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

Dirk Slock
EURECOM, BIOT Sophia-Antipolis, France, Email: slock@eurecom.fr

***Abstract*—Sparse Bayesian Learning (SBL) is a popular compressed sensing technique in which the sparsifying prior for the unknowns in the underdetermined linear system is modeled as a Gaussian scale mixture. This leads to a number of hyperparameters which involve at least the variance profile and the noise variance, but also possible parameters in the variance profile priors. These hyperparameters are typically determined by Type I or Type II Maximum Likelihood (ML) estimation. In this paper we introduce SURE SBL in which the hyperparameter optimization (and not estimation) is based on Stein's Unbiased Risk Estimator (SURE). Indeed the ultimate performance criterion is usually the Mean Squared Error (MSE) of the sparse parameters or the resulting signal model. We review the SURE approach and its use in the world of automatic control. Then we apply the SURE approach to the MSE of the sparse parameters (linear model input) and find that it yields the same hyperparameter optimization as by Type II ML. We finally propose the SURE approach at the level of the output of the linear model, where it leads to new hyperparameter adjustments.**

I. Introduction

Sparse signal reconstruction and compressed sensing (CS) has received an enormous amount of attention in recent years. Some applications include massive multi-input multi-output (MIMO) channel estimation [1], direction of arrival estimation [2], biomagnetic imaging [3], 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, using Bayesian techniques for sparse signal recovery (SSR) generally achieves the best performance. It is worth mentioning that [7] provides a detailed overview of the various SSR algorithms which fall under l_1 or l_2 norm minimization approaches such as Basis Pursuit, LASSO etc., and Sparse Bayesian Learning (SBL) methods. The authors justify the superior recovery performance of SBL compared to the conventional methods mentioned above. The SBL algorithm was first introduced by [8] and then proposed for the first time for SSR by [9]. In a Bayesian setting, the aim is to calculate the posterior distribution of the parameters $\boldsymbol{x}$ given some observations (data) and some a priori knowledge.

Compared to other state of the art techniques, the critical point about SBL is the hierarchical prior modeling which results in sparsification of the state $\boldsymbol{x}$. The Bayesian LASSO [10] uses a similar hierarchical modeling with a Gaussian-Exponential prior (equivalent to a Laplace prior) and it turns out to be a special case of the Student t prior in SBL.

In SBL, the unknown parameters $\boldsymbol{x}$ are modeled as decorrelated zero mean Gaussian[1] $\boldsymbol{x} \sim \mathcal{N}(0, (\text{Diag}(\boldsymbol{\xi}))^{-1})$ 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 $\boldsymbol{\xi}, \gamma$, 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 (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

[1]Notations: The operator $(\cdot)^T$ represents the matrix transpose. The pdf of a Gaussian random variable x with mean μ and variance σ^2 is written as $\mathcal{N}(x; \mu, \nu)$. x_k represents the k^{th} element of the vector $\boldsymbol{x}$. $KL(q||p)$ represents the Kullback-Leibler distance between the two distributions q, p. $\boldsymbol{A}_{n,:}$ represents the n^{th} row of matrix $\boldsymbol{A}$. blkdiag$(\cdot)$ represents the blockdiagonal part of a matrix. diag$(\mathbf{X})$ or Diag$(\boldsymbol{x})$ represents a vector obtained by the diagonal elements of the matrix $\mathbf{X}$ or the diagonal matrix obtained with the elements of $\boldsymbol{x}$ in the diagonal respectively. $\mathbf{1}_M$ represents a vector of length M with all ones as the elements. For a matrix $\boldsymbol{A}$, $\boldsymbol{A} \geq 0$ signifies it is non-negative definite. $\mathbf{I}_M$ represents the identity matrix of size M. tr$\{\boldsymbol{A}\}$ represents the trace of $\boldsymbol{A}$ (sum of diagonal elements). $\boldsymbol{A}_{i,j}$ represents element (i, j) of matrix $\boldsymbol{A}$.

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) involves a matrix inversion step (at each iteration in Type I ML, which is joint estimation of parameters $\boldsymbol{x}$ and hyperparameters), making it computationally complex even for moderately large datasets. An alternative approach to SBL is using variational approximation for Bayesian inference [19]. Variational Bayesian (VB) inference tries to find a factored approximation of the posterior distribution which maximizes the variational lower bound on $\ln\, p(\boldsymbol{y})$. [20] introduces a Fast version of SBL by alternatingly maximizing the variational posterior lower bound with respect to single (hyper)parameters. In [21], the authors introduce a Belief Propagation (BP) based SBL algorithm which turns out to be computationally more efficient. The authors use BP to infer the posterior pdf of $\boldsymbol{x}$ and the hyperparameters are estimated using the EM algorithm. The authors in [22] propose a MP approach combining BP and mean field (MF) approximations. MF is a special case of VB in which the partitioning of variables is pushed to the scalar granularity level. The benefits of the combined scheme can be summarized as follows: While the MF approach always admits a convergent implementation and the low complexity BP yields a good approximation of the posterior marginals if the factor graph has no cycles. The authors show that the MP fixed-point equations for a combination of BP and MF correspond to stationary points of one single constrained region-based free energy approximation and provide a clear rule stating how to couple the messages propagating in the BP and MF parts. Hence, it is advantageous to apply a combination of BP and the MF approximation on the same factor graph to exploit their respective virtues while limiting their drawbacks (MF has lower complexity than BP, but is potentially more suboptimal). However, [22] does not treat at all the topic of how to split nodes between BP and MF.

[23] uses the Approximate Message Passing (AMP) algorithm for LMMSE and introduces a non-parametric algorithm called NOPE that does not require any knowledge of the signal and noise powers (these two parameters are adjusted via SURE actually). The authors also prove that in the large system limit, NOPE achieves the same performance as that of the LMMSE equalizer.

Another approach appears in [24] (and former publications by the same authors), called the SPICE methodology, in which they adjust hyperparameters by covariance fitting using the 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 (20). 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).

II. Empirical Bayes via SURE: State of the Art

In this section, we provide a short overview of some Bayesian estimation schemes which are similar to SBL or of which SBL can be seen as a special case.

A. James-Stein Estimator

The Bayesian likelihood interpretation of (possibly overdetermined) Compressed Sensing can be written as

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 - 2\gamma^{-1}\, \ln p(\boldsymbol{x})\ . \tag{4}$$

Stein and James in their landmark paper [25] showed that for the linear Gaussian model with i.i.d. prior $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \xi^{-1}\mathbf{I})$, it is possible to construct a nonlinear estimate of $\boldsymbol{x}$ with lower (deterministic) MSE than that of ML for all values of the unknown true (deterministic) $\boldsymbol{x}$. This is an instance of (parametric) empirical Bayes in which the hyperparameters in the prior are estimated from the data also. A popular design strategy here is to minimize Stein's Unbiased Risk Estimate (SURE) [26], which is an unbiased estimate of the MSE. SURE directly approximates the MSE of an estimate from the data, without requiring knowledge of the hyperparameters (ξ). Stein's landmark discovery lead to the study of biased estimators that outperform minimum variance unbiased estimators (MVUE) in terms of MSE, see e.g. the work by Yonina Eldar [27]. Shrinkage estimators and penalized maximum likelihood (PML) estimators are examples of this. If the penalty term in PML can be interpreted as a prior log likelihood, then James-Stein is an instance of PML, which considers a Bayesian likelihood for deterministic parameters.

B. More General Covariance Gaussian Prior in Automatic Control

Another approach are the so-called kernel methods in linear system identification for (1) [28]. A good overview of Kernel methods, in connection with machine learning can be found in [29]. Traditional methods in that area are ML or prediction error methods (PEM) which are optimal in the large data limit. The kernel methods are an instance of PML and represent a generalization of James-Stein. The prior considered is $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{P})$ for some symmetric positive semidefinite kernel matrix $\boldsymbol{P}$. leading to

$$\widehat{\boldsymbol{x}} = \arg \min_{\boldsymbol{x}\in\mathcal{R}^M} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \frac{1}{\gamma}\boldsymbol{x}^T\boldsymbol{P}^{-1}\boldsymbol{x}. \tag{5}$$

The kernel $\boldsymbol{P}(\boldsymbol{\eta})$ is a parameterized family of matrices, where $\boldsymbol{\eta} \in \mathcal{R}^p$. $\boldsymbol{\eta}$ are the hyperparameters. Methods for hyperparameter estimation include cross-validation (CV), empirical Bayes (EB), C_p statistics and Stein's unbiased risk estimate (SURE). SBL can be interpreted as a special case with diagonal $\boldsymbol{P}$ with (inverse) diagonal elements $\boldsymbol{\eta}$.

1) Kernel hyperparameter estimation: In this subsection, we provide an overview of a few hyperparameter estimation techniques known in the literature, see [30] for more details.

The first approach is to estimate the hyperparameters using Empirical Bayes (EB = Type II ML):

$$\widehat{\boldsymbol{\eta}}_{EB} = \arg\min_{\boldsymbol{\eta}} f_{EB}(\boldsymbol{P}(\boldsymbol{\eta})),$$
$$f_{EB}(\boldsymbol{P}(\boldsymbol{\eta})) = \boldsymbol{y}^T\boldsymbol{Q}^{-1}\boldsymbol{y} + \ln\det(\boldsymbol{Q})\ ,\ \boldsymbol{Q} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T + \frac{1}{\gamma}\mathbf{I}_N. \tag{6}$$

There exists also two SURE methods, where the estimation problem gets formulated as below.

SURE 1: is based on minimizing the MSE of signal reconstruction ($MSE_x(\boldsymbol{P}) = \mathrm{E}(\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|^2)$):

$$\begin{aligned} SURE_x : \widehat{\boldsymbol{\eta}}_{Sx} &= \arg\min_{\boldsymbol{\eta}} f_{Sx}(\boldsymbol{P}(\boldsymbol{\eta})),\ \text{with} \\ f_{Sx}(\boldsymbol{P}(\boldsymbol{\eta})) &= \frac{1}{\gamma^2}\boldsymbol{y}^T\boldsymbol{Q}^{-T}\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-2}\boldsymbol{A}^T\boldsymbol{Q}^{-1}\boldsymbol{y} \\ &\quad + \frac{1}{\gamma}\mathrm{tr}\{2\boldsymbol{R}^{-1} - (\boldsymbol{A}^T\boldsymbol{A})^{-1}\}, \\ \text{with}\ \ \boldsymbol{R} &= \boldsymbol{A}^T\boldsymbol{A} + \frac{1}{\gamma}\boldsymbol{P}^{-1}. \end{aligned} \tag{7}$$

SURE 2: is based on minimizing the MSE of output prediction ($MSE_y(\boldsymbol{P}) = \mathrm{E}(\|\boldsymbol{A}\widehat{\boldsymbol{x}} + \boldsymbol{w}^* - \boldsymbol{y}\|^2)$), $\boldsymbol{w}^*$ is an independent copy of the noise $\boldsymbol{w}$:

$$\begin{aligned} SURE_y : \widehat{\boldsymbol{\eta}}_{Sy} &= \arg\min_{\boldsymbol{\eta}} f_{Sy}(\boldsymbol{P}(\boldsymbol{\eta})),\ \text{with} \\ f_{Sy}(\boldsymbol{P}(\boldsymbol{\eta})) &= \frac{1}{\gamma^2}\boldsymbol{y}^T\boldsymbol{Q}^{-T}\boldsymbol{Q}^{-1}\boldsymbol{y} + 2\frac{1}{\gamma}\mathrm{tr}\{\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T\boldsymbol{Q}^{-1}\} \end{aligned} \tag{8}$$

The MSE at the level of the sparse expansion coefficients $\boldsymbol{x}$ is perhaps not the most relevant, and neither the MSE at the level of the measurements $\boldsymbol{y}$. In general, the matrix $\boldsymbol{A}$ is the cascade of two matrices $\boldsymbol{A} = \boldsymbol{\Psi}\,\boldsymbol{\Phi}$ where $\boldsymbol{\Psi}$ is the measurement matrix and $\boldsymbol{\Phi}$ is the dictionary in which the representation of the signal $s = \boldsymbol{\Phi}\,\boldsymbol{x}$ becomes sparse. The main MSE of interest is that of the signal s. In this paper we explore the adjustment of the hyperparameters via the SURE MSE of s. We also explore the optimization of another hyperparameter which parametrizes a Generalized Gaussian prior, of which Gaussian and Laplace distributions are special cases. We compare to the standard SBL versions based on ML. It should also be possible to extend the large system analysis of [31] to the new setting.

III. Prior Variance Determination in Fast 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:

$$p_i = |\widehat{x}_i(p_i)|^2 + \sigma^2_{\widehat{x}_i(p_i)} = \lfloor |\widehat{x}_i(0)|^2 - \sigma^2_{\widehat{x}_i(0)} \rfloor_+ \tag{9}$$

where $\widehat{x}_i(p_i)$ and $\sigma^2_{\widehat{x}_i(p_i)}$ are the LMMSE estimate and the corresponding error variance for a priori variance p_i. The first line 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 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. Stein's Unbiased Risk Estimator: SURE Principle

Consider a simple additive white Gaussian noise model:

$$\boldsymbol{y} = \mathbf{z} + \boldsymbol{v} \tag{10}$$

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

$$\begin{aligned} \mathrm{MSE}_{\mathbf{z}} &= \mathrm{E}\,\|\widehat{\mathbf{z}} - \mathbf{z}\|^2 = \mathrm{E}\,\{\|\mathbf{z}\|^2 + \|\widehat{\mathbf{z}}\|^2 - 2\,\widehat{\mathbf{z}}^T\mathbf{z}\} \\ &\overset{(a)}{=} \mathrm{E}\,\{\|\mathbf{z}\|^2 + \|\widehat{\mathbf{z}}\|^2 - 2\,\widehat{\mathbf{z}}^T\boldsymbol{y} + 2\sigma^2\,\mathrm{tr}\{\tfrac{\partial\widehat{\mathbf{z}}^T}{\partial\boldsymbol{y}}\}\,\} \\ &= \mathrm{E}\,\{\|\mathbf{z}\|^2 - \|\boldsymbol{y}\|^2 + \|\widehat{\mathbf{z}} - \boldsymbol{y}\|^2 + 2\sigma^2\,\mathrm{tr}\{\tfrac{\partial\widehat{\mathbf{z}}^T}{\partial\boldsymbol{y}}\}\,\} \end{aligned} \tag{11}$$

where E is w.r.t. $\boldsymbol{v}$ ($\mathbf{z}$ is treated as deterministic) and (a) follows as a property of the Gaussian pdf [32]. 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 $\widehat{\mathbf{z}}$)

$$\begin{aligned} \widehat{\mathrm{MSE}}_{\mathbf{z}} &= \|\mathbf{z}\|^2 - \|\boldsymbol{y}\|^2 + \mathrm{SURE}_{\mathbf{z}}\ , \\ \mathrm{SURE}_{\mathbf{z}} &= \|\widehat{\mathbf{z}} - \boldsymbol{y}\|^2 + 2\sigma^2\,\mathrm{tr}\{\tfrac{\partial\widehat{\mathbf{z}}^T}{\partial\boldsymbol{y}}\}\,. \end{aligned} \tag{12}$$

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 $\boldsymbol{y}$.

V. 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

$$\widehat{x}_i(0) = x_i + \widetilde{x}_i(0) \tag{13}$$

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

$$\widehat{x}_i = \widehat{x}_i(p_i) = \frac{p_i}{p_i + \sigma^2}\,\widehat{x}_i(0) \tag{14}$$

Then we get

$$\begin{aligned} \mathrm{SURE}_{x_i}(p_i) &= \left(\frac{\sigma^2}{p_i + \sigma^2}\,\widehat{x}_i(0)\right)^2 + 2\frac{\sigma^2\,p_i}{p_i + \sigma^2} \\ &= \left(\frac{\sigma^2}{p_i + \sigma^2}\,\widehat{x}_i(0)\right)^2 - 2\frac{\sigma^4}{p_i + \sigma^2} + 2\sigma^2 \end{aligned} \tag{15}$$

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

$$\frac{\partial\mathrm{SURE}_{x_i}}{\partial p_i} = 2\sigma^4(p_i + \sigma^2 - \widehat{x}_i^2(0))/(p_i + \sigma^2)^3. \tag{16}$$

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

$$\frac{\partial\mathrm{SURE}_{x_i}}{\partial p_i}(p_i = 0) = 2(1 - \frac{\widehat{x}_i^2(0)}{\sigma^2}). \tag{17}$$

So, the minimum of $\text{SURE}_{x_i}(p_i)$ 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

$$p_i = \lfloor |\widehat{x}_i(0)|^2 - \sigma^2_{\widehat{x}_i(0)} \rfloor_+ \tag{18}$$

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.

VI. 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:

$$\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{x} + \boldsymbol{v}\;, \boldsymbol{v} \sim \mathcal{N}(\boldsymbol{v}; 0, \sigma^2\mathbf{I})\;,\; \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}; 0, \boldsymbol{P}) \tag{19}$$

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

$$\begin{aligned} \boldsymbol{x}|\boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{x};\, \boldsymbol{P}\boldsymbol{A}^T\boldsymbol{R}^{-1}\boldsymbol{y},\, \boldsymbol{P} - \boldsymbol{P}\boldsymbol{A}^T\boldsymbol{R}^{-1}\boldsymbol{A}\boldsymbol{P}) \\ &\sim \mathcal{N}(\boldsymbol{x};\, \boldsymbol{S}^{-1}\boldsymbol{A}^T\boldsymbol{y},\, \sigma^2\boldsymbol{S}^{-1}) \end{aligned} \tag{20}$$

where $\boldsymbol{R} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T + \sigma^2\mathbf{I}$ is the covariance matrix of $\boldsymbol{y}$ and the expressions with $\boldsymbol{S} = \boldsymbol{A}^T\boldsymbol{A} + \sigma^2\boldsymbol{P}^{-1}$ are valid only if $\boldsymbol{P}$ is non-singular, or by rewriting $\boldsymbol{S}^{-1} = \boldsymbol{P}(\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{P} + \sigma^2\mathbf{I})^{-1} = \boldsymbol{P}^{1/2}(\boldsymbol{P}^{1/2}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{P}^{1/2} + \sigma^2\mathbf{I})^{-1}\boldsymbol{P}^{1/2}$.

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 $\text{diag}(\boldsymbol{P})$). In that case the *estimation of the individual components* x_i and their prior power p_i is indeed important (previous section). The treatment of these components can be adressed jointly via the *Component-Wise Conditionally Unbiased (CWCU-)LMMSE* approach [33] which leads to e.g.

$$\widehat{\boldsymbol{x}}(0) = (\text{diag}(\boldsymbol{S}^{-1}\boldsymbol{A}^T\boldsymbol{A}))^{-1}\boldsymbol{S}^{-1}\boldsymbol{A}^T\boldsymbol{y}. \tag{21}$$

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.*

VII. 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 $\boldsymbol{A} = \mathbf{B}\mathbf{C}$. 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 $\boldsymbol{A}$ ($\mathbf{C}$) could be quite different from a superposition of estimations of the x_i. This SURE application is closely related to the SURE 2 criterion in Automatic Control mentioned earlier.

The estimation in the underdetermined linear model (fat $\boldsymbol{A}$) is related to the case of reduced rank (overdetermined) $\boldsymbol{A}$ discussed in [32].

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

$$\begin{aligned} &\text{SURE}_{\mathbf{z}}(\boldsymbol{P}) = \|\boldsymbol{y} - \widehat{\mathbf{z}}\|^2 + 2\sigma^2\,\text{tr}\{\tfrac{\partial\widehat{\mathbf{z}}^T}{\partial\boldsymbol{y}}\} = \sigma^4\boldsymbol{y}^T\boldsymbol{R}^{-2}\boldsymbol{y} \\ &+2\sigma^2\text{tr}\{\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T\boldsymbol{R}^{-1}\} = \|\boldsymbol{y}\|^2 + \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{S}^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{S}^{-1}\boldsymbol{A}^T\boldsymbol{y} \\ &-2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{S}^{-1}\boldsymbol{A}^T\boldsymbol{y} + 2\sigma^2\text{tr}\{\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{A}^T\} \end{aligned} \tag{22}$$

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

$$\begin{aligned} &\boldsymbol{S} = \boldsymbol{S}_{\bar{i}} + \tfrac{\sigma^2}{p_i}\boldsymbol{e}_i\boldsymbol{e}_i^T \\ &\Rightarrow\; \boldsymbol{S}^{-1} = \boldsymbol{S}_{\bar{i}}^{-1} - \boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i(\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i + \tfrac{p_i}{\sigma^2})^{-1}\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\,, \\ &\boldsymbol{S}^{-1}\boldsymbol{e}_i = \boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i(\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i + \tfrac{p_i}{\sigma^2})^{-1}\tfrac{p_i}{\sigma^2} \end{aligned} \tag{23}$$

Note that $\widehat{x}_i(0) = (\boldsymbol{e}_i^T\boldsymbol{S}^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{e}_i)^{-1}\boldsymbol{e}_i^T\boldsymbol{S}^{-1}\boldsymbol{A}^T\boldsymbol{y} = (\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{e}_i)^{-1}\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\boldsymbol{y}$.

Then we get

$$\begin{aligned} &\text{SURE}_{\mathbf{z}}(p_i; \boldsymbol{P}_{\bar{i}}) = \|\boldsymbol{y}\|^2 + \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{S}^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{S}^{-1}\boldsymbol{A}^T\boldsymbol{y} \\ &-2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{S}^{-1}\boldsymbol{A}^T\boldsymbol{y} + 2\sigma^2\text{tr}\{\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{A}^T\} \\ &= \|\boldsymbol{y}\|^2 - 2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\boldsymbol{y} \\ &+2(\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i + \tfrac{p_i}{\sigma^2})^{-1}(\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\boldsymbol{y})^2 \\ &-2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i(\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i + \tfrac{p_i}{\sigma^2})^{-1}(\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\boldsymbol{y}) \\ &+(\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i + \tfrac{p_i}{\sigma^2})^{-2}(\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\boldsymbol{y})^2\|\boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i\|^2 \\ &+2\sigma^2\text{tr}\{\boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\} - 2\sigma^2(\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i + \tfrac{p_i}{\sigma^2})^{-1}\|\boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i\|^2 \\ &= \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\boldsymbol{y}\|^2 + 2\,\sigma^2\text{tr}\{\boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\} \\ &-2\,\sigma^2(\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i + \tfrac{p_i}{\sigma^2})^{-1} \\ &(\|\boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i\|^2 - \boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{P}_{\bar{i}}^{-1}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i) \\ &\quad +(\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i + \tfrac{p_i}{\sigma^2})^{-2}(\boldsymbol{e}_i^T\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{A}^T\boldsymbol{y})^2\|\boldsymbol{A}\boldsymbol{S}_{\bar{i}}^{-1}\boldsymbol{e}_i\|^2 \\ &= \;a - \frac{2\,b}{c + p_i/\sigma^2} + \frac{d}{(c + p_i/\sigma^2)^2} \end{aligned} \tag{24}$$

where $\boldsymbol{P}_{\bar{i}}^{-1}$ should be interpreted as $(\boldsymbol{P}^{-1})_{\bar{i}}$ (hence with 0 in diagonal position i).

$\text{SURE}_{\mathbf{z}}(p_i; \boldsymbol{P}_{\bar{i}})$ is of similar structure as $\text{SURE}_{x_i}(p_i; \boldsymbol{P}_{\bar{i}})$. Hence we get

$$p_i = \sigma^2\lfloor\frac{d}{b} - c\rfloor_+\;. \tag{25}$$

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

Acknowledgements

EURECOM's research is partially supported by its industrial members: ORANGE, BMW, SAP, iABG, Norton

LifeLock, and by the projects EEMW4FIX and CellFree6G (French ANR), and 5G-OPERA (Franco-German 5G ecosystem project). The author would also like to acknowledge the early contributions of Dr. Christo Kurisummoottil Thomas to this work.

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] E. Riegler, G. E. Kirkelund, C. N. Manchón, and B. H. Fleury, "Merging Belief Propagation and the Mean Field Approximation: a Free Energy Approach ," *IEEE Trans. on Info. Theo.*, vol. 59, no. 1, Jan. 2013.

[23] 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.

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

[25] 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.

[26] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, Nov. 1981.

[27] Y. C. Eldar, "Rethinking Biased Estimation: Improving Maximum Likelihood and the Cramër-Rao Bound," in *Found. and Tren. in Sig. Process.*, 2008.

[28] G. Pillonetto and G. D. Nicolao, "A New Kernel-based Approach for Linear System Identification," *Automatica*, 2010.

[29] G. Pillonetto, F. Dinuzzo, T. Chen, G. D Nicolao, and L. Ljung, "Kernel Methods in System Identification, Machine Learning and Function Estimation: A Survey," *Automatica*, Feb. 2014.

[30] B. Mu, T. Chen, and L.Ljung, "On Asymptotic Properties of Hyperparameter Estimators for Kernel-based Regularization Methods," *Automatica*, May 2018.

[31] C. K. Thomas and D. Slock, "Posterior variance predictions in sparse Bayesian learning under approximate inference techniques," in *IEEE Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2020.

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

[33] 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.