Hyperparameter Optimization of Sparse Bayesian Learning based on Stein's Unbiased Risk Estimator

Fangqing Xiao, Dirk Slock
Communication Systems Department, EURECOM, France
fangqing.xiao@eurecom.fr, dirk.slock@eurecom.fr

***Abstract*—Sparse Bayesian Learning (SBL) stands as a widely utilized compressed sensing technique wherein the sparsity-inducing prior for the unknowns within the underdetermined linear system is characterized by a Gaussian scale mixture. This formulation results in several hyperparameters, which encompass the variance profile, noise variance, and potentially other parameters within the variance profile priors. Traditionally, these hyperparameters are determined via Type I or Type II Maximum Likelihood (ML) estimation methods. In this paper, we introduce SURE SBL, wherein the optimization of hyperparameters (as opposed to mere estimation) relies on Stein's Unbiased Risk Estimator (SURE). Notably, the primary performance criterion typically centers on the Mean Squared Error (MSE) of the sparse parameters or the resultant signal model. We conduct a review of the SURE approach. Subsequently, we apply the SURE approach to assess the MSE of the sparse parameters (the input to the linear model) and observe that it produces identical hyperparameter optimization outcomes as those obtained via Type II ML. Furthermore, we propose extending the SURE approach to the output level of the linear model. Remarkably, in the context of the large system limit, this extension yields equivalent hyperparameter optimization outcomes concerning the input to the linear model; however, when measurement noise is present, the results obtained by the two kinds of SURE optimizers diverge from those obtained through MSE optimization.**

I. Introduction

Sparse signal reconstruction and compressed sensing (CS) have garnered significant attention in recent years across various fields. Applications span from massive multi-input multi-output (MIMO) channel estimation [1], direction of arrival estimation [2], biomagnetic imaging [3], to tasks such as image restoration and echo cancellation. The compressed sensing (CS) problem can be formulated as:

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

where $\boldsymbol{y}$ are the observations or data, $\boldsymbol{A}$ is called the measurement or sensing matrix which in a first instance is known and is of dimension $N \times M$ with $N < M$, $\boldsymbol{x}$ is the M-dimensional sparse signal and $\boldsymbol{w}$ is the additive noise. In the exactly sparse case, the unknown $\boldsymbol{x}$ contains only K non-zero entries, with $K << M$. $\boldsymbol{w}$ is assumed to be a white Gaussian noise, $\boldsymbol{w} \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I})$ with precision (inverse variance) γ. To address this problem, a variety of algorithms such as Orthogonal Matching Pursuit (OMP) [4], basis pursuit [5] and the iterative re-weighted l_1 and l_2 algorithms [6] exist in the literature. Compared to these algorithms, employing Bayesian techniques for sparse signal recovery (SSR) typically achieves superior performance. Notably, [7] offers a comprehensive overview of various SSR algorithms falling under l_1 or l_2 norm minimization approaches, such as Basis Pursuit, LASSO, etc., as well as Sparse Bayesian Learning (SBL) methods. The authors substantiate the enhanced recovery performance of SBL compared to the conventional methods mentioned above.

The SBL algorithm was initially introduced by [8] and then first proposed for sparse signal recovery (SSR) by [9]. In a Bayesian framework, the objective is to compute the posterior distribution of the parameters $\boldsymbol{x}$ given some observations (data) and prior knowledge. One of the distinguishing features of SBL, compared to other state-of-the-art techniques, is its utilization of hierarchical prior modeling, leading to the sparsification of the state $\boldsymbol{x}$. The Bayesian LASSO, as described by [10], employs a similar hierarchical modeling approach with a Gaussian-Exponential prior (equivalent to a Laplace prior), which is revealed to be a special case of the Student-t prior used in SBL.

In SBL, the unknown parameters $\boldsymbol{x}$ are modeled as decorrelated zero-mean Gaussian [1]$\boldsymbol{x} \sim \mathcal{N}(0, (\mathrm{Diag}(\boldsymbol{\xi}))^{-1})$ with precision profile $\boldsymbol{\xi}$. The estimation of the hyperparameters $\boldsymbol{\xi}, \gamma$ and the sparse signal $\boldsymbol{x}$ is performed jointly. In one approach, the hyperparameters are estimated first using evidence maximization, which is referred to as the Type II Maximum Likelihood (ML) method [7], which is also an instance of Empirical Bayes (EB) estimation (i.e. Bayesian estimation with a parameterized prior in which the hyperparameters are estimated also). For a given estimate of $\widehat{\boldsymbol{\xi}}, \gamma$, the Gaussian posterior of $\boldsymbol{x}$ is formulated as $p(\boldsymbol{x}|\boldsymbol{y}, \widehat{\boldsymbol{\xi}}, \widehat{\gamma})$ and the mean of this posterior distribution is used as a Linear Mimimum Mean Squared Error (LMMSE) [11] point estimate of $\widehat{\boldsymbol{x}}$. In [12], the authors propose a Fast Marginalized ML (FMML) by alternating likelihood maximization w.r.t. the hyperparameters. Both previous approaches allow for a greedy (OMP-like, Orthogonal Matching Pursuit) initialization which improves convergence speed. Recently, Approximate Message Passing (AMP) [13], generalized AMP (GAMP) and vector AMP

[1] Notations: The operator $(\cdot)^T$ denotes the matrix transpose. The probability density function (pdf) of a Gaussian random variable x with mean μ and variance σ^2 is denoted as $\mathcal{N}(x; \mu, \nu)$. x_k denotes the k^{th} element of the vector $\boldsymbol{x}$. $KL(q||p)$ denotes the Kullback-Leibler distance between the distributions q and p. $\boldsymbol{A}_n$ represents the n^{th} column of the matrix $\boldsymbol{A}$. diag($\mathbf{X}$) or Diag($\boldsymbol{x}$) represents a vector obtained by extracting the diagonal elements of the matrix $\mathbf{X}$ or a diagonal matrix obtained with the elements of $\boldsymbol{x}$ on the diagonal, respectively. $\mathbf{I}_M$ represents a vector of length M with all elements set to one. For a matrix $\boldsymbol{A}$, $\boldsymbol{A} \geq 0$ indicates that it is non-negative definite. $\mathbf{I}_M$ denotes the identity matrix of size M. tr$\{\boldsymbol{A}\}$ represents the trace of $\boldsymbol{A}$ (the sum of its diagonal elements). $\boldsymbol{A}_{ij}$ denotes the element at row i and column j of matrix $\boldsymbol{A}$. $\lfloor a \rfloor_+$ denotes $max(0, a)$.

(VAMP) [14], [15], [16] were introduced to compute the posterior distributions in a message passing (MP) framework, with reduced complexity. The fundamental idea behind the derivation of AMP is the central limit theorem and Taylor series expansions, which allows to simplify the messages to be exchanged in MP and reduce their number. However, so far the Bayes optimality of these AMP algorithms has been shown only for i.i.d. or right orthogonally invariant $\boldsymbol{A}$, which severely limits their applicability. More recent attempts at obtaining converging versions of (G)AMP appear in [17], [18], where alternating constrained minimization of a large system limit of the Bethe Free Energy is pursued.

SBL (LMMSE) entails a matrix inversion step, particularly at each iteration in Type I ML, which involves the joint estimation of parameters $\boldsymbol{x}$ and hyperparameters. This characteristic renders it computationally complex, especially for moderately large datasets. An alternative approach to SBL involves utilizing variational approximation for Bayesian inference, as proposed by [19]. Variational Bayesian (VB) inference aims to discover a factored approximation of the posterior distribution that maximizes the variational lower bound on $\ln p(\boldsymbol{y})$. In a similar vein, [20] introduces a fast version of SBL by iteratively maximizing the variational posterior lower bound with respect to (w.r.t.) individual (hyper)parameters. Another notable approach is presented in [21], wherein the authors introduce a Belief Propagation (BP)-based SBL algorithm, which proves to be computationally more efficient. Here, BP is employed to infer the posterior probability density function (pdf) of $\boldsymbol{x}$, while the hyperparameters are estimated using the Expectation-Maximization (EM) algorithm. Furthermore, [22] utilizes the Approximate Message Passing (AMP) algorithm for LMMSE and introduces a non-parametric algorithm called NOPE, which doesn't necessitate any prior knowledge of the signal and noise powers. Notably, these parameters are adjusted via SURE. The authors also demonstrate that in the large system limit, NOPE achieves performance comparable to that of the LMMSE equalizer.

Another approach is presented in [23] (and previous publications by the same authors), known as the SPICE methodology. In this approach, hyperparameters are adjusted through covariance fitting using a weighted covariance fitting cost function.

$$\mathrm{tr}\{(\boldsymbol{y}\boldsymbol{y}^T-\boldsymbol{R})\boldsymbol{R}^{-1}(\boldsymbol{y}\boldsymbol{y}^T-\boldsymbol{R})\} \tag{2}$$

where $\boldsymbol{R}$ is the one appearing in (15). Now, (2) differs from the optimally weighted covariance fitting criterion

$$\mathrm{tr}\{(\boldsymbol{y}\boldsymbol{y}^T-\boldsymbol{R})\boldsymbol{R}^{-1}(\boldsymbol{y}\boldsymbol{y}^T-\boldsymbol{R})\boldsymbol{R}^{-1}\} \tag{3}$$

which leads to the same hyperparameter adjustments as Type II ML (EB).

In this paper, we have tried to analyze the use of SURE estimator for hyperparameter estimation of SBL and by analyzing $\mathrm{SURE}_{\boldsymbol{x}}$ and $\mathrm{SURE}_{\mathbf{z}}$, we have come up with the expression of the estimator w.r.t. the hyperparameters of SBL. And we analyze that both SURE optimizers have the same effect under the large system assumption, but diverge to MSE optimization when measurement is not negligible.

II. Stein's Unbiased Risk Estimator: SURE Principle

Consider a simple additive white Gaussian noise model:

$$\boldsymbol{y}=\mathbf{z}+\boldsymbol{w} \tag{4}$$

where $\boldsymbol{w}\sim\mathcal{N}(\boldsymbol{v};0,\sigma^2\mathbf{I}_M)$. Let $\widehat{\mathbf{z}}(\boldsymbol{y})$ be an estimator of $\mathbf{z}$. Then we get for the MSE

$$\begin{aligned}\mathrm{MSE}_{\mathbf{z}} &= \mathrm{E}\,\|\widehat{\mathbf{z}}-\mathbf{z}\|^2=\mathrm{E}\,\{\|\mathbf{z}\|^2+\|\widehat{\mathbf{z}}\|^2-2\,\widehat{\mathbf{z}}^T\mathbf{z}\}\\ &\overset{(a)}{=}\mathrm{E}\,\{\|\mathbf{z}\|^2+\|\widehat{\mathbf{z}}\|^2-2\,\widehat{\mathbf{z}}^T\boldsymbol{y}+2\sigma^2\,\mathrm{tr}\{\tfrac{\partial\widehat{\mathbf{z}}^T}{\partial\boldsymbol{y}}\}\,\}\\ &=\mathrm{E}\,\{\|\mathbf{z}\|^2-\|\boldsymbol{y}\|^2+\|\widehat{\mathbf{z}}-\boldsymbol{y}\|^2+2\sigma^2\,\mathrm{tr}\{\tfrac{\partial\widehat{\mathbf{z}}^T}{\partial\boldsymbol{y}}\}\,\}\end{aligned} \tag{5}$$

where E is w.r.t. $\boldsymbol{w}$ ($\mathbf{z}$ is treated as deterministic) and (a) follows as a property of the Gaussian pdf [24]. By dropping expectation, we get an instantaneous unbiased estimate of the MSE and the corresponding SURE function (which is the part of $\widehat{\mathrm{MSE}}$ that depends on $\widehat{\mathbf{z}}$)

$$\begin{aligned}\widehat{\mathrm{MSE}}_{\mathbf{z}} &= \|\mathbf{z}\|^2-\|\boldsymbol{y}\|^2+\mathrm{SURE}_{\mathbf{z}}\,,\\ \mathrm{SURE}_{\mathbf{z}} &= \|\widehat{\mathbf{z}}-\boldsymbol{y}\|^2+2\sigma^2\,\mathrm{tr}\{\tfrac{\partial\widehat{\mathbf{z}}^T}{\partial\boldsymbol{y}}\}\,.\end{aligned} \tag{6}$$

In $\mathrm{SURE}_{\mathbf{z}}$, the first term reflects the effect of bias in $\widehat{\mathbf{z}}$ whereas the second term reflects the variance of $\widehat{\mathbf{z}}$ and the noise effect in the first term due to replacing $\mathbf{z}$ by $\boldsymbol{y}$.

III. Prior Variance Determination in SBL Algorithms

Consider an analysis per component x_i in which we optimize over the prior variance p_i, keeping others $P_{\bar{i}}$ fixed. Then Variational Bayes, like EM, converges to:

$$\hat{p}_i=|\widehat{x}_i(p_i)|^2+\sigma^2_{\widetilde{x}_i(p_i)} \tag{7a}$$

$$=\lfloor|\widehat{x}_i(0)|^2-\sigma^2_{\widetilde{x}_i(0)}\rfloor_+ \tag{7b}$$

where $\widehat{x}_i(p_i)$ and $\sigma^2_{\widetilde{x}_i(p_i)}$ are the LMMSE estimate and the corresponding error variance for a priori variance p_i. The first line (7a) corresponds to the update equation at convergence of VB (or EM), yielding an implicit equation for p_i. The expression corresponds to the *orthogonality principle of LMMSE*: the prior variance equals the estimate variance plus the error variance, where the estimate variance is replaced by its instantaneous value.

The second line (7b) is the corresponding solution, which is also the estimate for p_i in Type II ML (EB). It is again an intuitive expression: *for an unbiased estimate*, the power in the estimate equals the prior power plus the estimation error variance.

IV. First SBL SURE application: component-wise x_i

Consider component i of the LMMSE estimate for $\boldsymbol{x}$ in SBL, $\widehat{x}_i(p_i)$. Then a simple instance of the previous additive noise model is

$$\widehat{x}_i(0)=x_i+\widetilde{x}_i(0) \tag{8}$$

where $\widetilde{x}_i(0)$ has variance $\sigma^2=\sigma^2_{\widetilde{x}_i(0)}$. We consider the LMMSE estimator

$$\widehat{x}_i=\widehat{x}_i(p_i)=\frac{p_i}{p_i+\sigma^2}\,\widehat{x}_i(0). \tag{9}$$

Then we get

$$\begin{aligned}\mathrm{SURE}_{x_i}(p_i) &= \left(\frac{\sigma^2}{p_i+\sigma^2}\,\widehat{x}_i(0)\right)^2 + 2\frac{\sigma^2\,p_i}{p_i+\sigma^2} \\ &= \left(\frac{\sigma^2}{p_i+\sigma^2}\,\widehat{x}_i(0)\right)^2 - 2\frac{\sigma^4}{p_i+\sigma^2} + 2\sigma^2\end{aligned} \tag{10}$$

where as a function of p_i, the first term is decreasing and the second term is increasing. We get

$$\frac{\partial \mathrm{SURE}_{x_i}}{\partial p_i} = 2\sigma^4(p_i+\sigma^2-\widehat{x}_i^2(0))/(p_i+\sigma^2)^3. \tag{11}$$

$\mathrm{SURE}_{x_i}(p_i)$ has a single extremum, a local minimum, at $p_i = \widehat{x}_i^2(0)-\sigma^2$. We have

$$\frac{\partial \mathrm{SURE}_{x_i}}{\partial p_i}(p_i=0) = 2(1-\frac{\widehat{x}_i^2(0)}{\sigma^2}). \tag{12}$$

So, the minimum of $\mathrm{SURE}_{x_i}(p_i)$ occurs at positive p_i when $\widehat{x}_i^2(0) > \sigma^2$, but at negative p_i in the opposite case. Hence, since we need $p_i \geq 0$, we get for the optimum

$$\hat{p}_i = \lfloor |\widehat{x}_i(0)|^2 - \sigma^2_{\widetilde{x}_i(0)} \rfloor_+ \tag{13}$$

which leads to exactly the same result as by VB or Type II ML (EB). This could be extended to the (non-Gaussian) Generalized Linear Model via GAMP.

V. SURE applied to SBL: Discussion

Consider now the linear model $\mathbf{z} = \boldsymbol{A}\boldsymbol{x}$ with diagonal Gaussian prior for $\boldsymbol{x}$: a simple additive white Gaussian noise model:

$$\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{x} + \boldsymbol{w}\;, \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{v}; 0, \sigma^2\mathbf{I})\,,\; \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}; 0, \boldsymbol{P}) \tag{14}$$

where $\boldsymbol{x}, \boldsymbol{v}$ are independent. By the Gauss-Markov theorem, the posterior for $\boldsymbol{x}$ is Gaussian again

$$\boldsymbol{x}|\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x};\, \boldsymbol{P}\boldsymbol{A}^T\boldsymbol{R}^{-1}\boldsymbol{y},\, \boldsymbol{P} - \boldsymbol{P}\boldsymbol{A}^T\boldsymbol{R}^{-1}\boldsymbol{A}\boldsymbol{P}) \tag{15}$$

where $\boldsymbol{R} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T + \sigma^2\mathbf{I}$ is the covariance matrix of $\boldsymbol{y}$.

In the SURE approach, the *Gaussian prior on $\boldsymbol{x}$* is not really considered as the true prior, but rather as a mechanism that leads to *biased estimates for $\boldsymbol{x}$ in a principled way*, allowing to *optimize the bias for MMSE*.

In some compressed sensing settings (e.g. DoA estimation), the important information is in the *support of $\boldsymbol{x}$* (or $\mathrm{diag}(\boldsymbol{P})$). In that case the *estimation of the individual components* x_i and their prior power p_i is indeed important (previous section).

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 [25]. 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:

$$\hat{x}_i(0) = \frac{\boldsymbol{A}_i^T(\sum_{j\neq i}^N p_j\boldsymbol{A}_j\boldsymbol{A}_j^T + \sigma^2\mathbf{I})^{-1}\boldsymbol{y}}{\boldsymbol{A}_i^T(\sum_{j\neq i}^N p_j\boldsymbol{A}_j\boldsymbol{A}_j^T + \sigma^2\mathbf{I})^{-1}\boldsymbol{A}_i}; \tag{16a}$$

$$\sigma^2_{\widetilde{x}_i(0)} = (\boldsymbol{A}_i^T(\sum_{j\neq i}^N p_j\boldsymbol{A}_j\boldsymbol{A}_j^T + \sigma^2\mathbf{I})^{-1}\boldsymbol{A}_i)^{-1}. \tag{16b}$$

Note that the *(partial) Bayesian modeling (of $\boldsymbol{x}_{\bar{i}}$) is a must here*, in the application of SURE, as *no deterministic estimate of $\boldsymbol{x}$ is possible in the underdetermined case.*

VI. Second SBL SURE application: linear model output $\mathbf{z} = \boldsymbol{A}\boldsymbol{x}$

In other compressed sensing settings (e.g. channel estimation with a superposition of multipath components), the important quantity is $s = \mathbf{C}\boldsymbol{x}$ in which *a signal s gets represented (approximated) as a superposition of atoms in a dictionary* $\mathbf{C}$. In this case, $\boldsymbol{x}$ is not as important as the resulting s. In compressed sensing, we cannot *measure* the whole of s but *only a projection (sketch)* $\mathbf{z} = \mathbf{B}s = \boldsymbol{A}\boldsymbol{x}$ with $\boldsymbol{A} = \mathbf{BC}$. for instance, in OFDM based wireless channel estimation, $\mathbf{B}$ may have the structure of a fat permutation submatrix and is semi-orthogonal. In such case, the MSE on $\mathbf{z}$ is representative of the MSE on s. Hence we *focus on the estimation of* $\mathbf{z}$, which in case of no RIP (Restricted Isometry Property) on $\boldsymbol{A}$ could be quite different from a superposition of estimations of the x_i.

The estimation in the underdetermined linear model (fat $\boldsymbol{A}$) is related to the case of reduced rank (overdetermined) $\boldsymbol{A}$ discussed in [24].

Hence with $\widehat{\mathbf{z}} = \boldsymbol{A}\widehat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T\boldsymbol{R}^{-1}\boldsymbol{y}$, parameterized by $\boldsymbol{P}$,

$$\begin{aligned}&\mathrm{SURE}_{\mathbf{z}}(\boldsymbol{P}) = \|\boldsymbol{y}-\widehat{\mathbf{z}}\|^2 + 2\sigma^2\,\mathrm{tr}\{\tfrac{\partial\widehat{\mathbf{z}}^T}{\partial\boldsymbol{y}}\} = \sigma^4\boldsymbol{y}^T\boldsymbol{R}^{-2}\boldsymbol{y} \\ &+2\sigma^2\mathrm{tr}\{\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T\boldsymbol{R}^{-1}\} = 2\sigma^2 + \sigma^4\boldsymbol{y}^T\boldsymbol{R}^{-2}\boldsymbol{y} - 2\sigma^4 tr\{\boldsymbol{R}^{-1}\}.\end{aligned} \tag{17}$$

Focusing on optimizing one p_i at a time, making explicit the dependence on p_i, we get

$$\begin{aligned}&\hat{p}_i = \arg\min_{p_i}\;\mathrm{SURE}_{\mathbf{z}}(\boldsymbol{P}) \\ &= \arg\min_{p_i}\;\boldsymbol{y}^T\boldsymbol{R}^{-2}\boldsymbol{y} - 2tr\{\boldsymbol{R}^{-1}\} \\ &= \arg\min_{p_i}\;(\boldsymbol{A}_i^T\boldsymbol{R}_{\tilde{i}}^{-1}\boldsymbol{A}_i + 1/p_i)^{-2}\boldsymbol{A}_i^T\boldsymbol{R}_{\tilde{i}}^{-2}\boldsymbol{A}_i(\boldsymbol{A}_i^T\boldsymbol{R}_{\tilde{i}}^{-1}\boldsymbol{y})^2 \\ &-2(\boldsymbol{A}_i^T\boldsymbol{R}_{\tilde{i}}^{-1}\boldsymbol{A}_i + 1/p_i)^{-1}(\boldsymbol{y}^T\boldsymbol{R}_{\tilde{i}}^{-2}\boldsymbol{A}_i\boldsymbol{A}_i^T\boldsymbol{R}_{\tilde{i}}^{-1}\boldsymbol{y} - \boldsymbol{A}_i\boldsymbol{R}_{\tilde{i}}^{-2}\boldsymbol{A}_i) \\ &+\boldsymbol{y}^T\boldsymbol{R}_{\tilde{i}}^{-2}\boldsymbol{y} \\ &= \arg\min_{p_i}\left[a - \frac{2\,b}{c+1/p_i} + \frac{d}{(c+1/p_i)^2}\right],\end{aligned} \tag{18}$$

where

$$b = \boldsymbol{y}^T\boldsymbol{R}_{\tilde{i}}^{-2}\boldsymbol{A}_i\boldsymbol{A}_i^T\boldsymbol{R}_{\tilde{i}}^{-1}\boldsymbol{y} - \boldsymbol{A}_i\boldsymbol{R}_{\tilde{i}}^{-2}\boldsymbol{A}_i; \tag{19a}$$

$$c = \boldsymbol{A}_i^T\boldsymbol{R}_{\tilde{i}}^{-1}\boldsymbol{A}_i; \tag{19b}$$

$$d = \boldsymbol{A}_i^T\boldsymbol{R}_{\tilde{i}}^{-2}\boldsymbol{A}_i(\boldsymbol{A}_i^T\boldsymbol{R}_{\tilde{i}}^{-1}\boldsymbol{y})^2. \tag{19c}$$

With the limit of p_i should always be non-negative, we get

$$\hat{p}_i = \lfloor\frac{b}{d-cb}\rfloor_+\;\;. \tag{20}$$

Though this expression requires further interpretation, it is expected that the assignment of power p_i in $\mathrm{SURE}_{\mathbf{z}}$ is (even) more affected (more sparsifying) in the case that $\boldsymbol{A}$ contains columns that are close to collinear.

VII. Large System Analysis

In our treatment of the linear regression problem (1), the vectors $\boldsymbol{y} = [y_1, \cdots, y_M]^T$, $\boldsymbol{x} = [x_1, \cdots, x_N]^T$, and $\boldsymbol{w} = [w_1, \cdots, w_M]^T$ are deterministic, while the matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ is also deterministic. However, it's crucial to note that we assume the components A_{ij} of $\boldsymbol{A}$ are realizations of independent and identically distributed (i.i.d.) Gaussian random variables $A_{ij} \sim \mathcal{N}(0, \frac{1}{M})$, which are drawn independently of $\boldsymbol{x}$ and $\boldsymbol{w}$. Throughout our analysis, we will primarily focus on the following large-system limit. *Definition 1:* The "large system limit" is defined as $M, N \to \infty$ with $M/N \to \alpha$ for some fixed sampling ratio $\alpha \in (0, \infty)$.

A. Preliminary

The analysis of the large system primarily relies on the deterministic equivalent proposed in [Wagner, 2012], which states:

Lemma 1: *Let $\boldsymbol{Q}$ be any Hermitian deterministic matrix and let $\boldsymbol{\Sigma} \in \mathbb{R}^{M \times M} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T = \sum_{i=1}^N p_i \boldsymbol{A}_i \boldsymbol{A}_i^T$, with diagonal matrix $\boldsymbol{P}$, and $\boldsymbol{A}$ containing N independent columns A_i with covariance matrix $\boldsymbol{\Theta}_i$. Also, assume that $\boldsymbol{Q}$, $\boldsymbol{\Theta}_i$ have uniformly bounded spectral norms. Then, for any $z > 0$ the following convergence result holds almost surely,*

$$\frac{1}{M} tr\{\boldsymbol{Q}(\boldsymbol{\Sigma} + z\mathbf{I})^{-1}\} - \frac{1}{M} tr\{\boldsymbol{Q}\mathbf{T}(z)\} \xrightarrow[M \to \infty]{a.s} 0, \tag{21}$$

with

$$\mathbf{T}(z) = \left(\sum_{i=1}^N \frac{d_i \boldsymbol{\Theta}_i}{1 + l_i(z)} + z\mathbf{I} \right)^{-1} \tag{22}$$

where $l_i(z) = l_i^{(\infty)}(z)$ is defined as the unique positive solution of

$$l_i(z) = tr \left\{ d_i \boldsymbol{\Theta}_i \left(\sum_{i=1}^N \frac{d_i \boldsymbol{\Theta}_i}{1 + l_i(z)} + z\mathbf{I} \right) \right\}^{-1}. \tag{23}$$

Also, in the appendix VI of [26], they defined as

Lemma 2: $\boldsymbol{A}_i^T \boldsymbol{\Sigma} \boldsymbol{A}_i - \frac{1}{M} tr\{\boldsymbol{\Sigma}\} \xrightarrow{N, M \to \infty} 0$ *when the elements of $\boldsymbol{A}_i$ are i.i.d. with zero mean and variance $1/M$ and independent of $\boldsymbol{A}_i$, and similarly when $\boldsymbol{y}$ is independent of $\boldsymbol{A}_i$, that $\boldsymbol{A}_i^T \boldsymbol{\Sigma} \boldsymbol{y} \xrightarrow{M, N \to \infty} 0$.*

Based on Lemma 1, it is obvious to define that

Lemma 3: *If $m^o(z) = \frac{1}{M} tr\{(\boldsymbol{\Sigma} + z\mathbf{I})^{-1}\}$, then $\frac{\partial m^o(z)}{\partial z} = -\frac{1}{M} tr\{(\boldsymbol{\Sigma} + z\mathbf{I})^{-2}\}$ and $\frac{\partial^2 m^o(z)}{\partial z^2} = \frac{1}{2} \frac{1}{M} tr\{(\boldsymbol{\Sigma} + z\mathbf{I})^{-3}\}$.*

Sketch of the proof: Lemma 3 is straightforward via algebra derivation.

B. Large System Analysis to $\hat{p}_i$ w.r.t. $\mathrm{SURE}_{\boldsymbol{x}}$

Theorem 1: Optimizing p_i from $\mathrm{SURE}_{\boldsymbol{x}}(p_i)$, we have the optimized p_i in (13) and $\widehat{x}_i(0), \widetilde{x}_i(0)$ in (16), under large system limit, the following convergence result holds almost surely

$$\hat{p}_i \xrightarrow{M, N \to \infty} \lfloor x_i^2 - \frac{1}{M} \sum_{j \neq i}^N \frac{p_j}{1 + l_j} - \frac{N}{M} \sigma^2 \rfloor_+, \tag{24}$$

where

$$l_j = \frac{p_j M}{N} \left(\sum_{j \neq i}^N \frac{p_j / N}{1 + l_j} + \sigma_v^2 \right)^{-1}. \tag{25}$$

Proof: For the sake of brevity, we define $\boldsymbol{R}_{\tilde{i}} = \sum_{j \neq i}^N p_j \boldsymbol{A}_j \boldsymbol{A}_j^T + \sigma^2 \mathbf{I}$. Then according to (16), we have

$$\begin{aligned} &\hat{x}_i(0)^2 = (\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{y})^2 / (\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i)^2 \\ &= (\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} (\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T \boldsymbol{A} + \boldsymbol{w}\boldsymbol{w}^T) \boldsymbol{R}_{\tilde{i}}^{-T} \boldsymbol{A}_i) / (\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i)^2 \\ &= \left\{ \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \left[\left(\textstyle\sum_{j=1}^N x_j \boldsymbol{A}_j \right) \left(\textstyle\sum_{j=1}^N x_j \boldsymbol{A}_j^T \right) + \boldsymbol{w}\boldsymbol{w}^T \right] \boldsymbol{R}_{\tilde{i}}^{-T} \boldsymbol{A}_i \right\} \\ &/ (\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i)^2. \end{aligned} \tag{26}$$

Each column $\boldsymbol{A}_i$ of $\boldsymbol{A}$ is independent each other and $\boldsymbol{w}$ is independent to all $\boldsymbol{A}_i$, therefore, according to Lemma 2, we can have

$$\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{w} = 0; \tag{27a}$$

$$\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_j = 0, \mathit{if}\ j \neq i; \tag{27b}$$

$$\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i = \frac{1}{N} tr\{\boldsymbol{R}_{\tilde{i}}^{-1}\}. \tag{27c}$$

Thus we have:

$$\hat{x}_i(0)^2 = (x_i^2 (\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i)^2) / (\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i)^2 = x_i^2. \tag{28}$$

According to Lemma 1, we can have

$$\begin{aligned} \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i &= \tfrac{1}{N} tr\{\boldsymbol{R}_{\tilde{i}}^{-1}\} \\ &= \tfrac{1}{N} tr \left\{ \left(\textstyle\sum_{i=1}^N \frac{p_i/N}{1+l_i} \mathbf{I} + \sigma^2 \mathbf{I} \right)^{-1} \right\} \\ &= \tfrac{M}{N} \left(\textstyle\sum_{i=1}^N \frac{p_i/N}{1+l_i} + \sigma^2 \right)^{-1}, \end{aligned} \tag{29}$$

where

$$l_j = \frac{p_j M}{N} \left(\sum_{j \neq i}^N \frac{p_j / N}{1 + l_j} + \sigma_v^2 \right)^{-1}. \tag{30}$$

Defining $m(\sigma^2) = \boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i$, according to Lemma 3, we have

$$\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-2} \boldsymbol{A}_i = -\frac{\partial m(\sigma^2)}{\partial \sigma^2} = \frac{N}{M} m^2(\sigma^2). \tag{31}$$

Then for $\sigma^2_{\widetilde{x}_i(0)}$, we have

$$\sigma^2_{\widetilde{x}_i(0)} = (\boldsymbol{A}_i^T \boldsymbol{R}_{\tilde{i}}^{-1} \boldsymbol{A}_i)^{-1} = 1/m(\sigma^2), \tag{32}$$

where $l_i(z)$ has already defined in (30). Combining (29)(31)(30)(32), we have

$$\hat{p}_i = \lfloor x_i^2 - \frac{1}{M} \sum_{j \neq i}^N \frac{p_j}{1 + l_j} - \frac{N}{M} \sigma^2 \rfloor_+ \tag{33}$$

as Theorem 1.

C. Large System Analysis to $\hat{p}_i$ w.r.t. $\mathrm{SURE}_{\mathbf{z}}$

Theorem 2: Optimizing p_i from $\mathrm{SURE}_{\mathbf{z}}(p_i)$ in (18), we have the optimized $\hat{p}_i$ in (13), under large system limit, the following convergence result holds almost surely

$$\hat{p}_i \xrightarrow{M,N\rightarrow\infty} \lfloor x_i^2 - \frac{1}{M}\sum_{j\neq i}^{N}\frac{p_j}{1+l_j} - \frac{N}{M}\sigma^2\rfloor_+, \tag{34}$$

where

$$l_j = \frac{p_j M}{N}\left(\sum_{j\neq i}^{N}\frac{p_j/N}{1+l_j} + \sigma_v^2\right)^{-1}. \tag{35}$$

Proof: Firstly, following the approach employed in the proof of Theorem 1 and leveraging Lemma 3, we obtain:

$$\frac{1}{2}\frac{\partial^2 m(\sigma^2)}{\partial\sigma^4} = \boldsymbol{A}_i^T\boldsymbol{R}_{\bar{i}}^{-3}\boldsymbol{A}_i = \frac{N^2}{M^2}m^3(\sigma^2). \tag{36}$$

Therefore, we can directly express b, c, and d in (19) as:

$$\begin{aligned} b &= x_i^2\boldsymbol{A}_i^T\boldsymbol{R}_{\bar{i}}^{-1}\boldsymbol{A}_i\boldsymbol{A}_i^T\boldsymbol{R}_{\bar{i}}^{-2}\boldsymbol{A}_i - \boldsymbol{A}_i^T\boldsymbol{R}_{\bar{i}}^{-2}\boldsymbol{A}_i \\ &= x_i^2\frac{N^2}{M^2}m^3(\sigma^2) - \frac{N}{M}m^2(\sigma^2); \end{aligned} \tag{37a}$$

$$c = \boldsymbol{A}_i^T\boldsymbol{R}_{\bar{i}}^{-1}\boldsymbol{A}_i = m(\sigma^2); \tag{37b}$$

$$d = x_i^2\boldsymbol{A}_i^T\boldsymbol{R}_{\bar{i}}^{-2}\boldsymbol{A}_i(\boldsymbol{A}_i^T\boldsymbol{R}_{\bar{i}}^{-1}\boldsymbol{A}_i)^2 = x_i^2\frac{N^2}{M^2}m^4(\sigma^2). \tag{37c}$$

Based on the aforementioned results and subsequent algebraic manipulation, we can express (20) as follows:

$$\hat{p}_i = \lfloor\frac{b}{d-cb}\rfloor_+ = \lfloor x_i^2 - m^{-1}(\sigma^2)\rfloor_+ . \tag{38}$$

As we can see, optimizing p_i from $\mathrm{SURE}_{\mathbf{z}}$ and $\mathrm{SURE}_{\boldsymbol{x}}$ leads to the same result under large system limit.

VIII. Hyperparameter optimization via MSE

For optimizing hyperparameter p_i from $\mathrm{MSE}_{\boldsymbol{x}}$, they can be expressed as:

$$\hat{p}_i = \arg\min_{p_i}\mathrm{MSE}_{\boldsymbol{x}} = \mathbf{E}\{\mathrm{SURE}_{\boldsymbol{x}}\}, \tag{39}$$

where E is w.r.t. $\boldsymbol{w}$ ($\boldsymbol{x}$ is treated as deterministic). With (10) and (11), (39) can be derived as:

$$\hat{p}_i = \lfloor\mathbf{E}_{\boldsymbol{w}}\{|\widehat{x}_i(0)|^2\} - \sigma^2_{\widetilde{x}_i(0)}\rfloor_+ , \tag{40}$$

where $\widehat{x}_i(0)$ and $\mathrm{MSE}_{\boldsymbol{x}}$ are defined in (16), respectively. **Theorem 3:** Optimizing p_i from $\mathrm{MSE}_{\boldsymbol{x}}(p_i)$ defined in (39), we have the optimized $\hat{p}_i$ in (40) and $\widehat{x}_i(0), \widetilde{x}_i(0)$ in (16), under large system limit, the following convergence result holds almost surely,

$$\hat{p}_i \xrightarrow{M,N\rightarrow\infty} \lfloor x_i^2 - \frac{1}{M}\sum_{j\neq i}^{N}\frac{p_j}{1+l_j}\rfloor_+, \tag{41}$$

where

$$l_j = \frac{p_j M}{N}\left(\sum_{j\neq i}^{N}\frac{p_j/N}{1+l_j} + \sigma_v^2\right)^{-1}. \tag{42}$$

Proof: According to (26) and with large system limit, $\mathbf{E}_{\boldsymbol{w}}\{|\widehat{x}_i(0)|^2\}$ can be calculated as:

$$\begin{aligned} \mathbf{E}_{\boldsymbol{w}}\{|\widehat{x}_i(0)|^2\} &= \frac{\boldsymbol{A}_i^T\boldsymbol{R}_{\bar{i}}^{-1}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A} + \sigma^2\mathbf{I})\boldsymbol{R}_{\bar{i}}^{-T}\boldsymbol{A}_i}{(\boldsymbol{A}_i^T\boldsymbol{R}_{\bar{i}}^{-1}\boldsymbol{A}_i)^2} \\ &= x_i^2 + \sigma^2\frac{\boldsymbol{A}_i^T\boldsymbol{R}_{\bar{i}}^{-2}\boldsymbol{A}_i}{(\boldsymbol{A}_i^T\boldsymbol{R}_{\bar{i}}^{-1}\boldsymbol{A}_i)^2} = x_i^2 + \frac{N}{M}\sigma^2. \end{aligned} \tag{43}$$

Afterwards by a simple derivation, similar to the proof of Theorem 1, we can prove the correctness of Theorem 3. According to Theorem 1,2,3, under the large system limit, the two SURE optimizers yield the same results as the MSE optimizer when the noise tends towards zero. However, discrepancies arise when the noise is significant.

IX. Conclusion

In this paper, we explore the utilization of Stein's unbiased risk estimation (SURE) for hyperparameter estimation within sparse Bayesian learning. We provide the derivation of expressions for SURE estimators w.r.t. hyperparameters of SBL based on linear model input and output. Additionally, we analyze the estimated parameters utilizing constraints from large systems. Notably, our analysis demonstrates that both SURE estimators yield equivalent outcomes when subjected to large system restrictions. Under the large system limit, the two SURE optimizers yield the same results as the MSE optimizer when the noise tends towards zero. However, discrepancies arise when the noise is significant.

Acknowledgements

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

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] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit ," *IEEE Trans. Inf. Theory*, December 2007.

[5] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit ," *SIAM J. Sci. Comput.*, vol. 20, no. 1, 1998.

[6] D. Wipf and S. Nagarajan, "Iterative reweighted l_1 and l_2 methods for finding sparse solutions ," *IEEE J. Sel. Topics Sig. Process.*, vol. 4, no. 2, April 2010.

[7] 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.

[8] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Mach. Learn. Res.*, vol. 1, 2001.

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

[10] T. Park and G. Casella, “The Bayesian Lasso,” *J. Amer. Statist. Assoc.*, Nov. 2008.

[11] T. Kailath, A.H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice-Hall, 2000.

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

[13] D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing ,” *PNAS*, vol. 106, Nov. 2009.

[14] S. Rangan, “Generalized approximate message passing for estimation with random linear mixing,” in *Proc. IEEE Int. Symp. Inf. Theory*, Saint Petersburg, Russia, August 2011.

[15] S. Rangan, P. Schniter, and A. Fletcher, “On the convergence of approximate message passing with arbitrary matrices,” in *Proc. IEEE Int. Symp. Inf. Theory*, 2014.

[16] S. Rangan, P. Schniter, and A. K. Fletcher, “Vector Approximate Message Passing,” *IEEE Trans. On Info. Theo.*, vol. 65, no. 10, Oct. 2019.

[17] D.T.M. Slock, “Convergent Approximate Message Passing by Alternating Constrained Minimization of Bethe Free Energy,” in *Proc. IEEE 7th Forum on Research and Technologies for Society and Industry Innovation (RTSI)*, Paris, France, Aug. 2022.

[18] D.T.M. Slock, “Convergent Approximate Message Passing,” in *IEEE Int’l Mediterranean Conf. Communications and Networking (MEDITCOM)*, Athens, Greece, Sept. 2022.

[19] M. J. Beal, “Variational Algorithms for Approximate Bayesian Inference,” in *Thesis, Univeristy of Cambridge, UK*, May 2003.

[20] D. Shutin, T. Buchgraber, S. R. Kulkarni, and H. V. Poor, “Fast variational sparse bayesian learning with automatic relevance determination for superimposed signals,” *IEEE Trans. on Sig. Process*, vol. 59, no. 12, December 2011.

[21] X. Tan and J. Li, “Computationally Efficient Sparse Bayesian Learning via Belief Propagation ,” *IEEE Trans. on Sig. Proc.*, vol. 58, no. 4, Apr. 2010.

[22] Ramina Ghods, Charles Jeon, Gulnar Mirza, Arian Maleki, and Christoph Studer, “Optimally-tuned nonparametric linear equalization for massive mu-mimo systems,” in *Proc. IEEE Int. Symp. Inf. Theory*, June 2017.

[23] P. Mattsson, D. Zachariah, and P. Stoica, “Tuned Regularized Estimators for Linear Regression via Covariance Fitting,” *arXiv2201.08756*, 2022.

[24] Y.C. Eldar, “Generalized SURE for Exponential Families: Applications to Regularization,” *IEEE Trans. Sig. Proc.*, Feb. 2009.

[25] 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.

[26] S. Wagner, R. Couillet, M. Debbah, and D. Slock, “Large System Analysis of Linear Precoding in Correlated MISO Broadcast Channels Under Limited Feedback,” *IEEE Trans. on Info. Theo.*, vol. 58, no. 7, pp. 4509–4537, 2012.