Towards hyperparameter optimization of sparse bayesian learning based on Stein's unbiased risk estimator

Fangqing Xiao, Dirk Slock

Towards Hyperparameter Optimizing of Sparse Bayesian Learning Based on Stein's Unbiased Risk Estimator

Fangqing Xiao, Dirk Slock
Communication Systems Department
Eurecom, France
Email: {fangqing.xiao, dirk.slock}@eurecom.fr

***Abstract*—Sparse Bayesian Learning (SBL) serves as a sparse signal recovery algorithm in compressed sensing, necessitating estimation of several hyperparameters. These can be optimized using Stein's Unbiased Risk Estimator (SURE), asymptotically equivalent to minimizing Mean Squared Error (MSE). In this paper, we analyze minimum MSE by optimizing hyperparameters via MSE. Additionally, we explore the potential of extending SBL's Gaussian prior to a generalized Gaussian prior by analyzing the Laplacian and uniform priors, which represent two special cases of the generalized Gaussian prior. Through simulation experiments, we observe that the Gaussian prior outperforms others for underestimated and deterministic signals, accurately recovering 0 with optimal hyperparameters optimized via MSE. For non-zero cases, the uniform prior demonstrates superior performance. Conversely, the Laplacian prior consistently performs worse than the other two cases, with its minimum MSE equivalent to the variance of extrinsic.**

I. Introduction

Sparse signal reconstruction (SSR) and compressed sensing (CS) have attracted considerable attention in recent years across diverse fields [1], [2], [3]. They can be formulated as:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{v}, \tag{1}$$

where $\boldsymbol{y}$ represents the observations or data, while $\boldsymbol{A}$ is referred to as the measurement or sensing matrix, initially known and of dimension $M \times N$ with $M < N$. The M-dimensional sparse signal is denoted by $\boldsymbol{x}$, and $\boldsymbol{v}$ represents the additive noise. In the case of exact sparsity, the unknown $\boldsymbol{x}$ contains only K non-zero entries, where $K << N$. The noise $\boldsymbol{v}$ is assumed to follow a white Gaussian distribution, $\boldsymbol{v} \sim \mathcal{N}(0, \gamma\mathbf{I})$, with variance γ. While $\boldsymbol{x}$ is deterministic yet sparse, directly estimating $\boldsymbol{x}$ poses an NP-hard problem.

To address this issue, the SBL algorithm was initially proposed for SSR by [4], [5]. Within a Bayesian framework, the goal is to compute the posterior distribution of the parameters $\boldsymbol{x}$ given observations (data) and prior knowledge. In SBL, the unknown deterministic parameters $\boldsymbol{x}$ are modeled as decorrelated zero-mean Gaussian, $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{P})$. The estimation of the hyperparameters $\boldsymbol{P}$ and the sparse signal $\boldsymbol{x}$ is performed jointly. One approach involves estimating the hyperparameters first using evidence maximization, known as the Type II Maximum Likelihood (ML) method [6], which is also an instance of Empirical Bayes (EB) estimation. Additionally, In [7] the authors propose a Fast Marginalized ML (FMML) technique by alternating likelihood maximization with respect to the hyperparameters. In our previous work [8], we introduced SURE-SBL, where hyperparameter optimization (not estimation) is based on Stein's Unbiased Risk Estimator (SURE) [9]. The ultimate performance criterion typically revolves around the Mean Squared Error (MSE) of the sparse parameters or the resultant signal model. However, directly analyzing the results optimized by SURE appears impractical due to the complexity introduced by $\boldsymbol{v}$, even with the aid of large system analysis (LSA). Since SURE is an asymptotically unbiased estimator based on MSE, another reasonable approach is to analyze the hyperparameters obtained based on optimizing MSE and examine the minimum MSE with respect to (w.r.t.) the input $\boldsymbol{x}$ and the output $\mathbf{z} = \boldsymbol{A}\boldsymbol{x}$.

Moreover, SBL can be considered a class of algorithms aiming to improve estimation by assuming an unknown distribution for parameters, which are jointly estimated for deterministic data. This resembles a regularization term within Maximum A Posteriori (MAP), albeit for the overarching goal of minimizing MSE, the MMSE estimator would have been preferable. In SBL, the assumed distribution is typically Gaussian, a subclass of the generalized Gaussian distribution (GGD) [10]. An intriguing inquiry arises when considering combining sub-Gaussian and super-Gaussian distributions. It is worth exploring which GGD yields better MSE. However, obtaining overall posterior distributions computationally is challenging, particularly for the GGD except the Gaussian case, due to high-dimensional integrals. Over the years, with the development of message passing algorithms [11], [12], [13], high-dimensional integrals for MMSE can be decomposed into scalar-level integrations of extrinsic Gaussian and prior distributions, which alleviates computational difficulty. In this paper, without delving into the details of calculating fixed point, we analyze the effect of Laplacian prior and uniform prior based on the same fixed extrinsics for scalar case. Even with statistical inference advancements, computing expectations regarding the noise $\boldsymbol{v}$ in calculating the MSE proves infeasible for non-Gaussian cases. Moreover, analytical solutions for hyperparameter optimization are elusive, necessitating reliance on numerical simulations. Through Monte Carlo experiments, we find that for SSR, the Gaussian prior excels

in recovering x_i when the signal is 0, with an optimized $p_i = 0$. Introducing the uniform prior minimizes MSE for non-zero signals, followed by the Gaussian prior. Conversely, the Laplacian prior yields the poorest results. Thus, for SSR and minimizing MSE, the Gaussian prior effectively balances accuracy and MSE minimization, contingent on accurately estimated hyperparameters. However, this conclusion warrants a more rigorous mathematical proof.

A. Notations

The notation $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents the Gaussian distribution function evaluated at $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\boldsymbol{A}_i$ denotes the i-th column vector of matrix $\boldsymbol{A}$. P_{ij} denotes $\boldsymbol{P}(i, j)$.

II. Generalized Gaussian Distribution

The probability density function (PDF) of the Generalized Gaussian Distribution (GGD) with zero mean is denoted as:

$$p(x; \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left[-\left(\frac{|x|}{\alpha}\right)^{\beta}\right], \alpha > 0, \beta > 0, \tag{2}$$

where $\Gamma(\cdot)$ denotes the Gamma function defined as $\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t}\, dt$. The GGD is characterized by two parameters: the scale parameter α and the shape parameter β. Here, α governs the width (standard deviation) of the curve, while β influences the sharpness of the GGD curve.

To elaborate further, it is evident that the GGD transitions into a Laplacian distribution when $\beta = 1$ and a standard Gaussian distribution when $\beta = 2$. When $\beta > 2$, the GGD represents a sub-Gaussian distribution with lighter tails compared to the Gaussian distribution, while for $\beta < 2$, it denotes a super-Gaussian distribution with heavier tails. As $\beta \to +\infty$, the GGD converges to a uniform distribution as follows:

$$\lim_{\beta \to +\infty} p(x; \alpha, \beta) = p(x, U(-\alpha, \alpha)) = \begin{cases} \dfrac{1}{2\alpha}, & |x| < \alpha; \\ 0, & |x| > \alpha. \end{cases} \tag{3}$$

Fig. 1 illustrates the GGDs' pdf shapes for various β values, with $\alpha = 1$. Super-Gaussian distributions, like the Laplacian distribution, tend to be sparser than standard Gaussian distributions, whereas sub-Gaussian distributions exhibit the opposite trend. As β approaches infinity, the GGD transitions into a uniform distribution, representing the extreme case of a sub-Gaussian distribution where no element holds greater significance.

III. Sparse Bayesian Learning (GGD with $\beta = 2$)

For estimating $\boldsymbol{x}$, SBL assumes that each element x_i of $\boldsymbol{x}$ follows an Automatic Relevance Prior (ARP). For normal SBL, ARP is modeled by a Gaussian distribution with zero mean and variance p_i, represented as:

$$p(x_i; p_i) = \mathcal{N}(x_i; 0, p_i), i = 1, \cdots, N; \tag{4}$$

where p_i is an unknown Gaussian variance optimized through the SBL algorithm. Typically, p_i tends towards zero (without noise) or approaches it (with noise). When $p_i = 0$, the corresponding estimated x_i is set to zero, thereby influencing solution sparsity significantly. Thus, optimizing p_i is paramount in SBL. To optimize each p_i, we aim to minimize MSE concerning $\boldsymbol{x}$ and $\mathbf{z}$. We define $\text{MSE}_{\boldsymbol{x}}$ and $\text{MSE}_{\mathbf{z}}$ as:

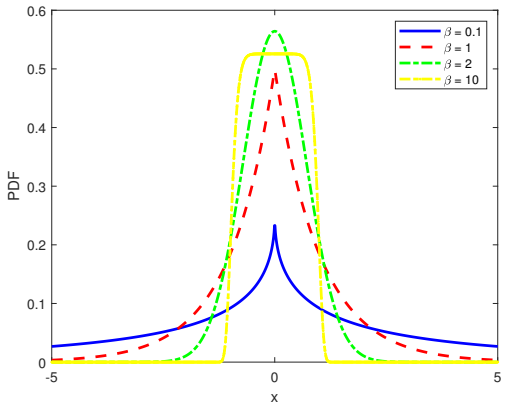

Fig. 1. Generalized Gaussian Probability Density Function With $\alpha = 1$

$$\text{MSE}_{\boldsymbol{x}} = \text{E}\|\widehat{\boldsymbol{x}}(\boldsymbol{P}) - \boldsymbol{x}\|^2; \tag{5a}$$

$$\text{MSE}_{\mathbf{z}} = \text{E}\|\widehat{\mathbf{z}}(\boldsymbol{P}) - \mathbf{z}\|^2 = \text{E}\|\widehat{\mathbf{z}}(\boldsymbol{P}) - \boldsymbol{A}\boldsymbol{x}\|^2, \tag{5b}$$

where E is w.r.t. $\boldsymbol{v}$ ($\boldsymbol{x}$ and $\mathbf{z}$ are treated as deterministic) and $\boldsymbol{P}$ is a diagonal matrix with $P_{ii} = p_i$.

A. Optimizing p_i by $\text{MSE}_{\boldsymbol{x}}$

By Gaussian-Markov theorem, the posterior of $\boldsymbol{x}$ is Gaussian with the pdf as:

$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x};\, \boldsymbol{P}\boldsymbol{A}^T\boldsymbol{R}^{-1}\boldsymbol{y},\, \boldsymbol{P} - \boldsymbol{P}\boldsymbol{A}^T\boldsymbol{R}^{-1}\boldsymbol{A}\boldsymbol{P}), \tag{6}$$

where $\boldsymbol{R} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T + \gamma\mathbf{I}$ is the covariance matrix of $\boldsymbol{y}$. In the context of estimating the i-th entry of the signal vector $\boldsymbol{x}$, we can follow the *Component-Wise Conditionally Unbiased (CWCU-)LMMSE* approach [14]. This approach assumes that the i-th entry of $\boldsymbol{x}$ is deterministic while the other entries are random. When considering only the i-th entry of the signal vector $\boldsymbol{x}$ to be deterministic (assume the prior variance to be $+\infty$), and treating the other entries as random variables, we can estimate the i-th entry of $\boldsymbol{x}$ and the associated error using the following equations:

$$r_i = x_i + w_i. \tag{7}$$

where r_i represents the CWCU-LMMSE estimated value, while w_i denotes a zero-mean Gaussian noise with variance ξ_i, equaling the variance of the CWCU-LMMSE estimator. Introducing the extrinsic x_i with pdf expressed as

$$p(x_i|r_i) = \mathcal{N}(x_i; r_i, \xi_i), \tag{8}$$

where r_i and ξ_i can be expressed as:

$$r_i = \frac{\boldsymbol{A}_i^T\boldsymbol{R}_{\tilde{i}}^{-1}\boldsymbol{y}}{\boldsymbol{A}_i^T\boldsymbol{R}_{\tilde{i}}^{-1}\boldsymbol{A}_i}, \quad \xi_i = (\boldsymbol{A}_i^T\boldsymbol{R}_{\tilde{i}}^{-1}\boldsymbol{A}_i)^{-1}, \tag{9}$$

where $\boldsymbol{R}_{\tilde{i}} = \boldsymbol{R} - p_i \boldsymbol{A}_i \boldsymbol{A}_i^T$. Combining the Gaussian prior information assumed in SBL to extrinsic $p(x_i|r_i)$, the posterior mean $\widehat{x}_i$ can be given as:

$$\widehat{x}_i = \frac{\int x_i \mathcal{N}(x_i; r_i, \xi_i)\mathcal{N}(x_i; 0, p_i)\mathrm{d}x_i}{\int \mathcal{N}(x_i; r_i, \xi_i)\mathcal{N}(x_i; 0, p_i)\mathrm{d}x_i} = \frac{p_i}{p_i + \xi_i} r_i. \quad (10)$$

At the fixed point, for all optimized p_j except p_i, optimizing p_i from $\mathrm{MSE}_{\boldsymbol{x}}$ and MSE_{x_i} yield the same result. For each x_i, the MSE_{x_i}, treating x_i as deterministic and replacing r_i by $x_i + v_i$, can be expressed as:

$$\mathrm{MSE}_{x_i}(p_i) = \mathrm{E}_{w_i}\|\widehat{x}_i - x_i\|^2 \quad (11a)$$

$$= \frac{p_i^2}{(p_i + \xi_i)^2}\xi_i + \frac{\xi_i^2}{(p_i + \xi_i)^2} x_i^2. \quad (11b)$$

Then the p_i can be optimized by minimizing $\mathrm{MSE}_{x_i}(p_i)$ as:

$$\hat{p}_i = \arg\min_{p_i} \mathrm{MSE}_{x_i}(p_i) = x_i^2. \quad (12)$$

Therefore, the minimum MSE (MMSE) with optimized $\hat{p}_i$ can be calculated as:

$$\mathrm{MMSE}_{x_i} = \frac{x_i^2 \xi_i}{x_i^2 + \xi_i}. \quad (13)$$

B. Optimizing p_i by $\mathrm{MSE}_{\mathbf{z}}$

Apart from optimizing p_i from MSE_{x_i}, another approach is via $\mathrm{MSE}_{\mathbf{z}}$. Let $\widehat{\mathbf{z}}$ be an MMSE estimator of $\mathbf{z}$ which can be expressed as:

$$\hat{\mathbf{z}} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T\boldsymbol{R}^{-1}\boldsymbol{y}. \quad (14)$$

Therefore the $\mathrm{MSE}_{\mathbf{z}}$ in (5b) can be represented as:

$$\mathrm{MSE}_{\mathbf{z}}(\boldsymbol{P}) = \gamma \mathrm{tr}\left\{[(\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T)^2 + \gamma \mathbf{z}\mathbf{z}^T]\boldsymbol{R}^{-2}\right\} \quad (15a)$$

$$= \gamma \mathrm{tr}\left\{[(\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T)^2 + \gamma \boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T]\boldsymbol{R}^{-2}\right\}. \quad (15b)$$

At the fixed point of $\boldsymbol{P}$, for each p_i, we can obtain:

$$\frac{\partial \mathrm{MSE}_{\mathbf{z}}(\boldsymbol{P})}{\partial p_i} = \mathrm{tr}\left[\frac{\partial \mathrm{MSE}_{\mathbf{z}}(\boldsymbol{P})}{\partial \boldsymbol{P}} \frac{\partial \boldsymbol{P}}{\partial p_i}\right] = 0. \quad (16)$$

After simple algebraic manipulation, (16) can be expressed as:

$$\gamma^2 \mathrm{tr}(\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{R}^{-2} \boldsymbol{A}(\boldsymbol{P} - \boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{A}^T\boldsymbol{R}^{-1}\boldsymbol{A}\boldsymbol{e}_i) = 0. \quad (17)$$

After simple algebraic manipulation, each solution $\hat{p}_i$ of (17) can be expressed as:

$$\hat{p}_i = \left\lfloor \frac{x_i^2 \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-2} \boldsymbol{A}_i \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i + Z_i}{\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-2} \boldsymbol{A}_i \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i + Y_i} \right\rfloor_+, \quad (18)$$

where

$$Z_i = \sum_{j\neq i}\sum_{k\neq i} x_j x_k \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-2} \boldsymbol{A}_j \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_k + \sigma_v^2 \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-3} \boldsymbol{A}_i - \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-2} \boldsymbol{A}_i;$$

$$Y_i = \sum_{j\neq i}\sum_{k\neq i} x_j x_k \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-2} \boldsymbol{A}_i \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_j \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_k - \sum_{j\neq i}\sum_{k\neq i} x_j x_k \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-2} \boldsymbol{A}_j \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_k + \sigma_v^2 (\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-2} \boldsymbol{A}_i \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-2} \boldsymbol{A}_i - \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-3} \boldsymbol{A}_i \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i).$$

An interesting point of discussion is to determine the conditions under which each optimized p_i from minimizing $\mathrm{MSE}_{\mathbf{z}}$ and $\mathrm{MSE}_{\boldsymbol{x}}$ would be the same. We observe that for $j \neq k \neq i$, in Z_i and Y_i, if

$$\sum_{j\neq i}\sum_{k\neq i,j} x_j x_k \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-2} \boldsymbol{A}_j \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_k = 0, \quad (20a)$$

$$\sum_{j\neq i}\sum_{k\neq i,j} x_j x_k \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-2} \boldsymbol{A}_i \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_j \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_k = 0, \quad (20b)$$

$$\sum_{j\neq i}\sum_{k\neq i,j} x_j x_k \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-2} \boldsymbol{A}_j \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_k = 0, \quad (20c)$$

then the each optimized $\hat{p}_i$ in (18) would equal to x_i^2. And the conditions outlined in (20) warrant further investigation through Large System Analysis.

IV. Sparse Bayesian Learning (GGD with $\beta = 1$)

We transition from using the Gaussian prior in SBL to the Laplacian prior, which corresponds to the GGD with $\beta = 1$. However, due to the high-dimensional integration involved, calculating the posterior $p(\boldsymbol{x}|\boldsymbol{y})$ directly is infeasible. Therefore, approximate methods [11], [12], [13] should be employed and the extrinsic pdf $p(x_i|r_i)$ can be approximated as a Gaussian pdf in (8). While obtaining the fixed point of the extrinsic is not the primary objective of this paper, for a fair comparison with SBL using different GGDs, we assume that the extrinsic pdf for the Laplacian prior has the same form as that of the Gaussian prior. For the sake of clarity, we define the Laplacian prior pdf by introducing $b_i = 1/\alpha_i$ as:

$$p(x_i; b_i) = \frac{b_i}{2} exp(-b_i|x_i|), b_i \geq 0. \quad (21)$$

Then optimizing b_i via $\mathrm{MSE}_{x_i}(b_i)$ can be expressed as:

$$\hat{b}_i = \arg\min_{b_i} \mathrm{MSE}_{x_i}(b_i) = \arg\min_{b_i} \mathrm{E}_{w_i}\|\widehat{x}_i(b_i) - x_i\|^2, \quad (22)$$

where $\widehat{x}_i(b_i)$ is the posterior mean which can be calculated as:

$$\widehat{x}_i(b_i) = \frac{\int x_i p(x_i; b_i)\mathcal{N}(x_i; r_i, \xi_i)dx_i}{\int p(x_i; b_i)\mathcal{N}(x_i; r_i, \xi_i)dx_i}. \quad (23a)$$

However, it is a bit complex to calculate as two times integration is needed. Therefore, we introduce a simple calculation way, define:

$$Z_i^{Laplacian} = \int p(x_i; b_i)\mathcal{N}(x_i; r_i, \xi_i) exp(\frac{r_i^2}{2\xi_i})dx_i \quad (24a)$$

$$= \frac{b_i}{4}(exp(A_-^2)(1 - \mathrm{erf}(A_-)) + exp(A_+^2)(1 + \mathrm{erf}(A_+))), \quad (24b)$$

where erf() is the Gaussian error function defined as

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z exp(-t^2)dt \quad (25)$$

and

$$A_- = \frac{r_i - \xi_i b_i}{\sqrt{2\xi_i}}; \quad (26a)$$

$$A_+ = \frac{r_i + \xi_i b_i}{\sqrt{2\xi_i}}. \quad (26b)$$

Then the posterior mean $\widehat{x}_i(b_i)$ can be calculated as:

$$\begin{aligned}\widehat{x}_i(b_i) &= \xi_i \frac{\partial \log Z_i^{Laplacian}}{\partial r_i} \\ &= r_i + q(b_i, x_i, w_i) \ , \end{aligned} \tag{27a}$$

where

$$q(b_i, x_i, w_i) = \xi_i b_i \frac{exp(r_i b_i)(1+\text{erf}(A_+)) - exp(-r_i b_i)(1-\text{erf}(A_-))}{exp(r_i b_i)(1+\text{erf}(A_+)) + exp(-r_i b_i)(1-\text{erf}(A_-))}. \tag{28}$$

Therefore, (22) can be represented as:

$$\hat{b}_i = \arg\min_{b_i} \mathrm{E}_{w_i} \left\{ 2 w_i q(b_i, x_i, w_i) + [q(b_i, x_i, w_i)]^2 \right\} + \xi_i. \tag{29}$$

However, (29) is not computationally feasible since the expectation cannot be calculated analytically, and obtaining a closed-form optimized value is hindered by the complexity of the expression. A naive approach is to approximate the expectation via Monte Carlo method, expressed as:

$$\hat{b}_i \approx \arg\min_{b_i} \frac{1}{L} \left[\sum_{l=1}^{L} 2 w_{il} q(b_i, w_{il}) + [q(b_i, w_{il})]^2 \right] + \xi_i. \tag{30}$$

where w_{il} is the generated random sample in $\mathcal{N}(w_i; 0, \xi_i)$ and L is the sample number.

V. Sparse Bayesian Learning (GGD with $\beta \to +\infty$)

When $\beta \to +\infty$, the GGD trends to be uniform distribution as we mentioned in (3). The posterior mean $\hat{x}_i(\alpha_i)$ with uniform prior in (3) and Gaussian extrinsic in (8) can be given as:

$$\hat{x}_i(\alpha_i) = \frac{\int_{-\alpha_i}^{\alpha_i} x_i \mathcal{N}(x_i; r_i, \xi_i) \mathrm{d}x_i}{\int_{-\alpha_i}^{\alpha_i} \mathcal{N}(x_i; r_i, \xi_i) \mathrm{d}x_i} = r_i + g(\alpha_i, r_i, \xi_i), \tag{31}$$

where

$$g(\alpha_i, r_i, \xi_i) = \sqrt{\frac{2}{\pi \xi_i}} \left\{ \frac{\exp\left[-\frac{(\alpha_i + r_i)^2}{2\xi_i}\right] - \exp\left[-\frac{(\alpha_i - r_i)^2}{2\xi_i}\right]}{\text{erf}\left[-\frac{(\alpha_i + r_i)}{\sqrt{2\xi_i}}\right] + \text{erf}\left[-\frac{(\alpha_i - r_i)}{\sqrt{2\xi_i}}\right]} \right\}. \tag{32}$$

Then optimizing α_i via $\text{MSE}_{x_i}(\alpha_i)$ can be expressed as:

$$\hat{\alpha}_i = \arg\min_{\alpha_i} \mathrm{E}_{w_{il}} \left\{ 2 w_i g(\alpha_i, x_i, w_{il}) + [g(\alpha_i, x_i, w_{il})]^2 \right\} + \xi_i. \tag{33}$$

However, (33) is not feasible since the expectation cannot be calculated analytically, and obtaining a closed-form optimized $\hat{\alpha}_i$ is hindered by the complexity of the expression. Employing the same technique as (29), we arrive at:

$$\hat{\alpha}_i \approx \arg\min_{b_i} \frac{1}{L} \left[\sum_{l=1}^{L} 2 w_{il} g(\alpha_i, x_i, w_{il}) + [g(\alpha_i, x_i, w_{il})]^2 \right] + \xi_i. \tag{34}$$

where w_{il} is the generated random sample in $\mathcal{N}(w_i; 0, \xi_i)$ and L is the sample number.

VI. Numerical Experiments

A. Simulation Setup

In the numerical experiments, we assess the performance of three different GGD priors: Laplacian prior ($\beta_i = 1$), Gaussian prior ($\beta_i = 2$), and uniform prior ($\beta_i = +\infty$), across varying levels of noise. We utilize a scalar model described in Equation (7). To approximate the expectation with respect to w_i, we employ Monte Carlo, as indicated in Equations (30) and (34). For $x_i = 0$, we set ξ_i to 0.001, 0.01, and 0.1; for $x_i = 1$, the signal-to-noise ratio (SNR) x_i^2/ξ_i is set to 5 dB, 10 dB, and 15 dB. We perform Monte Carlo trials L times, with L set to 100. For scenarios where $x_i \neq 0$, we utilize the MSE_{x_i} with optimized optimal Gaussian prior as the criterion, represented by a dashed line.

B. Results

1) $x_i = 0$: For $x_i = 0$, Figures 2 and 3 depict the mean MSE of the Laplacian and uniform priors, respectively, with varying hyperparameters α_i. The pentagrams represent the minimal MSE points for each prior. Remarkably, although there is no definitive approach to prove that the fixed point should be $\hat{b}_i = 0$, in our simulations, the minimum point consistently aligns with $\hat{b}_i = 0$. When $\hat{b}_i = 0$, the MSE of x_i equals ξ_i. While there is no analytic solution for the Laplacian prior, in our simulations, all minimum points exhibit smaller MSEs than ξ_i across different scenarios.

2) $x_i = 1$: For $x_i = 1$, Figures 4 and 5 display the MSE of the Laplacian and uniform priors, respectively, with varying hyperparameters α_i. Notably, the pentagrams denote the minimal MSE points for each prior. Similarly to the case when $x_i = 1$, for the Laplacian prior, the optimal $\hat{b}_i$ remains 0, resulting in an MSE of x_i equal to ξ_i. Additionally, across different scenarios, all minimum points for the Laplacian prior exhibit smaller MSEs than those of the optimal Gaussian prior.

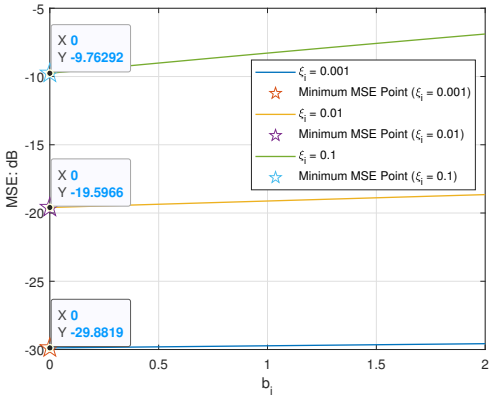

Fig. 2. MSE_{x_i} with $x_i = 0$ of Laplacian prior.

C. Discussion

The simulation results indicate that under various noise conditions, the minimum MSE linked with the Laplacian prior tends to be larger than that of the optimal Gaussian prior.

Conversely, the minimum MSE associated with the uniform prior tends to be smaller than that of the optimal Gaussian prior when $x_i \neq 0$. However, for $x_i = 0$, the optimal Gaussian prior outperforms other cases, as it precisely identifies 0 with a minimum MSE of 0.

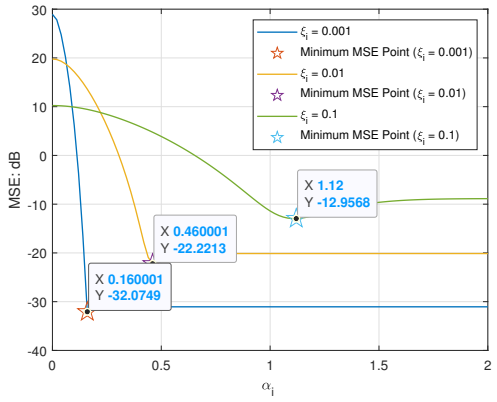

Fig. 3. MSE_{x_i} with $x_i = 0$ of uniform prior.

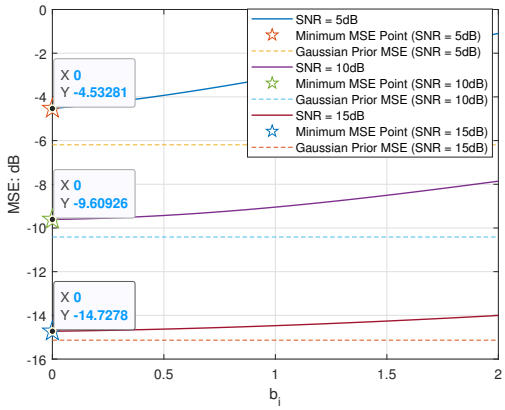

Fig. 4. MSE_{x_i} with $x_i = 1$ of Laplacian prior.

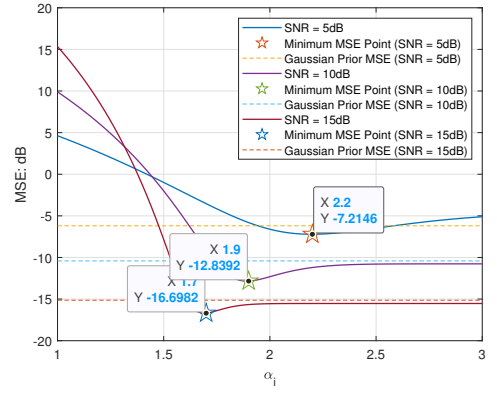

Fig. 5. MSE_{x_i} with $x_i = 1$ of uniform prior.

VII. Conclusion

In this paper, we investigate the impact of utilizing a generalized Gaussian distribution (GGD) prior in Sparse Bayesian Learning (SBL), particularly concerning the scenario where x is zero or non-zero. While both Laplacian, Gaussian, and uniform priors offer regularization benefits, our findings suggest that the Gaussian prior consistently outperforms the Laplacian and uniform priors in terms of minimizing mean squared error (MSE) under varying noise conditions when $x = 0$. This superiority arises from its ability to perfectly recover sparse cases with optimally optimized p. However, for non-zero x, the uniform prior demonstrates superior performance in minimizing MSE compared to the other two priors. It's important to note that these conclusions are drawn from simulations and lack analytic analysis. Further research could delve into additional factors influencing the selection of GGD priors by exploring different β values and examining the disparity between ideal optimal p_i and real estimated p_i derived from finite data.

Acknowledgements EURECOM's research is partially supported by its industrial members: ORANGE, BMW, SAP, iABG, Norton LifeLock, and by the Franco-German projects 5G-OPERA and CellFree6G.

References

[1] C. Qian, X. Fu, N. D. Sidiropoulos, and Y. Yang, "Tensor-based parameter estimation of double directional massive MIMO channel with dual-polarized antennas," in *ICASSP*, 2018.

[2] Z. Yang, L. Xie, and C. Zhang, "Off-Grid Direction of Arrival Estimation using Sparse Bayesian Inference," *IEEE Trans. On Sig. Process.*, vol. 61, no. 1, 2013.

[3] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic Source Imaging with FOCUSS: a Recursive Weighted Minimum Norm Algorithm," *J. Electroencephalog. Clinical Neurophysiol.*, vol. 95, no. 4, 1995.

[4] D. P. Wipf and B. D. Rao, "Sparse Bayesian Learning for Basis Selection ," *IEEE Trans. on Sig. Proc.*, vol. 52, no. 8, Aug. 2004.

[5] David Wipf, "Sparse Estimation with Structured Dictionaries," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, Eds., 2011, vol. 24.

[6] R. Giri and Bhaskar D. Rao, "Type I and type II bayesian methods for sparse signal recovery using scale mixtures," *IEEE Trans. on Sig Process.*, vol. 64, no. 13, 2018.

[7] Michael E. Tipping and Anita C. Faul, "Fast Marginal Likelihood Maximisation for Sparse Bayesian Models," in *AISTATS*, January 2003.

[8] Dirk Slock, "Sparse Bayesian Learning with Stein's Unbiased Risk Estimator based Hyperparameter Optimization," in *ACSSC*, 2022, pp. 857–861.

[9] W. James and C. M. Stein, "Estimation with quadratic loss," *Proc. of Four. Berk. Sympo. on Mathe. Stat. Prob., Berk.: Univ. of Calif. Press,*, 1961.

[10] Jihao Yin, Jianying Sun, and Xiuping Jia, "Sparse analysis based on generalized gaussian model for spectrum recovery with compressed sensing theory," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 8, no. 6, pp. 2752–2759, 2015.

[11] Zilu Zhao, Fangqing Xiao, and Dirk Slock, "Approximate Message Passing for Not So Large niid Generalized Linear Models," in *SPAWC*. IEEE, 2023, pp. 386–390.

[12] Sundeep Rangan, Philip Schniter, and Alyson K. Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.

[13] Thomas P. Minka, "Expectation propagation for approximate Bayesian inference," *arXiv preprint*, 2013.

[14] M. Triki and D. Slock, "Component-Wise Conditionally Unbiased Bayesian Parameter Estimation: General Concept and Applications to Kalman Filtering and LMMSE Channel Estimation," in *Proc. Asilomar Conf. on Sig., Sys., and Comp.*, Nov. 2005.