [x_1, \ldots, x_K]^T \sim \mathcal{N}(\mathbf{0}_{K\times 1}, \sigma_x^2 \mathbf{I}_K), \tag{2}$$

where σ_x^2 is the known amplitude variance. In addition, the measurements are contaminated by a noise vector $\mathbf{e}$ which is assumed statistically independent from $\mathbf{x}$.

B. Hierarchical Normal-Gamma representation

The i-th noise sample is assumed to be circular centered i.i.d. Gaussian according to

$$e_i|\gamma \sim \mathcal{N}\left(0, \frac{\sigma^2}{\gamma}\right), \tag{3}$$

where $\frac{\gamma}{\sigma^2}$ is usually called the noise precision, γ is an unknown hyper-parameter and σ^2 is a fixed scale parameter.

If the hyper-parameter is Gamma distributed according to

$$\gamma \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \tag{4}$$

where ν is the number of degrees of freedom, the joint distribution of (e_i, γ) follows a Normal-Gamma distribution [26] such as

$$(e_i, \gamma) \sim \text{NormalGamma}\left(0, \frac{1}{\sigma^2}, \frac{\nu}{2}, \frac{\nu}{2}\right). \tag{5}$$

The marginal distribution of the joint pdf over the hyper-parameter γ leads to a non-standardized Student's t-distribution, given by [11,27]

$$\mathcal{S}(e_i|0, \sigma^2, \nu) = \int_0^\infty \mathcal{N}\left(e_i|0, \frac{\sigma^2}{\gamma}\right) \mathcal{G}\left(\gamma|\frac{\nu}{2}, \frac{\nu}{2}\right) \mathrm{d}\gamma, \tag{6}$$

such that $e_i \sim \mathcal{S}(0, \sigma^2, \nu)$.

As $\nu \to \infty$, the distribution tends to a Gaussian with zero-mean and variance σ^2, while it becomes more heavy-tailed when ν is small [12,28]. With (3) and (4), and knowing that $\frac{1}{\gamma} \sim \mathcal{IG}(\frac{\nu}{2}, \frac{\nu}{2})$, we notice that the variance, noted σ_e^2 of each noise entry of $\mathbf{e}$, is given by the following expression

$$\sigma_e^2 = \mathbb{E}_\gamma \mathbb{E}_{e_i|\gamma}\left\{e_i^2\right\} = \sigma^2 \mathbb{E}_\gamma\left\{\frac{1}{\gamma}\right\} = \sigma^2 \frac{\nu}{\nu - 2}, \tag{7}$$

in which $\nu > 2$.

III. BCRB for Student's t-distribution

The vector of unknown parameters, denoted by $\boldsymbol{\theta}$, encompasses the amplitude vector and the noise hyper-parameter, *i.e.*

$$\boldsymbol{\theta} = [\mathbf{x}^T, \gamma]^T. \tag{8}$$

Given an independence assumption between $\mathbf{x}$ and γ, the joint pdf $p(\mathbf{y}, \boldsymbol{\theta})$ can be decomposed as

$$p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})p(\mathbf{x})p(\gamma). \tag{9}$$

Let us note $\hat{\boldsymbol{\theta}}$ an estimator of the unknown vector $\boldsymbol{\theta}$. Then, the mean square error (MSE), directly linked to the error covariance matrix, verifies the following inequality

$$\text{MSE}(\boldsymbol{\theta}) = \text{Tr}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\theta}}\left\{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T\right\}\right] \geq \text{Tr}\left[\mathbf{C}\right], \tag{10}$$

where $\mathbf{C}$ is the $(K+1) \times (K+1)$ BCRB matrix defined as the inverse of the Bayesian information matrix (BIM) $\mathbf{J}$. We can show that the BIM has a block-diagonal structure due to the independence between parameters. Thus, we write

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_{\mathbf{x},\mathbf{x}} & \mathbf{0}_{K\times 1} \\ \mathbf{0}_{1\times K} & \mathrm{J}_{\gamma,\gamma} \end{bmatrix}. \tag{11}$$

We assume an identifiable BHLM model so that, under weak regularity conditions [19], the BIM is given by

$$\mathbf{J} = \mathbb{E}_{\boldsymbol{\theta}}\left\{\mathbf{J}_D^{(\boldsymbol{\theta},\boldsymbol{\theta})}\right\} + \mathbf{J}_P^{(\boldsymbol{\theta},\boldsymbol{\theta})} + \mathbf{J}_{HP}^{(\boldsymbol{\theta},\boldsymbol{\theta})}, \tag{12}$$

in which

$$[\mathbf{J}_D^{(\boldsymbol{\theta},\boldsymbol{\theta})}]_{i,j} = \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}}\left\{-\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right\}, \tag{13}$$

$$[\mathbf{J}_P^{(\boldsymbol{\theta},\boldsymbol{\theta})}]_{i,j} = \mathbb{E}_{\mathbf{x}}\left\{-\frac{\partial^2 \log p(\mathbf{x})}{\partial\theta_i \partial\theta_j}\right\}, \tag{14}$$

$$[\mathbf{J}_{HP}^{(\boldsymbol{\theta},\boldsymbol{\theta})}]_{i,j} = \mathbb{E}_{\gamma}\left\{-\frac{\partial^2 \log p(\gamma)}{\partial\theta_i \partial\theta_j}\right\} \tag{15}$$

for $(i,j) \in \{1, \ldots, K+1\}^2$, and where $\mathbf{J}_D^{(\boldsymbol{\theta},\boldsymbol{\theta})}$ is the Fisher information matrix (FIM) on $\boldsymbol{\theta}$, $\mathbf{J}_P^{(\boldsymbol{\theta},\boldsymbol{\theta})}$ is the prior part of the BIM and $\mathbf{J}_{HP}^{(\boldsymbol{\theta},\boldsymbol{\theta})}$ is the hyper-prior part.

Correspondingly, we have

$$\mathbf{C} = \mathbf{J}^{-1} = \begin{bmatrix} \mathbf{C}_{\mathbf{x},\mathbf{x}} & \mathbf{0}_{K\times 1} \\ \mathbf{0}_{1\times K} & \mathrm{C}_{\gamma,\gamma} \end{bmatrix}. \tag{16}$$

Conditionally to $\boldsymbol{\theta}$, the observation vector $\mathbf{y}$ has the following Gaussian distribution

$$\mathbf{y}|\boldsymbol{\theta} \sim \mathcal{N}\Big(\boldsymbol{\mu}, \mathbf{R}\Big), \tag{17}$$

where $\boldsymbol{\mu} = \mathbf{A}\mathbf{x}$ and $\mathbf{R} = \frac{(\nu-2)\sigma_e^2}{\nu\gamma}\mathbf{I}_N$. In what follows, we directly make use of the Slepian-Bangs formula [29, p. 378]

$$[\mathbf{J}_D^{(\boldsymbol{\theta},\boldsymbol{\theta})}]_{i,j} = \left(\frac{\partial\boldsymbol{\mu}}{\partial\theta_i}\right)^T \mathbf{R}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_j} + \frac{1}{2}\text{Tr}\left[\frac{\mathbf{R}}{\partial\theta_i}\mathbf{R}^{-1}\frac{\mathbf{R}}{\partial\theta_j}\mathbf{R}^{-1}\right]. \tag{18}$$

This leads to

$$\mathbf{J}_D^{(\mathbf{x},\mathbf{x})} = \frac{\nu\gamma}{(\nu-2)\sigma_e^2}\mathbf{A}^T\mathbf{A}. \tag{19}$$

Using the fact that $\mathbf{R}^{-1} = \frac{\gamma}{\sigma^2}\mathbf{I}_N$, we obtain

$$J_D^{(\gamma,\gamma)} = \frac{\sigma^4}{2\gamma^4}\mathrm{Tr}\Big[\mathbf{R}^{-2}\Big] = \frac{N}{2\gamma^2}. \quad (20)$$

According to (2) and considering independent amplitudes, we have

$$-\log p(\mathbf{x}) = \sum_{i=1}^{K}\left(\frac{1}{2}\log(2\pi\sigma_x^2) + \frac{x_i^2}{2\sigma_x^2}\right). \quad (21)$$

Consequently,

$$\mathbf{J}_P^{(\mathbf{x},\mathbf{x})} = \frac{1}{\sigma_x^2}\mathbf{I}_K. \quad (22)$$

The BIM $\mathbf{J}$ is therefore composed of the following terms:

$$\mathbf{J}_{\mathbf{x},\mathbf{x}} = \mathbb{E}_\gamma\left\{\mathbf{J}_D^{(\mathbf{x},\mathbf{x})}\right\} + \mathbf{J}_P^{(\mathbf{x},\mathbf{x})}, \quad (23)$$

$$\mathrm{J}_{\gamma,\gamma} = \mathbb{E}_\gamma\left\{\mathrm{J}_D^{(\gamma,\gamma)}\right\} + J_{HP}^{(\gamma,\gamma)}. \quad (24)$$

The hyper-prior part of the BIM is given by

$$J_{HP}^{(\gamma,\gamma)} = \mathbb{E}_\gamma\left\{-\frac{\partial^2\log p(\gamma)}{\partial\gamma^2}\right\} = \frac{\nu-2}{2}\mathbb{E}_\gamma\left\{\frac{1}{\gamma^2}\right\}. \quad (25)$$

The second-order moment of an inverse-Gamma distributed random variable is given by

$$\mathbb{E}_\gamma\left\{\frac{1}{\gamma^2}\right\} = \frac{\nu^2}{(\nu-2)(\nu-4)}, \quad (26)$$

where $\nu > 4$. This finally leads to

$$\mathrm{J}_{\gamma,\gamma} = \frac{N\nu^2}{2(\nu-2)(\nu-4)} + \frac{\nu^2}{2(\nu-4)}. \quad (27)$$

Inverting the BIM, we obtain the BCRB for the amplitude parameters

$$\mathrm{BCRB}(\mathbf{x}) = \frac{\mathrm{Tr}\left[\mathbf{C}_{\mathbf{x},\mathbf{x}}\right]}{K} \quad \text{with } \mathbf{C}_{\mathbf{x},\mathbf{x}} = \sigma_x^2\left(r\mathbf{A}^T\mathbf{A} + \mathbf{I}_K\right)^{-1}, \quad (28)$$

where $r = \mathrm{SNR}\frac{\nu}{\nu-2}$ with $\mathrm{SNR} = \frac{\sigma_x^2}{\sigma_e^2}$ (signal-to-noise ratio).

IV. BCRB IN THE ASYMPTOTIC FRAMEWORK

Using (27), the asymptotic BCRB for the hyper-parameter γ is given by

$$\mathrm{C}_{\gamma,\gamma} = \frac{1}{\mathrm{J}_{\gamma,\gamma}} \overset{N\to\infty}{\longrightarrow} \mathrm{C}_{\gamma,\gamma}^{\infty}, \quad (29)$$

where

$$\mathrm{C}_{\gamma,\gamma}^{\infty} = \frac{2(\nu-4)}{\nu^2}. \quad (30)$$

A. RMT framework

In this section, we consider the context of large random matrices, *i.e.*, for $K, N \to \infty$ with $\frac{K}{N} \to \beta \in (0,1)$. The derived BCRB in this context is the asymptotic normalized BCRB defined by

$$\mathrm{BCRB}(\mathbf{x}) \overset{a.s.}{\to} \mathrm{BCRB}^{\infty}(\mathbf{x}). \quad (31)$$

Using (28) with [21, p. 11], we obtain

$$\mathrm{BCRB}^{\infty}(\mathbf{x}) = \sigma_x^2\left(1 - \frac{f(r,\beta)}{4r\beta}\right) \quad (32)$$

and $f(r,\beta) = \left(\sqrt{r(1+\sqrt{\beta})^2+1} - \sqrt{r(1-\sqrt{\beta})^2+1}\right)^2$.

B. Limit analytical expressions

- For $\beta \ll 1$, *i.e.*, $K \ll N$, after some manipulations and discarding the terms of order superior or equal to $O(\beta^2)$, we obtain

$$f(r,\beta) \approx \frac{4\beta r^2}{r+1}. \quad (33)$$

  Therefore, an asymptotic analytical expression of the BCRB, in the RMT framework, is given by

$$\mathrm{BCRB}^{\infty}(\mathbf{x}) \approx \frac{\sigma_x^2}{r+1} = \frac{(\nu-2)\sigma_x^2}{\nu(1+\mathrm{SNR})-2}. \quad (34)$$

- For small r, also meaning small SNR, according to the Neumann series expansion [30], we have $\left(r\mathbf{A}^T\mathbf{A}+\mathbf{I}_K\right)^{-1} \approx \mathbf{I}_K - r\mathbf{A}^T\mathbf{A}$ if the maximal eigenvalue $\lambda_{\max}(r\mathbf{A}^T\mathbf{A}) < 1$. Observe that $r\lambda_{\max}(\mathbf{A}^T\mathbf{A}) \overset{a.s.}{\to} r(1+\sqrt{\beta})^2$ [20]–[22]. In addition, if SNR is sufficiently small with respect to $(\nu-2)/(4\nu)$ then

$$\mathrm{BCRB}(\mathbf{x}) \approx \frac{\sigma_x^2}{K}\left(\mathrm{Tr}\left[\mathbf{I}_K\right] - r\mathrm{Tr}\left[\mathbf{A}^T\mathbf{A}\right]\right) \overset{a.s.}{\to} \sigma_x^2(1-r) = \frac{\sigma_x^2}{\nu-2}(\nu-2-\nu\mathrm{SNR}). \quad (35)$$

- For large r, also meaning large SNR, we have

$$\mathrm{BCRB}(\mathbf{x}) \approx \frac{\sigma_x^2}{rK}\left(\mathrm{Tr}\left[\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\right] - \frac{1}{r}\mathrm{Tr}\left[\left(\mathbf{A}^T\mathbf{A}\right)^{-2}\right]\right) \overset{a.s.}{\to} \frac{\sigma_x^2}{r}\left(\frac{1}{1-\beta} - \frac{1}{r}\frac{1}{(1-\beta)^3}\right) = \frac{(\nu-2)\sigma_x^2}{\nu\mathrm{SNR}(1-\beta)}\left(1 - \frac{\nu-2}{\nu\mathrm{SNR}(1-\beta)^2}\right), \quad (36)$$

  since [20]–[22]

$$\frac{1}{K}\mathrm{Tr}\left[\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\right] \overset{a.s.}{\to} \frac{1}{1-\beta}, \quad (37)$$

$$\frac{1}{K}\mathrm{Tr}\left[\left(\mathbf{A}^T\mathbf{A}\right)^{-2}\right] \overset{a.s.}{\to} \frac{1}{(1-\beta)^3}. \quad (38)$$

C. Comparison between two models with a target common SNR

We consider two different models:

$$(M_0): \ \mathbf{y}_0 = \mathbf{A}\mathbf{x} + \mathbf{e}_0 \ \text{ with } e_{i_0} \sim \mathcal{S}(0,\sigma_0^2,\nu_0), \quad (39)$$

$$(M_1): \ \mathbf{y}_1 = \mathbf{A}\mathbf{x} + \mathbf{e}_1 \ \text{ with } e_{i_1} \sim \mathcal{S}(0,\sigma_1^2,\nu_1). \quad (40)$$

Model (M_0) is the reference model and model (M_1) is the alternative one. According to (32), the asymptotic normalized BCRB for the k-th model with $k \in \{0,1\}$ is defined by

$$\mathrm{BCRB}_k^{\infty}(\mathbf{x}) = \sigma_x^2\left(1 - \frac{f(r_k,\beta)}{4r_k\beta}\right) \quad (41)$$

where $r_k = \mathrm{SNR}_k\frac{\nu_k}{\nu_k-2}$ with $\mathrm{SNR}_k = \frac{\sigma_x^2}{\sigma_{e_k}^2}$. A fair methodology to compare the bounds $\mathrm{BCRB}_0(\mathbf{x})$ and $\mathrm{BCRB}_1(\mathbf{x})$ is to impose a common target SNR for the models (M_0) and (M_1), *i.e.*, $\mathrm{SNR}_0 = \mathrm{SNR}_1$. A simple derivation shows that to reach

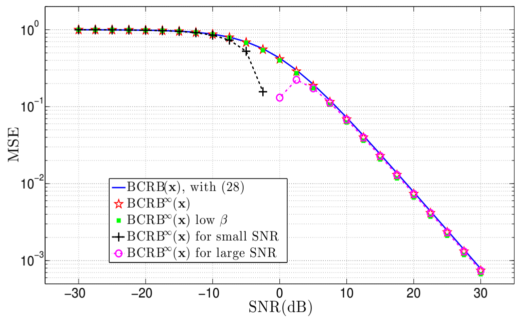

Fig. 1. $\mathrm{BCRB}(\mathbf{x})$ as a function of SNR in dB with specific limit approximations, in the RMT framework.

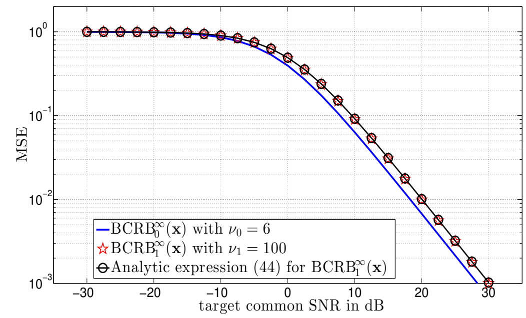

Fig. 2. Asymptotic normalized BCRBs for models (M_0) and (M_1) *vs.* a common SNR

the target SNR, we must have $r_1 = \frac{\nu_1(\nu_0-2)}{\nu_0(\nu_1-2)} r_0$. Specifically, the corresponding BCRBs are the following ones:

$$\mathrm{BCRB}_0^\infty(\mathbf{x}) = \sigma_x^2 \left(1 - \frac{f(r_0,\beta)}{4r_0\beta}\right), \tag{42}$$

$$\mathrm{BCRB}_1^\infty(\mathbf{x}) = \sigma_x^2 \left(1 - \frac{\nu_0(\nu_1-2) f(\frac{\nu_1(\nu_0-2)}{\nu_0(\nu_1-2)} r_0, \beta)}{4\nu_1(\nu_0-2) r_0 \beta}\right). \tag{43}$$

Recall that the Student's t-distribution is well known to promote noise outliers thanks to its heavy-tails property. Conversely, the Gaussian distribution does not share this behavior. So, an interesting scenario arises when $\nu_1 \to \infty$. In this case, the Student's t-distribution converges to the Gaussian one [10] and (43) tends to

$$\mathrm{BCRB}_1^\infty(\mathbf{x}) \overset{\nu_1 \to \infty}{=} \sigma_x^2 \left(1 - \frac{\nu_0 f\left(\frac{\nu_0-2}{\nu_0} r_0, \beta\right)}{4(\nu_0-2) r_0 \beta}\right). \tag{44}$$

D. Numerical simulations

In the following simulations, we consider $N = 100$ and $K = 10$ so that $\beta \ll 1$. The amplitude variance σ_x^2 is fixed to 1. In Fig. 1, we plot the BCRB of the amplitude vector $\mathbf{x}$, as defined by equations (28) and (32) (asymptotic expression), (34) (small β), (35) (small SNR) and (36) (large SNR), as a function of the SNR in dB. We choose $\nu = 6$.

We notice that $\mathrm{BCRB}(\mathbf{x})$ coincides precisely with its asymptotic expression in (32). Thus, the RMT framework predicts precisely the behavior of the BCRB of the amplitude as $K, N \to \infty$ with $\frac{K}{N} \to \beta$ and allows us to obtain a closed-form expression. Such limit remains correct even for values of N and K that are relatively not quite large. The expression of the BCRB obtained with (34) is a good approximation since here, we have $\beta = 0.1 \ll 1$. Finally, we notice that the curves obtained for low and high SNR approximate very well the BCRB of the amplitude, asymptotically.

In Fig. 2, as exposed in section IV-C, we consider two different models, with a different value for the number of degrees of freedom ν. We notice that a lower performance bound is achieved with $\nu_0 = 6$, especially in the low noise regime, than with $\nu_1 = 100$. Furthermore, the approximation

TABLE I
$\mathrm{C}_{\gamma,\gamma}^\infty$ AS A FUNCTION OF ν.

ν	10	30	50	70	100
$\mathrm{C}_{\gamma,\gamma}^\infty$ with (30)	0.1	0.06	0.04	0.03	0.02

in (44) is correct, since ν_1 has a large value. A low value for the number of degrees of freedom is well-adapted for the modelization of sparse (outlier) noise, characterized by a heavy-tailed distribution [31,32]. This large level in heavy-tailedness leads to robustness [1,33,34] while a Gaussian noise model (large degree of freedom) corresponds to a dense noise type. Thus, we can hope to achieve better estimation performances if we consider a model, which promotes sparsity and the presence of outliers in data.

Finally, in Table I, we represent the evolution of $\mathrm{C}_{\gamma,\gamma}^\infty$, given in (30), as a function of the number of degrees of freedom ν. We notice that a lower BCRB, for the noise hyper-parameter, is achieved as ν increases, *i.e.*, when the marginal distribution of the noise tends to a Gaussian.

V. Conclusion

This work discusses fundamental Bayesian lower bounds for multi-parameter robust estimation. More precisely, we consider a Bayesian linear model corrupted by a sparse noise following a Student's t-distribution. This class of prior can efficiently modelize outliers. Using the hierarchical Normal-Gamma representation of the Student's t-distribution, the Van Trees' Bayesian lower Bound (BCRB) is derived for unknown amplitude parameters and the noise hyper-parameter in an asymptotic context. By asymptotic, it means that the number of measurements and the number of unknown parameters grow to infinity at a finite rate. Consequently, closed-form expressions of the BCRB are obtained using some powerful results from the large Random Matrix Theory (RMT). Finally, a framework is provided to fairly compare two models corrupted by noises with different number of degrees of freedom for a fixed common target SNR. We recall that a small degree of freedom promotes outliers in the sense that the noise prior has heavy-tails. For the amplitude, a lower performance bound is achieved when the number of degrees of freedom is small.

REFERENCES

[1] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 61–80, 2012.

[2] K. Mitra, A. Veeraraghavan, and R. Chellappa, "Robust RVM regression using sparse outlier model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, 2010, pp. 1887–1894.

[3] ——, "Robust regression using sparse learning for high dimensional parameter estimation problems," in *IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, Dallas, TX, 2010, pp. 3846–3849.

[4] P. Zhuang, W. Wang, D. Zeng, and X. Ding, "Robust mixed noise removal with non-parametric Bayesian sparse outlier model," in *16th International Workshop on Multimedia Signal Processing (MMSP)*, Jakarta, Indonesia, 2014, pp. 1–5.

[5] G. E. Newstadt, A. O. Hero, and J. Simmons, "Robust spectral unmixing for anomaly detection." in *IEEE Workshop on Statistical Signal Processing (SSP)*, Gold Coast, VIC, 2014, pp. 109–112.

[6] M. Sundin, S. Chatterjee, and M. Jansson, "Combined modeling of sparse and dense noise improves Bayesian RVM," in *22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, 2014, pp. 1841–1845.

[7] ——, "Bayesian learning for robust Principal Component Analysis," in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, France, 2015, pp. 2361–2365.

[8] J. Luttinen, A. Ilin, and J. Karhunen, "Bayesian robust PCA of incomplete data," *Neural Processing Letters*, vol. 36, no. 2, pp. 189–202, 2012.

[9] D. Peel and G. J. McLachlan, "Robust mixture modelling using the t distribution," *Statistics and computing*, vol. 10, no. 4, pp. 339–348, 2000.

[10] S. Kotz and S. Nadarajah, *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

[11] J. Christmas, "Bayesian spectral analysis with Student-t noise," *IEEE Transactions on Signal Processing*, vol. 62, no. 11, pp. 2871–2878, 2014.

[12] H. Zhang, Q. M. J. Wu, T. M. Nguyen, and X. Sun, "Synthetic aperture radar image segmentation by modified Student's t-mixture model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 7, pp. 4391–4403, 2014.

[13] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Bayesian fusion of hyperspectral and multispectral images," in *IEEE Int. Conf. on Acoust., Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 3176–3180.

[14] N. L. Pedersen, C. N. Manchón, D. Shutin, and B. H. Fleury, "Application of Bayesian hierarchical prior modeling to sparse channel estimation," in *IEEE International Conference on Communications (ICC)*, Ottawa, ON, 2012, pp. 3487–3492.

[15] G. Kail, J.-Y. Tourneret, F. Hlawatsch, and N. Dobigeon, "Blind deconvolution of sparse pulse sequences under a minimum distance constraint: A partially collapsed Gibbs sampler method," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2727–2743, 2012.

[16] N. Dobigeon, J.-Y. Tourneret, and J. D. Scargle, "Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model," *IEEE Transactions on Signal Processing*, vol. 55, no. 2, pp. 414–423, 2007.

[17] A. Gelman, "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)," *Bayesian Analysis*, vol. 1, no. 3, pp. 515–534, 2006.

[18] J. Dahlin, F. Lindsten, T. B. Schön, and A. Wills, "Hierarchical Bayesian ARX models for robust inference," in *16th IFAC Symposium on System Identification (SYSID)*, Brussels, Belgium, 2012, pp. 131–136.

[19] H. L. Van Trees and K. L. Bell, *Bayesian bounds for parameter estimation and nonlinear filtering/tracking*. New York: Wiley-IEEE Press, 2007.

[20] J. W. Silverstein and Z. Bai, "On the empirical distribution of eigenvalues of a class of large dimensional random matrices," *Journal of Multivariate analysis*, vol. 54, no. 2, pp. 175–192, 1995.

[21] A. M. Tulino and S. Verdú, *Random matrix theory and wireless communications*. Foundations and Trends in Communications and Information Theory. Now Publishers Inc., 2004, vol. 1, no. 1.

[22] R. Couillet and M. Debbah, *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

[23] M. N. El Korso, R. Boyer, P. Larzabal, and B.-H. Fleury, "Estimation performance for the Bayesian hierarchical linear model," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 488–492, 2016.

[24] R. Prasad and C. R. Murthy, "Cramér-Rao-type bounds for sparse Bayesian learning," *IEEE Transactions on Signal Processing*, vol. 61, no. 3, pp. 622–632, 2013.

[25] V. V. Buldygin and Y. V. Kozachenko, *Metric characterization of random variables and random processes*. American Mathematical Soc., 2000, vol. 188.

[26] J. M. Bernardo and A. F. M. Smith, *Bayesian theory*. New York: J. Wiley, 1994.

[27] M. Svensén and C. M. Bishop, "Robust Bayesian mixture modelling," *Neurocomputing*, vol. 64, pp. 235–252, 2005.

[28] G. Sfikas, C. Nikou, and N. Galatsanos, "Robust image segmentation with mixtures of Student's t-distributions," in *International Conference on Image Processing (ICIP)*, vol. 1, Sant Antonio, TX, 2007, pp. I–273 – I–276.

[29] P. Stoica and R. L. Moses, *Spectral analysis of signals*. Pearson Prentice Hall, Upper Saddle River, NJ, 2005.

[30] G. B. Arfken and H. J. Weber, *Mathematical methods for physicists, sixth edition*. Academic press, 2005.

[31] G. Tzagkarakis and P. Tsakalides, "Bayesian compressed sensing of a highly impulsive signal in heavy-tailed noise using a multivariate Cauchy prior," in *17th European Signal Processing Conference*, Glasgow, Scotland, 2009, pp. 2293–2297.

[32] A. Amini, M. Unser, and F. Marvasti, "Compressibility of deterministic and random infinite sequences," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5193–5201, 2011.

[33] 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.

[34] N. Delannay, C. Archambeau, and M. Verleysen, "Improving the robustness to outliers of mixtures of probabilistic PCAs," *Advances in Knowledge Discovery and Data Mining*, vol. 5012, pp. 527–535, 2008.