Performance Analysis of Hyperparameter Optimization in Sparse Bayesian Learning via 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) is a widely-used framework for sparse signal reconstruction, yet its standard formulation optimizes model evidence rather than directly minimizing reconstruction error. In this paper, we reinterpret standard SBL as an approximate scheme for minimizing the mean squared error (MSE) in the input domain. Motivated by this insight, we derive novel hyperparameter update rules aimed at minimizing input-space MSE, and discuss the limitations of using Stein's unbiased risk estimate in underdetermined systems. To address this issue, we propose an alternative risk minimization framework based on output-space MSE, which admits an unbiased estimator. We derive closed-form coordinate-wise update rules for the regularization parameters and analyze their sparsity-promoting behavior. In particular, we identify a sufficient condition—termed the statistical orthogonality condition (SOC)—under which certain components are optimally pruned. This connects our framework to classical sparse recovery criteria. While our analysis sheds light on the emergence of sparsity via risk-based optimization, it also highlights open questions regarding the conditions under which SOC is satisfied, warranting further investigation.**

I. Introduction

Sparse signal reconstruction (SSR) and compressed sensing (CS) have attracted broad interest in recent years due to their wide applicability across signal processing, communications, and machine learning [1]–[5]. These problems are often modeled as:

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{v}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^M$ is the observation vector, $\mathbf{A} \in \mathbb{R}^{M\times N}$ is the known sensing matrix, $\mathbf{x}_0 \in \mathbb{R}^N$ is the unknown sparse signal with $K \ll N$ nonzero entries, and $\mathbf{v}$ is additive white Gaussian noise. Among the various Bayesian approaches to SSR, Sparse Bayesian Learning (SBL) [6]–[8] has emerged as a powerful and flexible framework. In SBL, $\mathbf{x}_0$ is modeled as a zero-mean Gaussian random vector with a diagonal covariance $\mathbf{P}$, and both the hyperparameters $\mathbf{P}$ and signal $\mathbf{x}_0$ are estimated in a hierarchical Bayesian setting. This estimation is typically achieved via evidence maximization, also known as Type II Maximum Likelihood (ML) or Empirical Bayes (EB) [9], with further acceleration provided by methods like Fast Marginalized Likelihood (FMML) [10].

Although effective in practice, the objective function in SBL is not directly aligned with the mean squared error (MSE)—a natural criterion in signal estimation tasks. In our previous work [11]–[13], we proposed SURE-SBL, which leverages Stein's Unbiased Risk Estimator (SURE) [14] to guide hyperparameter optimization with the goal of minimizing the MSE. This approach bridges the gap between Bayesian model evidence and risk-based performance measures. However, due to the presence of the unknown noise vector $\mathbf{v}$, the analysis of the MSE behavior under SURE-guided optimization remains analytically challenging.

This paper presents a deeper investigation of SBL from an MSE-centric perspective. We first show that standard SBL can be interpreted as an approximate algorithm for minimizing the input-space MSE, $\text{MSE}_{\mathbf{x}} := \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}_0\|^2$. Motivated by this insight, we derive alternative update rules aimed at more explicitly minimizing $\text{MSE}_{\mathbf{x}}$. However, due to the underdetermined nature of most SSR problems, Stein-based estimates of $\text{MSE}_{\mathbf{x}}$ (denoted $\text{SURE}_{\mathbf{x}}$) are biased and unreliable. To circumvent this limitation, we instead consider the output-space MSE, defined as $\text{MSE}_{\mathbf{z}} := \mathbb{E}\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{A}\mathbf{x}_0\|^2$, which admits an unbiased estimator $\text{SURE}_{\mathbf{z}}$ that can be evaluated from the observed data.

By minimizing $\text{SURE}_{\mathbf{z}}$, we derive a coordinate-wise update rule for the hyperparameters, which we show can lead to sparsity-promoting behavior in the recovered signal. In particular, we characterize the conditions under which certain hyperparameters diverge to infinity—effectively pruning their corresponding coefficients from the model. These sparsity patterns emerge naturally without explicitly enforcing ℓ_1 penalties or hard sparsity constraints. Moreover, we provide an analysis of when and why these sparse solutions arise, based on a tradeoff between data fidelity and model complexity. This includes connections to conditions such as statistical orthogonality between columns of $\mathbf{A}$. While our analysis explains part of the sparsity-inducing mechanism, a complete theoretical understanding of this behavior remains an open direction for future work.

II. MSE-Based Hyperparameter Optimization

A. Posterior Estimator and Preliminaries

For estimating $\mathbf{x}_0$, SBL assumes that each element x_i of $\mathbf{x}$ 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:

$$p(x_i; p_i) = \mathcal{N}(x_i; 0, p_i), \quad i = 1, \cdots, N; \tag{2}$$

where p_i is an unknown Gaussian variance optimized through the SBL algorithm.

Under the assumed Gaussian prior, the posterior mean estimator becomes:

$$\hat{\mathbf{x}} = \left(\mathbf{A}^\top \mathbf{A} + \mathbf{\Lambda}\right)^{-1} \mathbf{A}^\top \mathbf{y}, \quad \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \tag{3}$$

which corresponds to a coordinate-wise weighted ridge estimator, where each component of $\mathbf{x}$ is regularized with its own ℓ_2 penalty $\lambda_i = \sigma_v^2 / p_i$. Let $\mathbf{H} = \mathbf{A}^\top \mathbf{A} + \mathbf{\Lambda}$. Since σ_v^2 is known, optimizing λ_i and optimizing p_i are actually equivalent.

To isolate the effect of λ_i, we define

$$\mathbf{H}_{\bar{i}} = \mathbf{A}^\top \mathbf{A} + \sum_{j \neq i} \lambda_j \mathbf{e}_j \mathbf{e}_j^\top, \tag{4}$$

so that $\mathbf{H} = \mathbf{H}_{\bar{i}} + \lambda_i \mathbf{e}_i \mathbf{e}_i^\top$.

Applying the Sherman–Morrison formula yields

$$\mathbf{H}^{-1} = \mathbf{H}_{\bar{i}}^{-1} - \frac{\lambda_i}{1 + \lambda_i \alpha_i} \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1}, \tag{5}$$

where $\alpha_i = \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i$. By replace $\mathbf{y}$ by $\mathbf{A}\mathbf{x}_0 + \mathbf{v}$, the estimator $\hat{\mathbf{x}}$then decomposes as

$$\hat{\mathbf{x}} = \mathbf{H}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{x}_0 + \mathbf{H}^{-1} \mathbf{A}^\top \mathbf{v}. \tag{6}$$

B. Input-Space MSE$_x$

The mean squared error (MSE) with respect to $\mathbf{x}_0$ is given by

$$\mathrm{MSE}_x := \mathbb{E}\left[\|\hat{\mathbf{x}} - \mathbf{x}_0\|^2\right] = \mathrm{Bias}^2 + \mathrm{Var}, \tag{7}$$

where

$$\mathrm{Bias}^2 = \|(\mathbf{H}^{-1} \mathbf{A}^\top \mathbf{A} - \mathbf{I}) \mathbf{x}_0\|^2, \tag{8}$$

$$\mathrm{Var} = \sigma^2 \,\mathrm{Tr}(\mathbf{H}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{H}^{-1}). \tag{9}$$

Substituting the Sherman–Morrison identity into the bias term yields

$$\begin{aligned} \mathrm{Bias}^2 &= \|(\mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} - \mathbf{I}) \mathbf{x}_0\|^2 \\ &\quad - \frac{2\lambda_i}{1 + \lambda_i \alpha_i} \mathbf{x}_0^\top (\mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} - \mathbf{I})^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{x}_0 \\ &\quad + \left(\frac{\lambda_i}{1 + \lambda_i \alpha_i}\right)^2 \left(\mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{x}_0\right)^2 \cdot \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-2} \mathbf{e}_i. \end{aligned} \tag{10}$$

Similarly, the variance term becomes

$$\begin{aligned} \mathrm{Var} &= \sigma^2 \,\mathrm{Tr}(\mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{H}_{\bar{i}}^{-1}) \\ &\quad - \frac{2\lambda_i \sigma^2}{1 + \lambda_i \alpha_i} \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{H}_{\bar{i}}^{-2} \mathbf{e}_i \\ &\quad + \sigma^2 \left(\frac{\lambda_i}{1 + \lambda_i \alpha_i}\right)^2 \|\mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i\|^2 \cdot \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-2} \mathbf{e}_i. \end{aligned} \tag{11}$$

Combining the two parts, we obtain

$$\mathrm{MSE}_x(\lambda_i) = \mathrm{const} - \frac{2\lambda_i}{1 + \lambda_i \alpha_i} T_1^{(x)} + \left(\frac{\lambda_i}{1 + \lambda_i \alpha_i}\right)^2 T_2^{(x)}, \tag{12}$$

where

$$\begin{aligned} T_1^{(x)} &:= \mathbf{x}_0^\top (\mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} - \mathbf{I})^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{x}_0 \\ &\quad + \sigma^2 \cdot \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{H}_{\bar{i}}^{-2} \mathbf{e}_i, \end{aligned} \tag{13}$$

$$\begin{aligned} T_2^{(x)} &:= \left(\mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{x}_0\right)^2 \cdot \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-2} \mathbf{e}_i \\ &\quad + \sigma^2 \|\mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i\|^2 \cdot \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-2} \mathbf{e}_i. \end{aligned} \tag{14}$$

C. Output-Space MSE$_z$

We now consider the MSE measured in the output space:

$$\mathrm{MSE}_z := \mathbb{E}\left[\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{A}\mathbf{x}_0\|^2\right] = \mathrm{Bias}^2 + \mathrm{Var}, \tag{15}$$

where

$$\mathrm{Bias}^2 = \|\mathbf{A}(\mathbf{H}^{-1} \mathbf{A}^\top \mathbf{A} - \mathbf{I}) \mathbf{x}_0\|^2, \tag{16}$$

$$\mathrm{Var} = \sigma^2 \,\mathrm{Tr}(\mathbf{A} \mathbf{H}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{H}^{-1} \mathbf{A}^\top). \tag{17}$$

The squared bias term expands as:

$$\begin{aligned} \mathrm{Bias}^2 &= \left\|\mathbf{A}(\mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} - \mathbf{I}) \mathbf{x}_0\right\|^2 \\ &\quad - \frac{2\lambda_i}{1 + \lambda_i \alpha_i} \cdot \mathbf{x}_0^\top (\mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} - \mathbf{I})^\top \mathbf{A}^\top \mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i \\ &\quad \cdot \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{x}_0 \\ &\quad + \left(\frac{\lambda_i}{1 + \lambda_i \alpha_i}\right)^2 \left(\mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{x}_0\right)^2 \cdot \|\mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i\|^2. \end{aligned} \tag{18}$$

The variance term becomes:

$$\begin{aligned} \mathrm{Var} &= \sigma^2 \,\mathrm{Tr}(\mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top) \\ &\quad - \frac{2\lambda_i \sigma^2}{1 + \lambda_i \alpha_i} \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i \\ &\quad + \sigma^2 \left(\frac{\lambda_i}{1 + \lambda_i \alpha_i}\right)^2 \|\mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i\|^4. \end{aligned} \tag{19}$$

Combining bias and variance, we obtain:

$$\mathrm{MSE}_z(\lambda_i) = \mathrm{const} - \frac{2\lambda_i}{1 + \lambda_i \alpha_i} T_1^{(z)} + \left(\frac{\lambda_i}{1 + \lambda_i \alpha_i}\right)^2 T_2^{(z)}, \tag{20}$$

where:

$$\begin{aligned} T_1^{(z)} &:= \mathbf{x}_0^\top \mathbf{A}^\top (\mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top - \mathbf{I}) \mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{x}_0 \\ &\quad + \sigma^2 \cdot \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i, \end{aligned} \tag{21}$$

$$\begin{aligned} T_2^{(z)} &:= \left(\mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{A}^\top \mathbf{A} \mathbf{x}_0\right)^2 \cdot \|\mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i\|^2 \\ &\quad + \sigma^2 \cdot \|\mathbf{A} \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i\|^4. \end{aligned} \tag{22}$$

D. Hyperparameter Optimization

While the closed-form expression for the optimal λ_i^* derived earlier is exact, it is algebraically nontrivial due to the rational dependence on λ_i. To gain further insight, we consider an alternative derivation based on a change of variables that transforms the original objective into a simple quadratic form.

Recall that the MSE objective for either input-space in (12) or output-space (20) can be expressed generically as:

$$\mathrm{MSE}(\lambda_i) = \mathrm{const} - \frac{2\lambda_i}{1 + \lambda_i \alpha_i} T_1 + \left(\frac{\lambda_i}{1 + \lambda_i \alpha_i}\right)^2 T_2, \tag{23}$$

where T_1, T_2 are problem-dependent constants and $\alpha_i := \mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1} \mathbf{e}_i > 0$.

Define a transformed variable:

$$\theta := \frac{\lambda_i}{1 + \lambda_i \alpha_i}, \tag{24}$$

which is strictly increasing in λ_i over the interval $\lambda_i \in [0, \infty)$. This change is invertible:

$$\lambda_i = \frac{\theta}{1 - \alpha_i \theta}, \quad \text{valid for } \theta \in [0, 1/\alpha_i). \tag{25}$$

Substituting into the objective, we obtain:

$$\text{MSE}(\theta) = -2T_1\theta + T_2\theta^2 + \text{const}, \tag{26}$$

a convex quadratic function over the feasible interval $[0, 1/\alpha_i)$. The unconstrained minimizer is:

$$\theta^* = \frac{T_1}{T_2}. \tag{27}$$

We now consider three cases based on the sign and magnitude of T_1 and T_2:

Case 1: $T_1 \leq 0$: In this case, $\theta^* \leq 0$ lies outside the feasible domain. Since MSE(θ) is strictly increasing on $[0, 1/\alpha_i)$, the minimum is attained at $\theta = 0$, which corresponds to $\lambda_i^* = 0$.

Case 2: $T_1 > 0$, $T_2 > \alpha_i T_1$: Then $\theta^* = T_1/T_2 \in (0, 1/\alpha_i)$ lies within the feasible region, and the optimal regularization parameter is given by:

$$\lambda_i^* = \frac{\theta^*}{1 - \alpha_i \theta^*} = \frac{T_1}{T_2 - \alpha_i T_1}. \tag{28}$$

Case 3: $T_1 > 0$, $T_2 \leq \alpha_i T_1$: In this case, $\theta^* \geq 1/\alpha_i$ is infeasible. The function MSE(θ) is decreasing throughout the feasible domain, and the minimum is asymptotically attained as $\theta \to 1/\alpha_i^-$, which corresponds to $\lambda_i \to \infty$.

Therefore, the closed-form optimal λ_i^*, valid in both input- and output-space MSE formulations, is given by:

$$\lambda_i^* = \begin{cases} 0, & \text{if } T_1 \leq 0, \\ \dfrac{T_1}{T_2 - \alpha_i T_1}, & \text{if } T_1 > 0,\ T_2 > \alpha_i T_1, \\ \infty, & \text{if } T_1 > 0,\ T_2 \leq \alpha_i T_1. \end{cases} \tag{29}$$

This equivalent derivation not only simplifies the optimization process but also provides a clearer understanding of how regularization strength λ_i is governed by the balance between signal and noise contributions, encoded via T_1, T_2, and α_i.

III. SURE-Based Hyperparameter Optimization in the Input Domain

Since the system is underdetermined, it is not feasible to use SURE$_x$ directly as an unbiased proxy for the input-space MSE (MSE$_x$), due to the presence of components in $\mathbf{x}$ that are not recoverable from the observations $\mathbf{y}$. To address this issue, we investigate several alternative strategies: (i) using a component-wise SURE formulation by treating the rest of $\mathbf{x}$ as Gaussian noise; and (ii) approximating the analytical expressions $T_1^{(x)}$ and $T_2^{(x)}$ from observable quantities.

A. Component-Wise Estimation via SURE$_{x_i}$

We first derive a Stein Unbiased Risk Estimate (SURE) for estimating each individual component x_i under a scalar Gaussian observation model. Starting from the model:

$$\mathbf{y} = \mathbf{A}_i x_i + \sum_{j \neq i} \mathbf{A}_j x_j + \mathbf{v}, \tag{30}$$

we assume that the non-target components x_j $(j \neq i)$ are modeled as random variables with mean estimates $\hat{x}_j$ and zero-mean fluctuations $\tilde{x}_j$:

$$x_j = \hat{x}_j + \tilde{x}_j, \quad \tilde{x}_j \sim \mathcal{N}(0, \sigma_{\tilde{x}_j}^2). \tag{31}$$

Substituting into (30) gives:

$$\mathbf{y} - \sum_{j \neq i} \mathbf{A}_j \hat{x}_j = \mathbf{A}_i x_i + \sum_{j \neq i} \mathbf{A}_j \tilde{x}_j + \mathbf{v}. \tag{32}$$

This results in an effective scalar model:

$$r_i = x_i + w_i, \tag{33}$$

where w_i is an effective Gaussian noise with variance:

$$\sigma_{w_i}^2 = (\mathbf{A}_i^\top \mathbf{C}_i^{-1} \mathbf{A}_i)^{-1}, \tag{34}$$

$$r_i = \sigma_{w_i}^2 \mathbf{A}_i^\top \mathbf{C}_i^{-1} \mathbf{y}, \tag{35}$$

and

$$\mathbf{C}_i = \sum_{j \neq i} p_j \mathbf{A}_j \mathbf{A}_j^\top + \sigma_v^2 \mathbf{I}. \tag{36}$$

Under this model, the component-wise SURE for x_i is:

$$\text{SURE}_{x_i}(p_i) = \left(\frac{\sigma_{w_i}^2 r_i}{\sigma_{w_i}^2 + p_i} \right)^2 + 2 \frac{\sigma_{w_i}^2 p_i}{\sigma_{w_i}^2 + p_i}. \tag{37}$$

Differentiating with respect to p_i yields:

$$\frac{d}{dp_i} \text{SURE}_{x_i}(p_i) = \frac{2\sigma_{w_i}^4 (p_i + \sigma_{w_i}^2 - r_i^2)}{(p_i + \sigma_{w_i}^2)^3}, \tag{38}$$

which leads to the following update rule:

$$\hat{p}_i = \max(r_i^2 - \sigma_{w_i}^2, 0). \tag{39}$$

This form corresponds to the Type-II Maximum Likelihood (ML) update used in classical Sparse Bayesian Learning (SBL), and shows that SBL implicitly assumes the other coefficients are Gaussian-distributed, thereby avoiding the underdetermination issue.

B. Coupled Optimization via Sum of Component-Wise SUREs

The above update only optimizes SURE$_{x_i}$ with respect to p_i. However, since $\mathbf{C}_j$ (and thus SURE$_{x_j}$) also depends on p_i for $j \neq i$, a better approach is to jointly optimize:

$$p_i^* = \arg\min_{p_i} \text{SURE}_{x_i}(p_i) + \sum_{j \neq i} \text{SURE}_{x_j}(p_i). \tag{40}$$

Letting $\gamma_j = \mathbf{A}_j^\top \mathbf{C}_i^{-1} \mathbf{A}_j$, $z_j = \mathbf{A}_j^\top \mathbf{C}_i^{-1} \mathbf{y}$, we define:

$$\text{SURE}_{x_j}(p_i) = \left(\frac{\left[p_i + \sigma_{w_i}^2 (1 - p_j \gamma_j) \right] z_j}{\gamma_j (p_i + \sigma_{w_i}^2)} \right)^2 + \frac{2 p_j (p_i + \sigma_{w_i}^2 (1 - p_j \gamma_j))}{p_i + \sigma_{w_i}^2}. \tag{41}$$

Differentiating gives:

$$\frac{d}{dp_i}\text{SURE}_{x_j}(p_i) = \frac{1}{(p_i+\sigma_{w_i}^2)^3}\cdot(p_i C_{1,ij} + C_{2,ij}), \quad (42)$$

with:

$$C_{1,ij} = 2\sigma_{w_i}^2\left(p_j\cdot\frac{z_j^2}{\gamma_j} + p_j^2\gamma_j\right), \quad (43)$$

$$C_{2,ij} = 2\sigma_{w_i}^4\left(p_j\cdot\frac{z_j^2}{\gamma_j}(1-p_j\gamma_j) + p_j^2\gamma_j\right). \quad (44)$$

Combining with SURE_{x_i} yields the full gradient condition:

$$p_i(2\sigma_{w_i}^4 + \sum_{j\neq i} C_{1,ij}) = 2\sigma_{w_i}^4(r_i^2 - \sigma_{w_i}^2) + \sum_{j\neq i} C_{2,ij}, \quad (45)$$

leading to:

$$\hat{p}_i = \max\left(\frac{\sigma_{w_i}^2(r_i^2-\sigma_{w_i}^2) - \sum_{j\neq i} C_{2,ij}}{\sigma_{w_i}^2 + \sum_{j\neq i} C_{1,ij}}, 0\right). \quad (46)$$

C. Approximate Evaluation of $T_1^{(x)}$ and $T_2^{(x)}$

The input-space MSE in (12) depends on the quantities $T_1^{(x)}$ and $T_2^{(x)}$. We now show that they can be expressed using expectations over $\mathbf{y}$ and approximate posterior statistics:

$$T_1^{(x)} = \mathbb{E}_{\mathbf{v}}\left[\mathbf{e}_i^\top \mathbf{H}_{\bar{i}}^{-1}\mathbf{A}^\top\mathbf{y}\cdot\mathbf{e}_i^\top\mathbf{H}_{\bar{i}}^{-2}\mathbf{A}^\top\mathbf{y}\right] - \mathbf{e}_i^\top\mathbf{H}_{\bar{i}}^{-1}\mathbf{A}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top\mathbf{H}_{\bar{i}}^{-1}\mathbf{e}_i, \quad (47)$$

$$T_2^{(x)} = \mathbb{E}_{\mathbf{v}}\left[\left(\mathbf{e}_i^\top\mathbf{H}_{\bar{i}}^{-1}\mathbf{A}^\top\mathbf{y}\right)^2\cdot\mathbf{e}_i^\top\mathbf{H}_{\bar{i}}^{-2}\mathbf{e}_i\right]. \quad (48)$$

We propose two plug-in approximations for the covariance term $\mathbf{x}\mathbf{x}^\top$:

1) Using posterior mean:

$$\mathbf{x}\mathbf{x}^\top \approx \hat{\mathbf{x}}^{(t)}(\hat{\mathbf{x}}^{(t)})^\top, \quad (49)$$

where $\hat{\mathbf{x}}^{(t)} = \mathbf{H}^{-1}\mathbf{A}^\top\mathbf{y}$.

2) Using posterior second moment:

$$\mathbf{x}\mathbf{x}^\top \approx \hat{\mathbf{x}}^{(t)}(\hat{\mathbf{x}}^{(t)})^\top + \sigma^2\mathbf{H}^{-1}. \quad (50)$$

These substitutions allow us to form data-driven approximations to $T_1^{(x)}$ and $T_2^{(x)}$ and thus approximate MSE_x without needing the true $\mathbf{x}_0$.

D. Discussion and Open Questions

The derivations above provide multiple approaches for estimating the hyperparameters p_i via SURE_x-based strategies. However, it remains unclear which method consistently yields the best empirical or theoretical performance.

The simple component-wise update in (39) is attractive due to its efficiency and intuitive Bayesian justification—it mirrors the update in classical SBL, which assumes Gaussianity of the remaining components. Nevertheless, this assumption does not hold in truly sparse settings, and may lead to suboptimal estimates when significant structure exists in $\mathbf{x}_0$.

The coupled optimization approach, which considers the influence of p_i on all SURE_{x_j} for $j \neq i$, offers a more principled alternative but increases computational complexity and remains sensitive to the choice of prior estimates. Furthermore, both formulations still rely on Gaussian assumptions to sidestep the underdetermined nature of the problem, which may limit their applicability in practice.

The third route—approximating $T_1^{(x)}$ and $T_2^{(x)}$—offers a direct connection to the true MSE objective. While appealing in theory, its accuracy depends critically on the quality of the posterior statistics used for substitution. Whether the posterior mean alone suffices, or if second-order corrections are essential, remains an open question.

We believe a systematic comparison of these estimators, both theoretically (e.g., via bias-variance trade-offs) and empirically (e.g., in signal recovery benchmarks), is warranted. Moreover, extending these ideas beyond Gaussian assumptions or incorporating structure (e.g., block sparsity or correlated priors) could further enhance their practical utility.

IV. Risk Estimation via $\text{SURE}_{\mathbf{z}}$

A. Stein Risk Estimator in Output Space

We now derive an unbiased risk estimate for the output $\hat{\mathbf{z}} = \mathbf{A}\hat{\mathbf{x}}$, based on the SURE framework. Define the estimator

$$\hat{\mathbf{x}} = \mathbf{H}^{-1}\mathbf{A}^\top\mathbf{y}, \quad \hat{\mathbf{z}} = \mathbf{A}\hat{\mathbf{x}}, \quad \mathbf{H} = \mathbf{A}^\top\mathbf{A} + \mathbf{\Lambda}. \quad (51)$$

Then the Stein unbiased risk estimate of $\mathbb{E}[\|\hat{\mathbf{z}} - \mathbf{z}\|^2]$ is given by

$$\begin{aligned}\text{SURE}_{\mathbf{z}} &= \|\hat{\mathbf{z}} - \mathbf{y}\|^2 - M\sigma^2 + 2\sigma^2\cdot\text{Tr}\left(\frac{\partial\hat{\mathbf{z}}}{\partial\mathbf{y}}\right)\\ &= \|\mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top\mathbf{y} - \mathbf{y}\|^2 - M\sigma^2 + 2\sigma^2\cdot\text{Tr}(\mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top).\end{aligned} \quad (52)$$

Define a transformed variable $\beta_i := \frac{\lambda_i}{1+\lambda_i\alpha_i}$, which yields a quadratic SURE form:

$$\text{SURE}_{\mathbf{z}}(\beta_i) = \text{const} - 2C_1^{(z)}\beta_i + C_2^{(z)}\beta_i^2, \quad (53)$$

where

$$C_1^{(z)} := \mathbf{y}^\top(\mathbf{A}\mathbf{H}_{\bar{i}}^{-1}\mathbf{A}^\top - \mathbf{I})^\top\mathbf{A}\mathbf{H}_{\bar{i}}^{-1}\mathbf{e}_i\cdot\mathbf{e}_i^\top\mathbf{H}_{\bar{i}}^{-1}\mathbf{A}^\top\mathbf{y} + \sigma^2\|\mathbf{A}\mathbf{H}_{\bar{i}}^{-1}\mathbf{e}_i\|^2, \quad (54)$$

$$C_2^{(z)} := \|\mathbf{A}\mathbf{H}_{\bar{i}}^{-1}\mathbf{e}_i\|^2\cdot(\mathbf{e}_i^\top\mathbf{H}_{\bar{i}}^{-1}\mathbf{A}^\top\mathbf{y})^2. \quad (55)$$

This gives rise to the same thresholding rule as in the MSE case:

$$\lambda_i^* = \begin{cases}0, & \text{if } C_1^{(z)} \leq 0,\\ \frac{C_1^{(z)}}{C_2^{(z)} - \alpha_i C_1^{(z)}}, & \text{if } C_2^{(z)} > \alpha_i C_1^{(z)},\\ \infty, & \text{otherwise.}\end{cases} \quad (56)$$

B. Interpretation of Key Quantities

This coordinate-wise decision rule is controlled by three interpretable scalars:

(i) α_i (Stability/Variance Term): Represents the posterior variance of x_i if all other λ_j are fixed. A large α_i suggests that the i-th coordinate is weakly identifiable from the data.

(ii) $C_2^{(z)}$ (Distortion/Cost Term): Quantifies the increase in output energy caused by retaining $\hat{x}_i$, combining squared bias and variance of $\hat{x}_i$ in the projected space.

(iii) $C_1^{(z)}$ (Error Reduction Term): Represents the amount by which prediction error can be reduced by allowing $\hat{x}_i$ to participate in $\hat{\mathbf{z}}$. If $C_1^{(z)}$ is small or negative, including x_i contributes little and should be pruned.

These terms together form an implicit sparsity mechanism: coordinates with small $C_1^{(z)}$ or large $C_2^{(z)}$ tend to be suppressed by large λ_i^*.

C. Connection to Sparse Recovery via Statistical Orthogonality

Suppose $x_i = 0$, i.e., the true signal does not contain this component. Denote the residual prediction error excluding $\hat{x}_i$ as:

$$\mathbf{r}_{\bar{i}} = \sum_{j \neq i} \mathbf{A}_j (\hat{x}_j - x_j). \tag{57}$$

Then, minimizing $\text{MSE}_{\mathbf{z}}$ w.r.t. $\hat{x}_i$ leads to:

$$\hat{x}_i^* = -\frac{\mathbb{E}[\langle \mathbf{A}_i, \mathbf{r}_{\bar{i}} \rangle]}{\|\mathbf{A}_i\|^2}. \tag{58}$$

Thus, the sufficient condition for $\hat{x}_i^* = 0$ is:

$$\mathbb{E}[\langle \mathbf{A}_i, \mathbf{r}_{\bar{i}} \rangle] = 0. \tag{59}$$

We refer to this as the *statistical orthogonality condition* (SOC), as it ensures that the residual signal is, on average, uncorrelated with the component direction $\mathbf{A}_i$. When SOC holds, it is optimal to set $\lambda_i^* \to \infty$, effectively pruning the i-th coordinate and enforcing $\hat{x}_i = 0$.

This insight provides a theoretical foundation for sparsity promotion in risk-based estimators: pruning occurs not merely because $\hat{x}_i$ is small, but because the component $\mathbf{A}_i$ fails to align—statistically—with the residual signal generated by other coordinates.

It is important to emphasize, however, that while $\lambda_i \to \infty$ implies $\hat{x}_i = 0$ and hence removes any contribution of $\mathbf{A}_i$ to the output prediction, this alone does not guarantee SOC. Indeed, once $\hat{x}_i = 0$, the residual $\mathbf{r}_{\bar{i}}$ becomes independent of $\hat{x}_i$, and may still have nonzero statistical correlation with $\mathbf{A}_i$. Therefore, SOC is not a consequence of setting $\lambda_i \to \infty$; rather, it must *precede and justify* this pruning decision.

This subtlety reveals an interesting open question: under what structural or statistical conditions on the sensing matrix $\mathbf{A}$ and signal $\mathbf{x}_0$ does SOC naturally emerge? In the context of sparse recovery, this relates closely to classical incoherence and null-space conditions that guarantee uniqueness of sparse solutions. A deeper investigation into how SOC connects to these conditions—especially in high-dimensional or random design regimes—could lead to novel theoretical insights and principled regularization schemes. We leave this as a direction for future work.

V. Conclusion

We have presented an MSE-centric reinterpretation of Sparse Bayesian Learning (SBL), demonstrating that standard SBL implicitly approximates the minimization of input-domain MSE. Building on this insight, we explored several variants of SBL, including direct minimization of $\text{MSE}_{\mathbf{x}}$ and the use of $\text{SURE}_{\mathbf{x}}$ as a proxy risk. Due to the limitations of $\text{SURE}_{\mathbf{x}}$ in underdetermined systems, we proposed a tractable alternative based on $\text{SURE}_{\mathbf{z}}$. Our analysis shows that coordinate-wise minimization of $\text{SURE}_{\mathbf{z}}$ leads to a closed-form update for each hyperparameter, capable of promoting sparsity without explicit ℓ_1 regularization. We introduced the statistical orthogonality condition (SOC) as a sufficient criterion for pruning coefficients, thereby connecting risk-based learning with structural conditions commonly used in sparse recovery theory. Despite these insights, several theoretical questions remain unresolved—particularly regarding the typical scenarios in which SOC holds, and how this relates to design properties of the sensing matrix and signal model. Further research is needed to establish a comprehensive understanding of sparsity mechanisms under SURE-guided optimization.

Acknowledgements EURECOM's research is partially supported by its industrial members: ORANGE, BMW, SAP, iABG, Norton LifeLock, by the Franco-German project 5G-OPERA (BPI), the French project YACARI (PEPR-5G), the EU INFRA project CONVERGE, and by a Huawei France funded Chair towards Future Wireless Networks.

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] Z. Zhao, F. Xiao, and D. Slock, "Approximate Message Passing for Not So Large niid Generalized Linear Models," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2023, pp. 386–390.

[5] Z. Zhao and D. Slock, "Extrinsics and Linearized Component-Wise Conditionally Unbiased MMSE Estimation in Approximate Message Passing," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, 2025, pp. 871–875.

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

[7] C. K. Thomas and D. Slock, "SAVED - Space Alternating Variational Estimation for Sparse Bayesian Learning with Parametric Dictionaries," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, 2018, pp. 1985–1989.

[8] ——, "SAVE - Space Alternating Variational Estimation for Sparse Bayesian Learning," in *Proc. IEEE Data Sci. Workshop (DSW)*, 2018, pp. 11–15.

[9] R. Giri and B. 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.

[10] M. E. Tipping and A. C. Faul, "Fast Marginal Likelihood Maximisation for Sparse Bayesian Models," in *AISTATS*, January 2003.

[11] D. Slock, "Sparse Bayesian Learning with Stein's Unbiased Risk Estimator based Hyperparameter Optimization," in *Proc. 56th Asilomar Conf. Signals, Syst., Comput.*, 2022, pp. 857–861.

[12] F. Xiao and D. Slock, "Towards Hyperparameter Optimizing of Sparse Bayesian Learning Based on Stein's Unbiased Risk Estimator," in *Proc. IEEE Inf. Theory Workshop (ISIT-W)*, 2024, pp. 1–5.

[13] ——, "Single Snapshot Direction of Arrival Estimation Using the EP-SURE-SBL Algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2025, pp. 1–5.

[14] 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.