On MMSE and MAP Denoising Under Sparse Representation Modeling Over a Unitary Dictionary[☆]

J.S. Turek[a], I. Yavneh[a], M. Protter[a], M. Elad[a]

[a] *Technion, Israel Institute of Technology, Computer Science Department, Haifa 32000, Israel*

Abstract

Among the many ways to model signals, a recent approach that draws considerable attention is sparse representation modeling. In this model, the signal is assumed to be generated as a random linear combination of a few atoms from a pre-specified dictionary. In this work we analyze two Bayesian denoising algorithms – the Maximum-Aposteriori Probability (MAP) and the Minimum-Mean-Squared-Error (MMSE) estimators, under the assumption that the dictionary is unitary. It is well known that both these estimators lead to a scalar shrinkage on the transformed coefficients, albeit with a different response curve. In this work we start by deriving closed-form expressions for these shrinkage curves and then analyze their performance. Upper bounds on the MAP and the MMSE estimation errors are derived. We tie these to the error obtained by a so-called oracle estimator, where the support is given, establishing a worst-case gain-factor between the MAP/MMSE estimation errors and the oracle's performance. These denoising algorithms are demonstrated on synthetic signals and on true data (images).

Keywords: Sparse representations, MAP, MMSE, Unitary dictionary, Shrinkage, Bayesian estimation, Oracle

[☆]This research was supported by the European Community's FP7-FET program, SMALL project, under grant agreement no. 225913, and by the Israel Science Foundation (ISF) grant number 1031/08.

Email addresses: javiert@cs.technion.ac.il (J.S. Turek), irad@cs.technion.ac.il (I. Yavneh), matanpr@cs.technion.ac.il (M. Protter), elad@cs.technion.ac.il (M. Elad)

1. Introduction

A classical and long-studied subject in signal processing is denoising. This task considers a given measurement signal $\mathbf{y} \in \mathbb{R}^n$ obtained from a clear signal $\mathbf{w} \in \mathbb{R}^n$ by an additive contamination of the form $\mathbf{y} = \mathbf{w} + \mathbf{v}$. We shall restrict our discussion to zero mean i.i.d. Gaussian noise vectors $\mathbf{v} \in \mathbb{R}^n$, with each entry drawn at random from the normal distribution $\mathcal{N}\left(0, \sigma^2\right)$. The denoising goal is to recover $\mathbf{w}$ from $\mathbf{y}$.

An effective denoising algorithm assumes knowledge about the noise characteristics, like the above description, and introduces some assumptions about the class of signals to which $\mathbf{w}$ belongs, that is, a-priori knowledge about the signal. There is a great number of algorithms today, corresponding to a variety of signal models. Among these, a recently emerging group of techniques relies on sparse and redundant representations for modeling the signals [4].

A signal $\mathbf{w}$ is said to have a sparse representation over a known dictionary, $\mathbf{D} \in \mathbb{R}^{n \times m}$, if there exists a sparse vector $\mathbf{x} \in \mathbb{R}^m$ such that $\mathbf{w} = \mathbf{D}\mathbf{x}$. The vector $\mathbf{x}$ is the representation of $\mathbf{w}$, having a number of non-zeros, $\|\mathbf{x}\|_0 = k$, which is much smaller than its length, m. Thus, $\mathbf{x}$ describes how to construct $\mathbf{w}$ as a linear combination of a few columns (also referred to as atoms) of $\mathbf{D}$. In general, the dictionary may be redundant, containing more atoms than the signal dimension ($m \geq n$).

Assuming that $\mathbf{w} = \mathbf{D}\mathbf{x}$ with a sparse representation $\mathbf{x}$, how can one recover $\mathbf{w}$ from the noisy measurement $\mathbf{y}$? By posing a prior probability density function over $\mathbf{x}$, one can derive the exact Maximum-A'posteriori Probability (MAP) estimator for this task. This becomes a search for the support of the sparse representation $\hat{\mathbf{x}}$ that maximizes the posterior probability. This problem is computationally complex, as it generally requires an exponential sweep over all the possible sparse supports [17]. Therefore, approximation methods are often employed, such as the Orthogonal Matching Pursuit (OMP) [15] and the Basis Pursuit (BP) [6].

While MAP estimation promotes seeking a single sparse representation to explain the measurements, recent work has shown that better results[1] are possible using the Minimum Mean Square Error (MMSE) estimator [14, 20, 12]. These works develop MMSE estimators, showing that they lead to a weighted average of all the possible representations that may explain the signal, with weights related to their probabilities. Just like MAP in the general setting, this estimation is infeasible to compute, and thus various approximations are proposed [14, 20, 12].

A well known and celebrated result in signal processing is the fact that the MAP estimator mentioned above admits a closed-form simple formula, in the special case where the dictionary $\mathbf{D}$ is square and unitary [11, 21, 16]. This formula, known as a *shrinkage* operation, yields the estimate $\hat{\mathbf{w}}$ by applying a simple 1D operation on the entries of the vector $\mathbf{D}^T\mathbf{y}$. The denoised signal is then obtained by multiplication by $\mathbf{D}$. Shrinkage tends to eliminate small entries, while leaving larger ones almost intact.

Our recent work reported in [18, 19] aimed to develop an MMSE closed-form formula for the unitary case. With a specific prior model on $\mathbf{x}$, a recursive formula for this task was developed. Thus, at least in principle, the implications from this work are that one need not turn to approximations, as this formula is

[1] In the ℓ_2-error sense, which is often the measure used to assess performance.

easily computable, leading to the exact MMSE. While such a result is very encouraging, it does not provide a truly simple technique of the form that MAP enjoys. Furthermore, due to its recursive nature, this algorithm suffers from instability problems that hinder its use for high-dimensional signals.

In the present work, we propose a modified prior model for the sparse representation vector $\mathbf{x}$. We show that this change leads to a simplified MMSE formula, which, just as for the MAP, becomes a scalar shrinkage, albeit with a different response curve. As such, this exact MMSE denoising exhibits no numerical sensitivities as in [18, 19], and thus it can operate easily in any dimension.

The core idea that MMSE estimation for the unitary case leads to a shrinkage algorithm has been observed before [7, 8, 9, 1, 2]. Here we adopt a distinct approach in the derivation, which also gives us exact and simple expressions for MAP and MMSE shrinkage curves, and their expected ℓ_2-errors. We use these as a stepping-stone towards the development of upper bounds on the MAP and the MMSE estimation errors.

A fundamental and key question that has attracted attention in recent years is the proximity between practical pursuit[2] results and the oracle performance. The oracle is an estimator that knows the true support, thus giving an ultimate result which can be used as a gold-standard for assessing practical pursuit performance. For example, the work reported in [5] shows that the Danzig Selector algorithm is a constant (and log) factor away from the oracle result. Similar claims for the BP, the OMP, and even the thresholding algorithms, are made in [3].

In both these papers, the analysis is deterministic and non-Bayesian, which is different from the point of view taken in this paper. In this work we tie the MAP and the MMSE errors for the unitary case to the error obtained by an oracle estimator. We establish worst-case gain-factors of the MAP and the MMSE errors relative to the oracle error. This gives a clear ranking of these algorithms, and states clearly their nearness to the ideal performance.

The paper is organized as follows. In section 2 we describe the signal model we shall use throughout this work. For completeness of the presentation, we also derive the MAP and MMSE estimators for the general case in this section. In Section 3 we turn to the unitary case and present the ideal MAP and MMSE estimators, showing how both lead to shrinkage operations. Section 4 is devoted to the development of the performance behavior of the MAP and MMSE estimates, and the upper bounds on their errors. Section 5 presents numerical experiments, demonstrating the proposed algorithms in action. In Section 6 we conclude the paper.

2. Background

2.1. The Signal Model

We consider a generative signal model that resembles the one presented in [20]. In this model, each atom has a prior probability P_i of participating in the support of each signal, and $(1 - P_i)$ of not appearing. One can think of the support selection stage as performing biased coin-tosses of m coins, with the i^{th} coin having

[2]Pursuit is a generic name given to algorithms that aim to estimate $\mathbf{x}$.

a probability P_i of "heads" and $(1-P_i)$ for "tails". The coins that turn up ("heads") constitute the support $\mathcal{S}$ for this signal. Thus, the a priori probability for any support $\mathcal{S}$ is given by

$$P(\mathcal{S}) = \prod_{i\in\mathcal{S}} P_i \cdot \prod_{j\notin\mathcal{S}} (1-P_j). \tag{1}$$

It is important to note that, as opposed to the model used in [12, 18, 19], here it is not possible to explicitly prescribe the cardinality of the support, nor is it possible to limit it (as even the empty and full supports may arise by chance). If, for some i, P_i equals 0, all the supports that contain element i have zero probability. Similarly, if we have $P_i = 1$, then all the supports that do not select the i^{th} atom also have zero probability. Hence, in our study we only need to consider values $0 < P_i < 1$ for all i, and this is assumed henceforth.

We further assume that, given the support $\mathcal{S}$, the coefficients in $\mathbf{x}$ on this support are drawn as i.i.d. Gaussian random variables[3] with zero mean and variance σ_x^2,

$$\mathbf{x}|\mathcal{S} \sim \mathcal{N}\left(0, \sigma_x^2 \mathbf{I}_{|\mathcal{S}|}\right), \tag{2}$$

where $\mathbf{I}_{|\mathcal{S}|}$ is the identity matrix of size $|\mathcal{S}|$.

We measure the vector $\mathbf{y} \in \mathbb{R}^n$, a noisy linear combination of atoms from $\mathbf{D}$ with coefficients $\mathbf{x} \in \mathbb{R}^m$, namely, $\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{v}$, where the noise $\mathbf{v}$ is assumed to be white Gaussian with variance σ^2, i.e., $\mathbf{v} \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$, and the columns of $\mathbf{D}$ are normalized.

¿From the model assumptions made above, it can be seen [13] that $\mathbf{y}$ and $\mathbf{x}$ are jointly Gaussians for a given support,

$$\left[\begin{array}{c}\mathbf{y}\\ \mathbf{x}\end{array}\right] \Bigg| \mathcal{S} \sim \mathcal{N}\left(\left[\begin{array}{c}\mathbf{0}\\ \mathbf{0}\end{array}\right], \left[\begin{array}{cc}\mathbf{C}_{\mathcal{S}} & \sigma_x^2 \mathbf{D}_{\mathcal{S}}\\ \sigma_x^2 \mathbf{D}_{\mathcal{S}}^T & \sigma_x^2 \mathbf{I}_{|\mathcal{S}|}\end{array}\right]\right), \tag{3}$$

where

$$\mathbf{C}_{\mathcal{S}} = \sigma_x^2 \mathbf{D}_{\mathcal{S}} \mathbf{D}_{\mathcal{S}}^T + \sigma^2 \mathbf{I}_n, \tag{4}$$

and $\mathbf{D}_{\mathcal{S}} \in \mathbb{R}^{n\times|\mathcal{S}|}$ is comprised of the columns of the matrix $\mathbf{D}$ that appear in the support $\mathcal{S}$. Hence, the marginal p.d.f. $P(\mathbf{y}|\mathcal{S})$ is Gaussian and it is given by

$$\mathbf{y}|\mathcal{S} \sim \mathcal{N}\left(0, \mathbf{C}_{\mathcal{S}}\right). \tag{5}$$

Using properties of the Multivariate Gaussian p.d.f. (see [13, p. 325]), we have that the likelihood $P(\mathbf{y}|\mathbf{x}, \mathcal{S})$ and the posterior p.d.f. $P(\mathbf{x}|\mathbf{y}, \mathcal{S})$ are also Gaussian, namely

$$\mathbf{y}|\mathbf{x}, \mathcal{S} \sim \mathcal{N}\left(\mathbf{D}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}, \sigma^2 \mathbf{I}_n\right) \tag{6}$$

$$\mathbf{x}|\mathbf{y}, \mathcal{S} \sim \mathcal{N}\left(\frac{1}{\sigma^2}\mathbf{Q}_{\mathcal{S}}^{-1}\mathbf{D}_{\mathcal{S}}^T\mathbf{y}, \mathbf{Q}_{\mathcal{S}}^{-1}\right), \tag{7}$$

[3] In fact, we may suggest a broader model of the form $P(x_i) \sim \exp\{-f(x_i/\sigma_x)\}$, for an arbitrary function $f(\cdot)$, thus keeping the model very general. It appears that with this change one can still obtain MMSE-shrinkage. Furthermore, one may also study the sensitivity of MMSE/MAP shrinkage-curves under perturbations of $f(\cdot)$, and even find the worst choice of this function, that leads to the maximal expected error in MMSE – all these are left to future work, as we mainly focus here on the Gaussian model.

where the sub-vector $\mathbf{x}_{\mathcal{S}}$ is comprised of the elements of $\mathbf{x}$ whose indices are in the support $\mathcal{S}$, and

$$\mathbf{Q}_{\mathcal{S}} = \frac{1}{\sigma_x^2}\mathbf{I}_{|\mathcal{S}|} + \frac{1}{\sigma^2}\mathbf{D}_{\mathcal{S}}^T\mathbf{D}_{\mathcal{S}}. \tag{8}$$

There is a direct link between the matrices $\mathbf{Q}_{\mathcal{S}}$ and $\mathbf{C}_{\mathcal{S}}$, expressed using the matrix inversion lemma,

$$\mathbf{C}_{\mathcal{S}}^{-1} = \frac{1}{\sigma^2}\mathbf{I}_n - \frac{1}{\sigma^4}\mathbf{D}_{\mathcal{S}}\mathbf{Q}_{\mathcal{S}}^{-1}\mathbf{D}_{\mathcal{S}}^T. \tag{9}$$

2.2. MAP/MMSE Estimators – The General Case

2.2.1. The Oracle Estimator

The first estimator we derive is the oracle. This estimator assumes knowledge of the chosen support for $\mathbf{x}$, information that is unknown in the actual problem. Therefore it cannot be obtained in practice. Nevertheless, it gives us a reference performance quality to compare against. The oracle can target the minimization of the MSE[4]. A well-known and classical result states that the MMSE estimator is equal to the conditional mean of the unknown, conditioned on the known parts, and thus in our case it is $E\{\mathbf{x}|\mathbf{y},\mathcal{S}\}$. As the support $\mathcal{S}$ is known, we need to estimate $\mathbf{x}_{\mathcal{S}}$, the sub-vector of non-zero entries of $\mathbf{x}$, so the estimator is given by

$$\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle} = E\{\mathbf{x}_{\mathcal{S}}|\mathbf{y},\mathcal{S}\} = \frac{1}{\sigma^2}\mathbf{Q}_{\mathcal{S}}^{-1}\mathbf{D}_{\mathcal{S}}^T\mathbf{y}, \tag{10}$$

where this equality comes from the expectation of the probability distribution in (7).

2.2.2. Maximum A-Posteriori Estimator (MAP)

The MAP estimator proposes an estimate $\hat{\mathbf{x}}$ that maximizes the posterior probability. As the model mixes discrete probabilities P_i with continuous ones $P(\mathbf{x}|\mathcal{S})$, the MAP should be carefully formulated, otherwise, the most probable estimate would be the zero vector. Thus, we choose instead to maximize the posterior of the support,

$$\mathcal{S}^{MAP} = \arg\max_{\mathcal{S}} P(\mathcal{S}|\mathbf{y}), \tag{11}$$

and only then compute the corresponding estimate $\hat{\mathbf{x}}_{\mathcal{S}^{\mathcal{MAP}}}$. We know from Equation (7), that $\mathbf{x}|\mathbf{y},\mathcal{S}$ behaves as a normal distribution, and thus the estimate $\hat{\mathbf{x}}_{\mathcal{MAP}}$ is given by the oracle in (10) with the specific support $\mathcal{S}^{MAP}$. Using Bayes's rule, Equation (11) leads to

$$P(\mathcal{S}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathcal{S})P(\mathcal{S})}{P(\mathbf{y})}. \tag{12}$$

Since $P(\mathbf{y})$ does not depend on $\mathcal{S}$, it affects this expression only as a normalizing factor. Using the expressions of the probabilities in the numerator that are given by Equations (5) and (1), respectively, we obtain

$$P(\mathcal{S}|\mathbf{y}) \propto \frac{1}{\sqrt{\det(\mathbf{C}_{\mathcal{S}})}}\exp\left\{-\frac{1}{2}\mathbf{y}^T\mathbf{C}_{\mathcal{S}}^{-1}\mathbf{y}\right\}\cdot\prod_{i\in\mathcal{S}}P_i\cdot\prod_{j\notin\mathcal{S}}1-P_j \equiv t_{\mathcal{S}}, \tag{13}$$

[4]Or MAP – in fact, the two are the same in this case due to the Gaussianity of $\mathbf{x}|\mathbf{y},\mathcal{S}$.

where we have introduced the notation $t_{\mathcal{S}}$ for brevity of later expressions. Returning to our MAP goal posed in Equation (11), applying a few simple algebraic steps on the expression for $P(\mathcal{S}|\mathbf{y})$ leads to the following penalty function, which should be maximized with respect to the support $\mathcal{S}$,

$$Val(\mathcal{S}) = \frac{1}{2}\left\|\frac{1}{\sigma^2}\mathbf{Q}_{\mathcal{S}}^{-1/2}\mathbf{D}_{\mathcal{S}}^{T}\mathbf{y}\right\|_2^2 - \frac{1}{2}\log\det\mathbf{C}_{\mathcal{S}} + \sum_{i\in\mathcal{S}}\log\left(P_i\right) + \sum_{j\notin\mathcal{S}}\log\left(1-P_j\right), \tag{14}$$

over all 2^m possible supports. Once found, we obtain the MAP estimation by using the oracle formula from Equation (10), which computes $\hat{\mathbf{x}}_{\mathcal{S}}$ for this support.

2.2.3. Minimum Mean Square Error Estimator (MMSE)

The MMSE estimate is given by the conditional expectation, $E\left\{\mathbf{x}|\mathbf{y}\right\}$,

$$\hat{\mathbf{x}}^{MMSE} = E\left\{\mathbf{x}|\mathbf{y}\right\} = \int_{\mathbf{x}}\mathbf{x}P(\mathbf{x}|\mathbf{y})d\mathbf{x}. \tag{15}$$

Marginalizing the posterior probability $P(\mathbf{x}|\mathbf{y})$ over all possible supports $\mathcal{S}\in\Omega$, we have

$$P(\mathbf{x}|\mathbf{y}) = \sum_{\mathcal{S}\in\Omega}P(\mathbf{x}|\mathbf{y},\mathcal{S})P(\mathcal{S}|\mathbf{y}). \tag{16}$$

Plugging Equation (16) into Equation (15) yields

$$\begin{aligned}\hat{\mathbf{x}}^{MMSE} &= \sum_{\mathcal{S}\in\Omega}P(\mathcal{S}|\mathbf{y})\int_{\mathbf{x}}\mathbf{x}P(\mathbf{x}|\mathbf{y},\mathcal{S})d\mathbf{x} \\ &= \sum_{\mathcal{S}\in\Omega}P(\mathcal{S}|\mathbf{y})E\left\{\mathbf{x}|\mathbf{y},\mathcal{S}\right\} \\ &= \sum_{\mathcal{S}\in\Omega}P(\mathcal{S}|\mathbf{y})\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}.\end{aligned} \tag{17}$$

Equation (17) shows that the MMSE estimator is a weighted average of all the "oracle" solutions, each with a different support and weighted by its probability. Finally, we substitute the expression t_s developed in Equation (13) into Equation (17), and get the formula for MMSE estimation,

$$\hat{\mathbf{x}}^{MMSE} = \frac{1}{t}\sum_{\mathcal{S}\in\Omega}t_{\mathcal{S}}\cdot\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}, \tag{18}$$

where $t=\sum_{\mathcal{S}\in\Omega}t_{\mathcal{S}}$ is the overall normalizing factor.

2.3. Estimator Performance – The General Case

We conclude this background section by discussing the expected Mean-Squared-Error (MSE) induced by each of the estimators developed above. Our goal is to obtain clear expressions for these errors, which will later serve when we develop similar and simpler expressions for the unitary case.

We start with the performance of the oracle estimator, as the oracle is central to the derivation of MAP and MMSE errors. The oracle's expected MSE is given by

$$\begin{aligned}E\left\{\left\|\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}-\mathbf{x}_{\mathcal{S}}\right\|_2^2\middle|\mathbf{y}\right\} &= E\left\{\left\|\mathbf{Q}_{\mathcal{S}}^{-1}\frac{1}{\sigma^2}\mathbf{D}_{\mathcal{S}}^{T}\mathbf{y}-\mathbf{x}_{\mathcal{S}}\right\|_2^2\middle|\mathbf{y}\right\} \\ &= E\left\{\left\|\mathbf{Q}_{\mathcal{S}}^{-1}\frac{1}{\sigma^2}\mathbf{D}_{\mathcal{S}}^{T}\left(\mathbf{D}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}+\mathbf{v}\right)-\mathbf{x}_{\mathcal{S}}\right\|_2^2\middle|\mathbf{y}\right\} = \mathrm{trace}\left(\mathbf{Q}_{\mathcal{S}}^{-1}\right),\end{aligned} \tag{19}$$

where we have used Equation (8), and the fact that $\mathbf{y} = \mathbf{D}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}} + \mathbf{v}$.

Our analysis continues with the expected error for a general estimate $\hat{\mathbf{x}}$, observing that it can be written as

$$\begin{aligned} E\left\{ \left\|\hat{\mathbf{x}}-\mathbf{x}\right\|_2^2 \middle| \mathbf{y} \right\} &= \int_{\mathbf{x}\in\mathbb{R}^m} \left\|\hat{\mathbf{x}}-\mathbf{x}\right\|_2^2 P(\mathbf{x}|\mathbf{y})d\mathbf{x} \\ &= \sum_{\mathcal{S}\in\Omega} P(\mathcal{S}|\mathbf{y}) \int_{\mathbf{x}\in\mathbb{R}^m} \left\|\hat{\mathbf{x}}-\mathbf{x}\right\|_2^2 P(\mathbf{x}|\mathbf{y},\mathcal{S})d\mathbf{x}, \end{aligned} \tag{20}$$

where we have used the marginalization proposed in Equation (16). We add and subtract the oracle estimate $\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}$ that corresponds to the support $\mathcal{S}$ into the norm term, yielding

$$\begin{aligned} \int_{\mathbf{x}\in\mathbb{R}^m} \left\|\hat{\mathbf{x}}-\mathbf{x}\right\|_2^2 P(\mathbf{x}|\mathbf{y},\mathcal{S})d\mathbf{x} &= \int_{\mathbf{x}\in\mathbb{R}^m} \left\|\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}-\mathbf{x}\right\|_2^2 P(\mathbf{x}|\mathbf{y},\mathcal{S})d\mathbf{x} \\ &\quad + \int_{\mathbf{x}\in\mathbb{R}^m} \left\|\hat{\mathbf{x}}-\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}\right\|_2^2 P(\mathbf{x}|\mathbf{y},\mathcal{S})d\mathbf{x}. \end{aligned} \tag{21}$$

Note that the integral over the cross-term $\left(\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}-\mathbf{x}\right)^T\left(\hat{\mathbf{x}}-\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}\right)$ vanishes, since the term $\left(\hat{\mathbf{x}}-\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}\right)$ is deterministic and can thus be moved outside the integration, while the expression remaining inside the integral is zero, since the oracle estimate is the expected $\mathbf{x}$ over this domain and with this support.

Continuing with Equation (21), the first term represents the MSE of an oracle for a given support $\mathcal{S}$, as derived in Equation (19). In the second term, the norm factor does not depend on the integral variable $\mathbf{x}$, and thus it may be pulled outside the integration. The remaining part is equal to one. Therefore,

$$\int_{\mathbf{x}\in\mathbb{R}^m} \left\|\hat{\mathbf{x}}-\mathbf{x}\right\|_2^2 P(\mathbf{x}|\mathbf{y},\mathcal{S})d\mathbf{x} = \text{trace}\left(\mathbf{Q}_{\mathcal{S}}^{-1}\right) + \left\|\hat{\mathbf{x}}-\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}\right\|_2^2. \tag{22}$$

Returning to the overall expected MSE as in Equation (20), using the fact that $P(\mathcal{S}|\mathbf{y}) = t_{\mathcal{S}}/t$, as developed in Equation (13), we have

$$E\left\{ \left\|\hat{\mathbf{x}}-\mathbf{x}\right\|_2^2 \right\} = \frac{1}{t}\sum_{\mathcal{S}\in\Omega} t_{\mathcal{S}} \cdot \left[\text{trace}\left(\mathbf{Q}_{\mathcal{S}}^{-1}\right) + \left\|\hat{\mathbf{x}}-\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}\right\|_2^2\right]. \tag{23}$$

By plugging $\hat{\mathbf{x}} = \hat{\mathbf{x}}^{MMSE}$ into this expression, we get the MMSE error. Note that if we minimize the above with respect to $\mathbf{x}$, we get the MMSE estimate formula exactly, as expected, since the MMSE is the solution that leads to the smallest error.

Observe that (23) can be written differently by adding and subtracting $\hat{\mathbf{x}}^{MMSE}$ inside the norm term, giving

$$\begin{aligned} E\left\{ \left\|\hat{\mathbf{x}}-\mathbf{x}\right\|_2^2 \middle| \right\} &= \frac{1}{t}\sum_{\mathcal{S}\in\Omega} t_{\mathcal{S}} \cdot \text{trace}\left(\mathbf{Q}_{\mathcal{S}}^{-1}\right) + \frac{1}{t}\sum_{\mathcal{S}\in\Omega} t_{\mathcal{S}} \left\|\hat{\mathbf{x}}-\hat{\mathbf{x}}^{MMSE}+\hat{\mathbf{x}}^{MMSE}-\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}\right\|_2^2 \\ &= \frac{1}{t}\sum_{\mathcal{S}\in\Omega} t_{\mathcal{S}} \cdot \text{trace}\left(\mathbf{Q}_{\mathcal{S}}^{-1}\right) + \left\|\hat{\mathbf{x}}-\hat{\mathbf{x}}^{MMSE}\right\|_2^2 + \frac{1}{t}\sum_{\mathcal{S}\in\Omega} t_{\mathcal{S}} \left\|\hat{\mathbf{x}}^{MMSE}-\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}\right\|_2^2 \\ &= \left\|\hat{\mathbf{x}}-\hat{\mathbf{x}}^{MMSE}\right\|_2^2 + E\left\{ \left\|\hat{\mathbf{x}}^{MMSE}-\mathbf{x}\right\|_2^2 \middle| \mathbf{y} \right\}. \end{aligned} \tag{24}$$

In this derivation, the cross-term $(\hat{\mathbf{x}}-\hat{\mathbf{x}}^{MMSE})^T(\hat{\mathbf{x}}^{MMSE}-\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle})$ drops out, since in this summation the term $(\hat{\mathbf{x}}-\hat{\mathbf{x}}^{MMSE})^T$ can be positioned outside the summation, and then, using Equation (18), it is easily

shown that we are left with an expression that equals $\hat{\mathbf{x}}^{MMSE} - \hat{\mathbf{x}}^{MMSE} = 0$. We have then a general error formula for any estimator, given by equation (24). In particular, this means that the error for the MAP estimate can be calculated by

$$E\left\{\left\|\hat{\mathbf{x}}^{MAP}-\mathbf{x}\right\|_2^2\middle|\mathbf{y}\right\} = \left\|\hat{\mathbf{x}}^{MAP}-\hat{\mathbf{x}}^{MMSE}\right\|_2^2 + E\left\{\left\|\hat{\mathbf{x}}^{MMSE}-\mathbf{x}\right\|_2^2\middle|\mathbf{y}\right\}. \tag{25}$$

3. MAP & MMSE Estimators for a Unitary Dictionary

The derivation of MAP and MMSE for a general dictionary leads to prohibitive computational tasks. As we shall see next, when using unitary dictionaries, we are able to avoid these demanding computations, and instead obtain closed-form solutions for each one of the estimators. Furthermore, the two resulting algorithms are very similar, both having a shrinkage structure.

While this claim about MAP and MMSE leading to shrinkage is not new [7, 8, 9, 1, 2], our distinct development of the closed-form shrinkage formulae will lead to a simple computational process for the evaluation of the MAP and the MMSE, which will facilitate the performance analysis derived in Section 4.

3.1. The Oracle

Just as for the general dictionary case, we start by deriving an expression for the oracle estimation. In this case, we assume that the dictionary $\mathbf{D}$ is a unitary matrix, and thus $\mathbf{D}^T\mathbf{D} = \mathbf{I}$. Moreover, it is easily seen that $\mathbf{D}_{\mathcal{S}}^T\mathbf{D}_{\mathcal{S}} = \mathbf{I}_{|\mathcal{S}|}$, which will simplify our expressions. We start by simplifying the matrix $\mathbf{Q}_{\mathcal{S}}$ defined in (8),

$$\mathbf{Q}_{\mathcal{S}} = \frac{1}{\sigma_x^2}\mathbf{I}_{|\mathcal{S}|} + \frac{1}{\sigma^2}\mathbf{D}_{\mathcal{S}}^T\mathbf{D}_{\mathcal{S}} = \frac{\sigma_x^2+\sigma^2}{\sigma_x^2\sigma^2}\mathbf{I}_{|\mathcal{S}|}. \tag{26}$$

The oracle solution, as given in Equation (10), becomes

$$\hat{\mathbf{x}}^{Oracle} = \frac{1}{\sigma^2}\mathbf{Q}_{\mathcal{S}}^{-1}\mathbf{D}_{\mathcal{S}}^T\mathbf{y} = c^2\beta_{\mathcal{S}}, \tag{27}$$

where we have defined the constant $c^2 = \sigma_x^2/(\sigma_x^2+\sigma^2)$ and the vector $\beta_{\mathcal{S}} = \mathbf{D}_{\mathcal{S}}^T\mathbf{y}$. The oracle estimator has thus been reduced to a simple matrix by vector multiplication.

3.2. The MAP – Unitary Case

We turn to the MAP estimation, which requires to first find the optimal support $\mathcal{S}$ based on Equations (13) and (14), and then plug it into the oracle expression as given in Equation (10) to get the estimate.

We proceed by simplifying the expression $\det(\mathbf{C}_{\mathcal{S}})$ in Equations (13) and (14). The matrix $\mathbf{C}_{\mathcal{S}}$ is defined in Equation (4) as $\mathbf{C}_{\mathcal{S}} = \sigma_x^2\mathbf{D}_{\mathcal{S}}\mathbf{D}_{\mathcal{S}}^T + \sigma^2\mathbf{I}_n$. Denoting by $\mathbf{W}_{\mathcal{S}}$ a diagonal matrix with ones and zeros on its main diagonal matching the support[5] $\mathcal{S}$, we obtain

$$\begin{aligned}\det(\mathbf{C}_{\mathcal{S}}) &= \det\left(\sigma_x^2\mathbf{D}_{\mathcal{S}}\mathbf{D}_{\mathcal{S}}^T + \sigma^2\mathbf{I}_n\right)\\ &= \det\left(\mathbf{D}\right)\cdot\det\left(\sigma_x^2\mathbf{W}_{\mathcal{S}} + \sigma^2\mathbf{I}_n\right)\cdot\det\left(\mathbf{D}^T\right)\\ &= \left(\sigma_x^2+\sigma^2\right)^{|\mathcal{S}|}\left(\sigma^2\right)^{n-|\mathcal{S}|} = \left(1-c^2\right)^{-|\mathcal{S}|}\sigma^{2n}.\end{aligned} \tag{28}$$

[5] $(\mathbf{W}_{\mathcal{S}})_{ii}$ is 1 if $i \in \mathcal{S}$, and 0 elsewhere.

Plugging this result into Equation (13), and using the relation between $\mathbf{C}_\mathcal{S}$ and $\mathbf{Q}_\mathcal{S}$ in Equation (9), yields

$$\begin{aligned} P(\mathcal{S}|\mathbf{y}) &\propto \exp\left\{\frac{c^2}{2\sigma^2}\mathbf{y}^T\mathbf{D}_\mathcal{S}\mathbf{D}_\mathcal{S}^T\mathbf{y} - \frac{1}{2}\log\left(1-c^2\right)^{-|\mathcal{S}|}\right\}\prod_{i\in\mathcal{S}}P_i\cdot\prod_{j\notin\mathcal{S}}1-P_j \\ &\propto \exp\left\{\frac{c^2}{2\sigma^2}\left\|\beta_\mathcal{S}\right\|_2^2\right\}\prod_{i\in\mathcal{S}}P_i\sqrt{1-c^2}\cdot\prod_{j\notin\mathcal{S}}1-P_j \\ &\propto \prod_{i\in\mathcal{S}}\exp\left\{\frac{c^2}{2\sigma^2}\beta_i^2\right\}P_i\sqrt{1-c^2}\cdot\prod_{j\notin\mathcal{S}}1-P_j. \end{aligned} \tag{29}$$

Taking into account that $0 < P_i < 1$, we can rewrite this expression as

$$\begin{aligned} P(\mathcal{S}|\mathbf{y}) &\propto \prod_{i\in\mathcal{S}}\exp\left\{\frac{c^2}{2\sigma^2}\beta_i^2\right\}\frac{P_i}{1-P_i}\sqrt{1-c^2}\cdot\prod_{j=1}^{n}1-P_j \\ &\propto \prod_{i\in\mathcal{S}}\exp\left\{\frac{c^2}{2\sigma^2}\beta_i^2\right\}\frac{P_i}{1-P_i}\sqrt{1-c^2} = \prod_{i\in\mathcal{S}}q_i, \end{aligned} \tag{30}$$

where we have defined

$$q_i = \exp\left\{\frac{c^2}{2\sigma^2}\beta_i^2\right\}\frac{P_i}{1-P_i}\sqrt{1-c^2}. \tag{31}$$

We further define $g_i = q_i/(1+q_i)$ (which implies that $q_i = g_i/(1-g_i)$), and substitute this into Equation (30). Adding now the necessary normalization factor we get

$$\begin{aligned} P(\mathcal{S}|\mathbf{y}) &= \left(\sum_{\mathcal{S}^*\in\Omega}\prod_{i\in\mathcal{S}^*}q_i\right)^{-1}\prod_{i\in\mathcal{S}}q_i \\ &= \left(\sum_{\mathcal{S}^*\in\Omega}\prod_{i\in\mathcal{S}^*}\frac{g_i}{1-g_i}\right)^{-1}\prod_{i\in\mathcal{S}}\frac{g_i}{1-g_i} \\ &= \left(\frac{\sum_{\mathcal{S}^*\in\Omega}\prod_{i\in\mathcal{S}^*}g_i\prod_{j\notin\mathcal{S}^*}(1-g_j)}{\prod_{k=1}^{n}(1-g_k)}\right)^{-1}\frac{\prod_{i\in\mathcal{S}}g_i\prod_{j\notin\mathcal{S}}(1-g_j)}{\prod_{k=1}^{n}(1-g_k)} \\ &= \left(\sum_{\mathcal{S}^*\in\Omega}\prod_{i\in\mathcal{S}^*}g_i\prod_{j\notin\mathcal{S}^*}(1-g_j)\right)^{-1}\prod_{i\in\mathcal{S}}g_i\prod_{j\notin\mathcal{S}}(1-g_j)\,. \end{aligned} \tag{32}$$

The following observation will facilitate a further simplification of this expression:

Proposition 1. *Let Ω be the set of all possible subsets of n indices, and let g_i be values associated with each index, such that $0 \le g_i \le 1$. Then,*

$$\sum_{\mathcal{S}\in\Omega}\prod_{i\in\mathcal{S}}g_i\cdot\prod_{j\notin\mathcal{S}}(1-g_j) = 1. \tag{33}$$

Proof. Consider the following experiment: a set of n independent coins are tossed, with the i^{th} coin having a probability g_i for "heads" and $(1-g_i)$ for "tails". The probability of a specific set of $\mathcal{S}$ coins turning up "heads" (and the rest turning up "tails") is $\prod_{i\in\mathcal{S}}g_i\cdot\prod_{j\notin\mathcal{S}}(1-g_j)$. For any one toss of the n coins, exactly one of these combinations will be the outcome. Therefore, the sum of these probabilities over all the combinations must be 1. □

Using this proposition, the normalization term in Equation (32) vanishes, as it is equal to 1 ($0 < g_i \leq 1$ since $g_i = \frac{q_i}{1+q_i}$ and $q_i \geq 0$ for every i). We therefore obtain

$$P(\mathcal{S}|\mathbf{y}) = \prod_{i\in\mathcal{S}} g_i \prod_{j\notin\mathcal{S}} (1-g_j)\,. \tag{34}$$

The optimization task (11) can now be written as

$$\begin{aligned}
\mathcal{S}^{\mathcal{MAP}} &= \arg\max_{\mathcal{S}\in\Omega} \prod_{i\in\mathcal{S}} g_i \prod_{j\notin\mathcal{S}} (1-g_j) \\
&= \arg\max_{\mathcal{S}\in\Omega} \prod_{i\in\mathcal{S}} \frac{q_i}{1+q_i} \prod_{j\notin\mathcal{S}} \left(1-\frac{q_j}{1+q_j}\right) \\
&= \arg\max_{\mathcal{S}\in\Omega} \frac{\prod_{i\in\mathcal{S}} q_i \prod_{j\notin\mathcal{S}} 1}{\prod_{k=1}^{n}(1+q_k)} = \arg\max_{\mathcal{S}\in\Omega} \prod_{i\in\mathcal{S}} \exp\left\{\frac{c^2}{2\sigma^2}\beta_i^2\right\} \frac{P_i}{1-P_i}\sqrt{1-c^2}.
\end{aligned} \tag{35}$$

Interpreting this expression, we see that every element in the support influences the penalty in one of two ways:

- **If it is part of the support**: Multiply the expression by $\sqrt{1-c^2}\frac{P_i}{1-P_i}\exp\left\{\frac{c^2}{2\sigma^2}\beta_i^2\right\}$, or
- **If it is not in the support**: Multiply the expression by 1.

As we aim to maximize the expression in Equation (35), the support will contain all the elements i such that $\sqrt{1-c^2}\cdot\frac{P_i}{1-P_i}\cdot\exp\left\{\frac{c^2}{2\sigma^2}\beta_i^2\right\} > 1$. (In the case that no such element exists, the support should be empty and the solution is therefore $\hat{\mathbf{x}}^{MAP} = \mathbf{0}$.) Once these elements are found, all we have to do is to multiply their value β_i by c^2 and this is the MAP estimate.

Stated differently, this means that after computing the transformed vector $\beta = \mathbf{D}^T\mathbf{y}$, we test each of its entries, and set the MAP estimate for the i^{th} entry to be

$$\hat{x}_i^{MAP} = \psi_{MAP}(\beta_i) = \begin{cases} c^2\beta_i & |\beta_i| > \frac{\sqrt{2}\sigma}{c}\sqrt{\log\left(\frac{1-P_i}{\sqrt{1-c^2}P_i}\right)} \\ 0 & \text{otherwise} \end{cases}. \tag{36}$$

This is the shrinkage algorithm mentioned earlier – each entry is handled independently of the others, passing through a scalar shrinkage curve that nulls small entries and keeps large ones intact (up to the multiplication by c^2). There is no trace of the exhaustive and combinatorial search that characterizes MAP in the general case, and this simple algorithm yields the exact MAP estimation.

3.3. The MMSE – The Unitary Case

Equation (18) shows the presence of the oracle in the MMSE estimation. Similarly to MAP, we make use of the unitary oracle estimate in Equation (27). Note that $\beta_{\mathcal{S}}$ may be written as

$$\beta_{\mathcal{S}} = \sum_{k=1}^{n} \mathbf{I}_{\mathcal{S}}(k)\beta_k\mathbf{e}_k, \tag{37}$$

where $\mathbf{e}_k$ is the k^{th} vector in the canonical basis, and $\mathbf{I}_{\mathcal{S}}(k)$ is an indicator function ($\mathbf{I}_{\mathcal{S}}(k) = 1$ if $k \in \mathcal{S}$, and zero otherwise). While this may seem like a cumbersome change, it will prove valuable in later derivations. Starting from Equation (18), substituting the expression developed for $P(\mathcal{S}|\mathbf{y})$ in Equation (34) into Equation (18), and using Equation (37), we obtain the following expression for the unitary MMSE estimator,

$$\begin{aligned}\hat{\mathbf{x}}^{MMSE} &= \sum_{\mathcal{S}\in\Omega}\left[\prod_{i\in\mathcal{S}} g_i \prod_{j\notin\mathcal{S}} (1-g_j)\, c^2 \cdot \left(\sum_{k=1}^{n} \mathbf{I}_{\mathcal{S}}(k)\beta_k \mathbf{e}_k\right)\right] \\ &= c^2 \sum_{k=1}^{n}\left[\sum_{\mathcal{S}\in\Omega} \mathbf{I}_{\mathcal{S}}(k) \prod_{i\in\mathcal{S}} g_i \prod_{j\notin\mathcal{S}} (1-g_j)\right]\beta_k \mathbf{e}_k. \end{aligned} \tag{38}$$

We introduce now another observation, similar to the one posed in Proposition 1. This will be used to further simplify the above expression.

Proposition 2. *Let Ω be the set of all possible subsets of n indices, and let g_i be values associated with each index, such that $0 \le g_i \le 1$. Then,*

$$\sum_{\mathcal{S}\in\Omega} \mathbf{I}_{\mathcal{S}}(k) \prod_{i\in\mathcal{S}} g_i \cdot \prod_{j\notin\mathcal{S}} (1-g_j) = g_k. \tag{39}$$

PROOF. In the spirit of the coin tossing interpretation described in the proof of Proposition 1, the multiplication by the expression $\mathbf{I}_{\mathcal{S}}(k)$ implies that only toss outcomes where the k^{th} coin turns up "heads" are included in the summation. Thus, the overall probability of those is exactly the probability that the k^{th} coin turn up "heads", which is g_k as claimed. A somewhat more formal way to pose this rationale is by observing that

$$\begin{aligned}\sum_{\mathcal{S}\in\Omega} \mathbf{I}_{\mathcal{S}}(k) \prod_{i\in\mathcal{S}} g_i \cdot \prod_{j\notin\mathcal{S}} (1-g_j) &= \sum_{\mathcal{S}\in\Omega \ s.t. \ k\in\mathcal{S}} \prod_{i\in\mathcal{S}} g_i \cdot \prod_{j\notin\mathcal{S}} (1-g_j) \\ &= g_k \cdot \sum_{\mathcal{S}\in\Omega_k} \prod_{i\in\mathcal{S}} g_i \cdot \prod_{j\notin\mathcal{S}} (1-g_j)\,. \end{aligned}$$

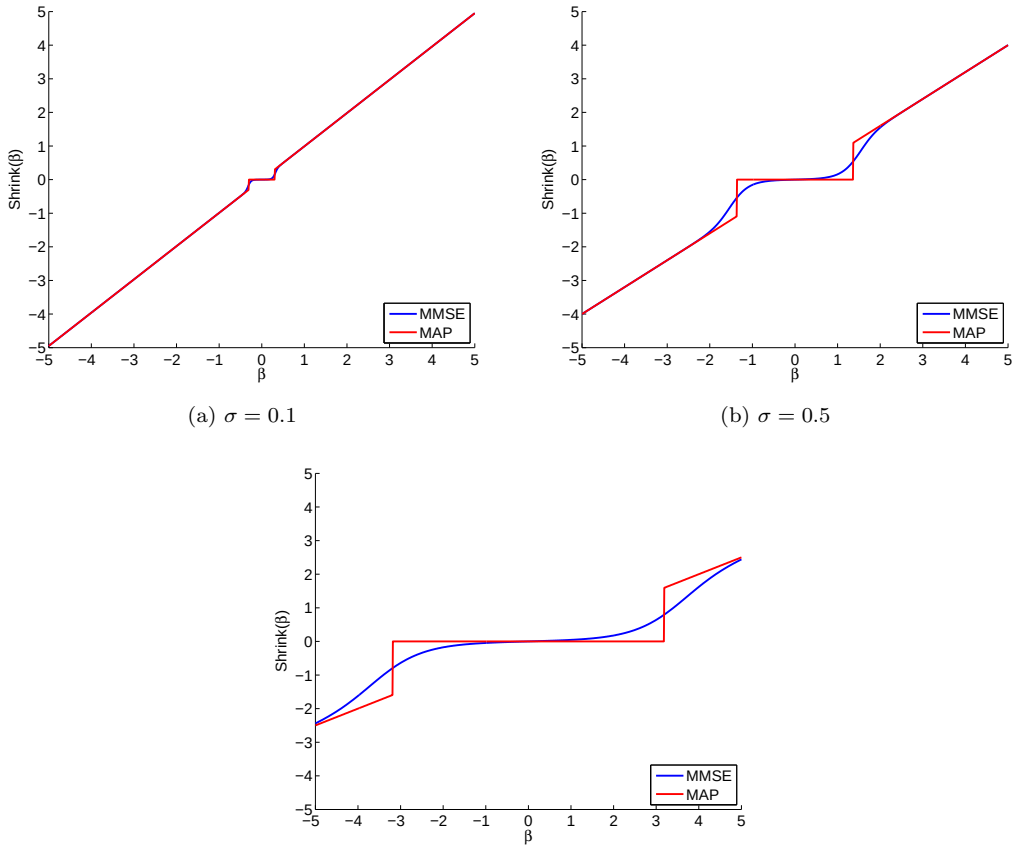
The last summation is over the set Ω_k, that contains all the supports in Ω and do not contain the k^{th} entry. Thus, for the remaining $n-1$ elements, this summation is complete, just as posed in Proposition 1, and therefore the overall expression equals g_k. □

Returning to the MMSE expression in Equation (38), and using this equality, we get a far simpler MMSE expression of the form

$$\hat{\mathbf{x}}^{MMSE} = c^2 \sum_{k=1}^{n} g_k \beta_k \mathbf{e}_k = c^2 \sum_{k=1}^{n} \frac{q_k}{1+q_k} \beta_k \mathbf{e}_k. \tag{40}$$

This is an explicit formula for MMSE estimation. The estimation is computed by first calculating $\beta = \mathbf{D}^T\mathbf{y}$, and then simply multiplying each entry β_k by $c^2 q_k/(1+q_k)$ (which is a function of β_k as well). Explicitly, the MMSE estimate is given elementwise by

$$\hat{x}_i^{MMSE} = \psi_{MMSE}(\beta_i) = \frac{\exp\left\{\frac{c^2}{2\sigma^2}\beta_i^2\right\}\frac{P_i}{1-P_i}\sqrt{1-c^2}}{1+\exp\left\{\frac{c^2}{2\sigma^2}\beta_i^2\right\}\frac{P_i}{1-P_i}\sqrt{1-c^2}} \cdot c^2\beta_i. \tag{41}$$



(a) $\sigma = 0.1$

(b) $\sigma = 0.5$

(c) $\sigma = 1$

Figure 1: Shrinkage functions for MMSE and MAP estimators ($P_i = 0.1$, $\sigma_x = 1$).

This operation has the form of a scalar shrinkage operation, just like MAP. For $|\beta_i| \ll \sigma/c$ this formula leads to $\hat{x}_i^{MMSE} \approx 0$, whereas for $|\beta_i| \gg \sigma/c$ the outcome is $\hat{x}_i^{MMSE} \approx c^2\beta_i$ (just like the MAP). Thus, the expression multiplying $c^2\beta_i$ here serves as a soft-shrinkage[6] operation, which replaces the hard-shrinkage practiced in the MAP. Figure 1 shows the various shrinkage functions obtained for each estimator.

4. Performance Analysis

4.1. Deriving the Estimators' MSE

Our main goal in this work is to develop error expressions for the different estimators in the unitary regime, exploiting the general derivations of section 2.3. We start by calculating the error for an oracle solution $\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}$. Using Equation (26) we obtain

$$E\left\{\left\|\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle} - \mathbf{x}\right\|_2^2\right\} = \text{trace}\left(\mathbf{Q}_{\mathcal{S}}^{-1}\right) = |\mathcal{S}|c^2\sigma^2 = \sum_{k=1}^{n} \mathbf{I}_{\mathcal{S}}(k)c^2\sigma^2, \tag{42}$$

[6]This should not be confused with the term soft-thresholding obtained when minimizing an ℓ_1 penalty.

where the indicator function is the same as previously used in (37). The last equality will become useful for our later development.

Turning to the MMSE estimator, recall the general expected-MSE expression in Equation (23),

$$E\left\{\left\|\hat{\mathbf{x}}^{MMSE}-\mathbf{x}\right\|_2^2\right\}=\sum_{\mathcal{S}\in\Omega}P(\mathcal{S}|\mathbf{y})\cdot\left[|\mathcal{S}|c^2\sigma^2+\left\|\hat{\mathbf{x}}^{MMSE}-\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}\right\|_2^2\right]. \tag{43}$$

Using the unitary MMSE estimator expression in Equation (40) and that of the oracle solution in Equation (27), we further develop the second term in the expression above, and obtain

$$\begin{aligned}\left\|\hat{\mathbf{x}}^{MMSE}-\hat{\mathbf{x}}_{\mathcal{S}}^{Oracle}\right\|_2^2 &= \left\|c^2\sum_{k=1}^{n}g_k\beta_k\mathbf{e}_k-c^2\sum_{k=1}^{n}\mathbf{I}_{\mathcal{S}}(k)\beta_k\mathbf{e}_k\right\|_2^2\\ &= \sum_{k=1}^{n}c^4\left[g_k-\mathbf{I}_{\mathcal{S}}(k)\right]^2\beta_k^2\\ &= \sum_{k=1}^{n}c^4\left[g_k^2-2g_k\mathbf{I}_{\mathcal{S}}(k)+\mathbf{I}_{\mathcal{S}}(k)\right]\beta_k^2.\end{aligned} \tag{44}$$

Plugging this expression back into Equation (43), together with the expression for $P(\mathcal{S}|\mathbf{y})$ in Equation (34), gives

$$\begin{aligned}\text{MSE}\left(\hat{\mathbf{x}}^{MMSE}\right) &= \sum_{\mathcal{S}\in\Omega}P(\mathcal{S}|\mathbf{y})\left\{c^2\sigma^2\sum_{k=1}^{n}\mathbf{I}_{\mathcal{S}}(k)+c^4\sum_{k=1}^{n}\left[g_k^2-2g_k\mathbf{I}_{\mathcal{S}}(k)+\mathbf{I}_{\mathcal{S}}(k)\right]\beta_k^2\right\}\\ &= \sum_{k=1}^{n}\left\{c^2\sigma^2\sum_{\mathcal{S}\in\Omega}\mathbf{I}_{\mathcal{S}}(k)\prod_{i\in\mathcal{S}}g_i\prod_{j\notin\mathcal{S}}(1-g_j)\right.\\ &\quad\left.+c^4\beta_k^2\sum_{\mathcal{S}\in\Omega}\prod_{i\in\mathcal{S}}g_i\prod_{j\notin\mathcal{S}}(1-g_j)\left[g_k^2-2g_k\mathbf{I}_{\mathcal{S}}(k)+\mathbf{I}_{\mathcal{S}}(k)\right]\right\}\\ &= \sum_{k=1}^{n}c^2\sigma^2g_k+c^4\beta_k^2\left(g_k-g_k^2\right).\end{aligned} \tag{45}$$

Here we have exploited Proposition 2. Interestingly, the property $|\mathcal{S}|=\sum_{k=1}^{n}\mathbf{I}_{\mathcal{S}}(k)$ and Proposition 2 yield the relationship

$$\begin{aligned}E\left\{|\mathcal{S}|\right\} &= \sum_{\mathcal{S}\in\Omega}P(\mathcal{S}|y)|\mathcal{S}|\\ &= \sum_{\mathcal{S}\in\Omega}\left(\sum_{k=1}^{n}\mathbf{I}_{\mathcal{S}}(k)\right)\prod_{i\in\mathcal{S}}g_i\prod_{j\notin\mathcal{S}}(1-g_j)\\ &= \sum_{k=1}^{n}\left[\sum_{\mathcal{S}\in\Omega}\mathbf{I}_{\mathcal{S}}(k)\prod_{i\in\mathcal{S}}g_i\prod_{j\notin\mathcal{S}}(1-g_j)\right]=\sum_{k=1}^{n}g_k.\end{aligned} \tag{46}$$

This implies that the MMSE error can be alternatively written as

$$\text{MSE}\left(\hat{\mathbf{x}}^{MMSE}\right)=c^2\sigma^2E\left\{|\mathcal{S}|\right\}+c^4\sum_{k=1}^{n}\beta_k^2\left(g_k-g_k^2\right), \tag{47}$$

suggesting that the error is composed of an "oracle" error[7], and an additional part that is necessarily positive (since $0 < g_k < 1$). As an extreme example, if the elements of the vector β tend to be either very high or very low (compared to σ/c), then the g_k tend to the extremes as well. In such a case, the second term nearly vanishes, and the performance is close to that of the oracle.

We next study the MAP performance. Recall Equation (23), and note that $\hat{\mathbf{x}}^{MAP}$ may be written as

$$\hat{\mathbf{x}}^{MAP} = \sum_{k=1}^{n} \mathbf{I}_{\mathcal{MAP}}(k) c^2 \beta_k \mathbf{e}_k, \tag{48}$$

where $\mathbf{I}_{\mathcal{MAP}}(k)$ is an indicator function for the MAP support. Exploiting Propositions 1 and 2, we obtain the following expression for the MAP mean-squared-error,

$$\begin{aligned}
\text{MSE}\left(\hat{\mathbf{x}}^{MAP}\right) &= \sum_{\mathcal{S}\in\Omega} P(\mathcal{S}|\mathbf{y}) \left\{ c^2\sigma^2 \sum_{k=1}^{n} \mathbf{I}_{\mathcal{S}}(k) + c^4 \sum_{k=1}^{n} \left[\mathbf{I}_{\mathcal{MAP}}(k) - 2\mathbf{I}_{\mathcal{S}}(k)\mathbf{I}_{\mathcal{MAP}}(k) + \mathbf{I}_{\mathcal{S}}(k)\right] \beta_k^2 \right\} \\
&= \sum_{k=1}^{n} \Bigg\{ c^2\sigma^2 \sum_{\mathcal{S}\in\Omega} \mathbf{I}_{\mathcal{S}}(k) \prod_{i\in\mathcal{S}} g_i \prod_{j\notin\mathcal{S}} (1-g_j) \\
&\quad + c^4\beta_k^2 \sum_{\mathcal{S}\in\Omega} \prod_{i\in\mathcal{S}} g_i \prod_{j\notin\mathcal{S}} (1-g_j) \left[\mathbf{I}_{\mathcal{MAP}}(k) + \mathbf{I}_{\mathcal{S}}(k) - 2\mathbf{I}_{\mathcal{S}}(k)\mathbf{I}_{\mathcal{MAP}}(k)\right] \Bigg\} \\
&= \sum_{k=1}^{n} c^2\sigma^2 g_k + c^4\beta_k^2 \left[g_k + \mathbf{I}_{\mathcal{MAP}}(k)(1-2g_k)\right].
\end{aligned} \tag{49}$$

Analyzing the difference between the MMSE and MAP errors, in Equations (45) and (49) respectively, we find that only the last terms in each are different: $-g_k^2$ versus $\mathbf{I}_{\mathcal{MAP}}(k)(1-2g_k)$, respectively. Obviously, this implies $\text{MSE}\left(\hat{\mathbf{x}}^{MMSE}\right) \leq \text{MSE}\left(\hat{\mathbf{x}}^{MAP}\right)$, because $-g_k^2 \leq \mathbf{I}_{\mathcal{MAP}}(k)(1-2g_k)$ for any k, and regardless of the value of $\mathbf{I}_{\mathcal{MAP}}(k)$ (zero or one).

In order to further understand the estimators' performance given in the Equations (45) and (49), we turn now to a further analysis of these expressions and derive worst-case upper-bounds for them. The bounds we are about to build do not depend on the dimension of the signal, but rather on the problem parameters (σ, σ_x, P_i) alone. We begin with the MMSE, then turn to the MAP, and finally compare and discuss the resulting bounds.

4.2. MMSE Performance Bound

Referring to Equation (45), which describes the error associated with the MMSE approximation, we shall denote by MSE_1 the first term,

$$MSE_1 = c^2\sigma^2 \sum_{k=1}^{n} g_k. \tag{50}$$

As mentioned before, this is the expected MSE of the oracle (given $\mathbf{y}$). The second term, denoted by MSE_2, is given by

$$MSE_2 = c^4 \sum_{k=1}^{n} \beta_k^2 g_k (1-g_k). \tag{51}$$

[7] See the similarity between the first term here and the one posed in Equation (42).

This is the additional error due to the fact that the support is unknown. We would like to bound the ratio $r = MSE_2/MSE_1$, as this immediately yields a bound $(r+1)$ on the MMSE error in terms of the expected oracle error. Our goal is thus to characterize the worst ratio

$$\max_{\mathbf{y}\in R^n} r = \max_{\mathbf{y}\in R^n} \frac{MSE_2}{MSE_1}, \tag{52}$$

that is, the worst (largest) ratio over all conceivable signals $\mathbf{y}$, where the dependence on $\mathbf{y}$ enters via the β_k's. In order to characterize this ratio, we shall need the following simple lemma:

Lemma 3. *Let $(a_k, b_k), k = 1, \ldots, n$, be pairs of positive real numbers. Let m be the index of a pair whose ratio is maximal, i.e.,*

$$\frac{a_k}{b_k} \leq \frac{a_m}{b_m} \qquad \text{for all } k \geq 1. \tag{53}$$

Then

$$\frac{\sum_{k=1}^n a_k}{\sum_{j=1}^n b_j} \leq \frac{a_m}{b_m},$$

with equality occurring only if $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \cdots = \frac{a_n}{b_n}$.

Proof. By (53), $a_k b_m \leq a_m b_k$ for all $k \geq 1$, with equality obtained only if $a_k/b_k = a_m/b_m$. Summing up all these inequalities, we obtain

$$b_m \sum_{k=1}^n a_k \leq a_m \sum_{j=1}^n b_j \,,$$

hence,

$$\frac{\sum_{k=1}^n a_k}{\sum_{j=1}^n b_j} \leq \frac{a_m}{b_m},$$

as claimed, with equality occurring only if $\frac{a_i}{b_i} = \frac{a_m}{b_m}$, for every i. □

Returning to our task of bounding MSE_2/MSE_1, we observe that this ratio can be written as

$$\frac{MSE_2}{MSE_1} = \frac{c^4 \sum_{k=1}^n \beta_k^2 g_k (1-g_k)}{c^2\sigma^2 \sum_{k=1}^n g_k}, \tag{54}$$

which is of the same form as the ratio appearing in the Lemma. This leads us to the following Theorem:

Theorem 4. *Denote $G_k = \sqrt{1-c^2}P_k/(1-P_k)$, and let m be the index corresponding to an a priori least likely atom, i.e., $P_m = \min_{1\leq k\leq n} P_k$ and hence, $G_m = \min_{1\leq k\leq n} G_k$. Denote $f_{MMSE}(s) = \frac{2s}{1+G_m e^s}$, and define (implicitly) $s^\star = \arg\max_{s\geq 0} f_{MMSE}(s)$. Then*

1. *$r^\star = f_{MMSE}(s^\star)$ is an upper-bound on the ratio MSE_2/MSE_1.*
2. *The worst ratio, $r^\star$, satisfies the explicit bound*

$$r^\star \leq \begin{cases} 2\ln\left(\frac{1}{4G_m}\right) & G_m < \frac{1}{4e^2} \approx 0.034 \\ \frac{2}{\sqrt{G_m e}} & G_m \geq \frac{1}{4e^2} \approx 0.034 \end{cases}. \tag{55}$$

PROOF. Starting with the first claim, we embark from Equation (54) and exploit Lemma 3 to obtain

$$\begin{aligned}\frac{MSE_2}{MSE_1} &= \frac{c^4\sum_{k=1}^n \beta_k^2 g_k(1-g_k)}{c^2\sigma^2\sum_{k=1}^n g_k} \\ &\leq \frac{c^2}{\sigma^2}\cdot\max_{1\leq k\leq n}\frac{\beta_k^2 g_k(1-g_k)}{g_k} \\ &\leq \frac{c^2}{\sigma^2}\cdot\max_{1\leq k\leq n}\beta_k^2(1-g_k). \end{aligned} \tag{56}$$

Recalling that $g_k = q_k/(1+q_k)$, the definition of q_k in (31), and the definition of G_k above, we have

$$\begin{aligned}1-g_k = \frac{1}{1+q_k} &= \frac{1}{1+\frac{P_k}{1-P_k}\sqrt{1-c^2}\exp\left\{\frac{c^2}{2\sigma^2}\beta_k^2\right\}} \\ &= \frac{1}{1+G_k\exp\left\{c^2\beta_k^2/2\sigma^2\right\}}. \end{aligned} \tag{57}$$

Plugging this into Equation (56) and denoting $s = c^2\beta_k^2/2\sigma^2$, we obtain

$$\frac{MSE_2}{MSE_1} \leq \max_{1\leq k\leq n}\frac{2s}{1+G_k\exp\{s\}}. \tag{58}$$

This is a monotonically decreasing function of G_k for any fixed value of $s \geq 0$ (note that s must be non-negative, due to its definition). Thus, the maximum over the indices $1 \leq k \leq n$ is obtained for the index m for which G_k is the smallest. Therefore,

$$\max_{\beta}\frac{MSE_2}{MSE_1} \leq \max_{s\geq 0}\frac{2s}{1+G_m\exp\{s\}} = f_{MMSE}(s^\star) = r^\star, \tag{59}$$

as claimed.

Turning to the second claim of the theorem, we desire to bound $f_{MMSE}(s)$ from above. To this end, we maximize the alternative function $\overline{f}(s)$ that bounds $f_{MMSE}(s)$ from above point-wise:

$$f_{MMSE}(s) = \frac{2s}{1+G_m e^s} \leq \frac{2s}{\max\left(1, 2\sqrt{G_m e^s}\right)} \equiv \overline{f}(s). \tag{60}$$

Here we have used the facts that (i) the arithmetic mean $(1+G_m e^s)/2$ is necessarily larger than the geometric one, $\sqrt{G_m e^s}$, and (ii) $1+G_m e^s \geq 1$.

The switch-over in the denominator of $\overline{f}(s)$ occurs when $G_m e^s = 1/4$, which takes place for $s = s_0 \equiv \ln(1/4G_m)$. For $s \leq s_0$, $\overline{f}(s) = 2s$, which is monotonically increasing. For $s \geq s_0$, $\overline{f}(s) = s/\sqrt{G_m e^s}$, whose derivative is given by $\overline{f}'(s) = (1-s/2)/\sqrt{G_m e^s}$. Thus, if $s_0 \geq 2$, the maximum of $\overline{f}(s)$ occurs at $s = s_0$, being $\overline{f}(s_0) = 2s_0 = 2\ln(1/4G_m)$. Otherwise, the maximum occurs at $s = 2$, being $\overline{f}(2) = 2/\sqrt{G_m e}$. This proves the explicit upper bound on $r^\star$, as given in Equation (55). □

Figure 2 shows the functions $f_{MMSE}(s)$ and its upper bound $\overline{f}(s)$ for two possible values of G_m: 0.01 and 0.1. These two cases correspond to the two options covered in Equation (55). As can be seen, for $G_m = 0.01 < 0.034$, the maximum point is obtained on the linear part of $\overline{f}(s)$, whereas in the case of $G_m = 0.1 > 0.034$, the maximum is obtained for $s = 2$. Figure 3 presents the value of $r^\star$ as a function of G_m. This figure also shows the upper-bound on this value as given in Equation (55), and the two sub-functions that comprise it.

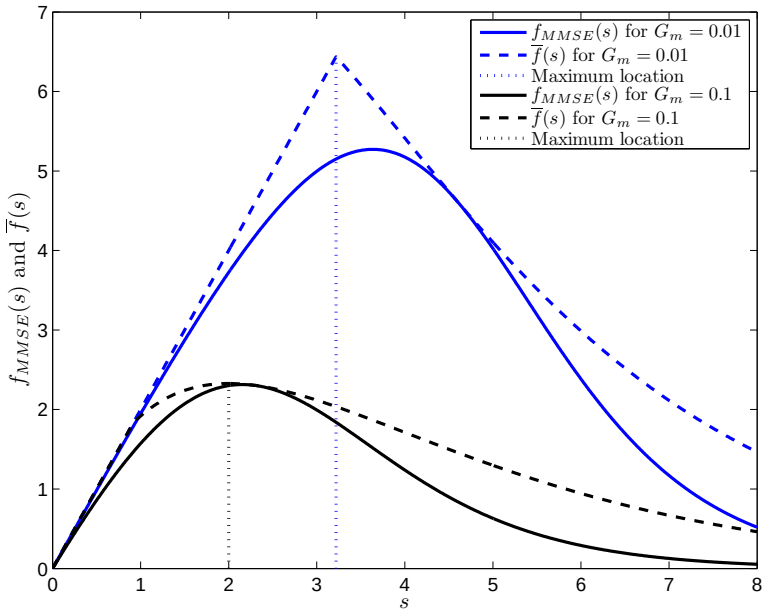


Figure 2: Graph plot of the function $f_{MMSE}(s)$ and its upper-bounding function $\overline{f}(s)$ (the solid and the dashed lines, respectively), exhibiting the two cases, where the maximum changes given the value G_m.

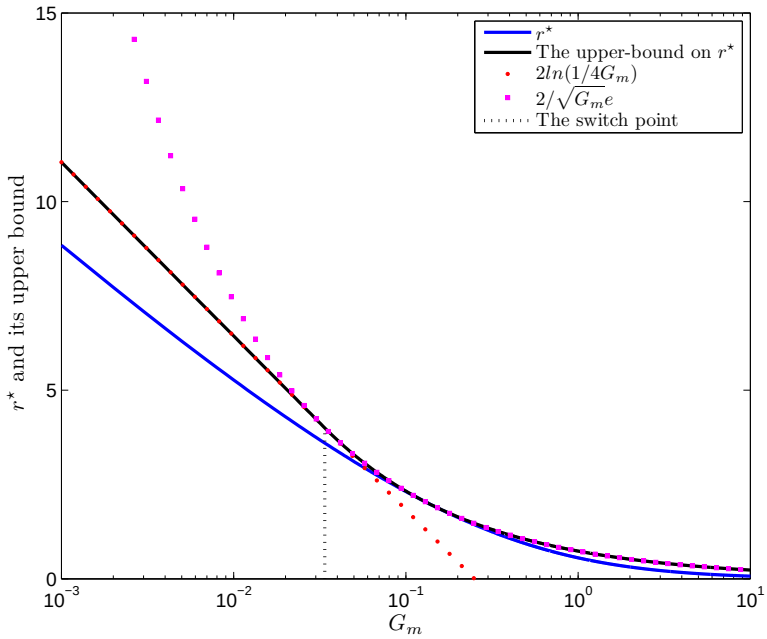


Figure 3: The worst ratio $r^\star$ and its upper bound, as given in Equation (55). This graph also shows the two portions of this bounding function, and the location of the switch between them.

Corollary 5. *The expected error for the MMSE estimator is bounded for any signal* $\mathbf{y}$ *by*

$$\text{MSE}\left(\hat{\mathbf{x}}^{MMSE}\right) \leq \text{MSE}\left(\hat{\mathbf{x}}^{Oracle}\right) \cdot \begin{cases} 1 + 2\ln\frac{1}{4G_m} & G_m \leq \frac{1}{4}e^{-2} \\ 1 + \frac{2}{\sqrt{G_m e}} & G_m \geq \frac{1}{4}e^{-2} \end{cases}. \tag{61}$$

PROOF. Follows from Theorem 4. □

What happens when all the probabilities P_k are equal? In such a case we obtain that $G_1 = G_2 = \cdots = G_n$. From Equation (56), which uses Lemma 3, it is obvious that the worst-ratio $r^\star$ becomes a tight upper-bound on MSE_2/MSE_1, since all the terms in the numerator and the denominator summations are equal. Furthermore, the worst-case β_k's are all equal to $\pm\sigma\sqrt{s^\star}/c$.

4.3. MAP Performance Bound

We next develop an upper-bound on the error associated with the MAP estimate in Equation (49). While MSE_1 remains the same as in Equation (50), the term that corresponds to MSE_2 for the MAP becomes

$$MSE_2 = c^4 \sum_{k=1}^{n} \beta_k^2 g_k \left[1 + \frac{\mathbf{I}_{\mathcal{MAP}}(k)(1-2g_k)}{g_k}\right].$$

Continuing with the same definitions as in the previous section, we prove a similar theorem for the expected MSE of the MAP estimator.

Theorem 6. *Denote* $G_k = \sqrt{1-c^2}P_k/(1-P_k)$*, and let* m *be the index corresponding to an a priori least likely atom, i.e.,* $P_m = \min_{1\leq k\leq n} P_k$ *and hence,* $G_m = \min_{1\leq k\leq n} G_k$*. Define the function*

$$f_{MAP}(s) = \begin{cases} 2s & G_m e^s < 1 \\ \frac{2s}{G_m e^s} & G_m e^s \geq 1 \end{cases}, \tag{62}$$

and define (implicitly) $s^\star = \arg\max_{s\geq 0} f_{MAP}(s)$*. Then*

1. $r^\star = f_{MAP}(s^\star)$ *is an upper-bound on the ratio* MSE_2/MSE_1*.*
2. *The worst ratio,* $r^\star$*, satisfies the explicit bound*

$$r^\star = \begin{cases} 2\ln\frac{1}{G_m} & G_m < e^{-1} \approx 0.368 \\ \frac{2}{G_m e} & G_m \geq e^{-1} \approx 0.368 \end{cases}. \tag{63}$$

PROOF. The proof follows the same lines as that of Theorem 4. Starting with the ratio r, we exploit Lemma 3 and obtain

$$\begin{aligned} \frac{MSE_2}{MSE_1} &= \frac{c^4 \sum_{k=1}^{n} \beta_k^2 g_k \left[1 + \mathbf{I}_{\mathcal{MAP}}(k)\frac{1-2g_k}{g_k}\right]}{c^2\sigma^2 \sum_{k=1}^{n} g_k} \\ &\leq \frac{c^2}{\sigma^2} \cdot \max_{1\leq k\leq n} \frac{\beta_k^2 g_k \left[1 + \mathbf{I}_{\mathcal{MAP}}(k)\frac{1-2g_k}{g_k}\right]}{g_k} \\ &\leq \frac{c^2}{\sigma^2} \cdot \max_{1\leq k\leq n} \beta_k^2 \left[1 + \mathbf{I}_{\mathcal{MAP}}(k)\frac{1-2g_k}{g_k}\right]. \end{aligned} \tag{64}$$

Again using the relation $g_k = q_k/(1+q_k)$ and the definition of q_k from (31), we have that

$$\frac{1-2g_k}{g_k} = \frac{1}{q_k} - 1 = \frac{1}{G_k \exp\left\{\frac{c^2\beta_k^2}{2\sigma^2}\right\}} - 1 = \frac{1}{G_k \exp\{s\}} - 1, \tag{65}$$

where we have used the definition of s as before ($s = c^2\beta_k^2/2\sigma^2$). Plugged back into Equation (64), we obtain

$$\frac{MSE_2}{MSE_1} \leq \max_{1\leq k\leq n} 2s\left[1 + \mathbf{I}_{\mathcal{MAP}}(k)\left(\frac{1}{G_k\exp\{s\}} - 1\right)\right]. \tag{66}$$

For any fixed value of s, the maximum over the indices $1 \leq k \leq n$ is obtained for the index m for which G_k is the smallest. Therefore, maximizing this expression with respect to both k and s yields

$$\begin{aligned}\max_{\beta}\frac{MSE_2}{MSE_1} &\leq \max_{s\geq 0} 2s\left[1 + \mathbf{I}_{\mathcal{MAP}}(m)\left(\frac{1}{G_m\exp\{s\}} - 1\right)\right] \\ &\leq \max_{s\geq 0}\begin{cases} 2s & G_m\exp\{s\} < 1 \\ \frac{2s}{G_m e^s} & G_m\exp\{s\} \geq 1 \end{cases}, .\end{aligned} \tag{67}$$

Here we have used the fact that $\mathbf{I}_{\mathcal{MAP}}(m) = 1$ when the atom m is part of the MAP support, which takes place if $q_m \geq 1$ (see the discussion after Equation (35)).

We turn to the second claim of the theorem, and calculate explicitly the value $s^\star$ for which $f_{MAP}(s^\star) = r^\star$ is maximized. The switch-over between the two cases of $f_{MAP}(s)$ occurs when $G_m e^s = 1$, that is, $s = s_0 \equiv \ln(1/G_m)$. For $s \leq s_0$, $f_{MAP}(s) = 2s$, which is monotonically increasing. For $s \geq s_0$, $f_{MAP}(s) = 2s/(G_m e^s)$, whose derivative is given by $f'(s) = (2-2s)/(G_m e^s)$. Thus, if $s_0 > 1$, the maximum of f occurs at $s^\star = s_0$, that is, $f_{MAP}(s_0) = 2\ln(1/G_m)$. Otherwise, the maximum occurs at $s^\star = 1$ with $f_{MAP}(1) = 2/(G_m e)$. This proves the explicit upper bound $r^\star$ as given in Equation (63). □

Figure 4 shows two examples of $f_{MAP}(s)$ for two possible values of G_m: 0.2 and 0.8. These two cases correspond to the two options covered in Equation (63). As can be seen, for $G_m = 0.2 < 0.368$, the maximum point of f_{MAP} is obtained at the switch-over point, whereas in the case of $G_m = 0.8 > 0.368$, the maximum is found at $s = 1$.

Figure 5 presents the value of $r^\star$ as a function of G_m for both the MAP and the MMSE. This figure also shows the two sub-functions that construct $r^\star$ for the MAP, as described in Equation (63).

Corollary 7. *The expected MSE error for the MAP estimator is bounded for any signal* $\mathbf{y}$ *by*

$$\mathrm{MSE}\left(\hat{\mathbf{x}}^{MAP}\right) \leq \mathrm{MSE}\left(\hat{\mathbf{x}}^{Oracle}\right)\cdot\begin{cases} 1 + 2\ln\frac{1}{G_m} & G_m \leq e^{-1} \\ 1 + \frac{2}{G_m e} & G_m \geq e^{-1} \end{cases}. \tag{68}$$

Proof. Follows from Theorem 6. □

When all the probabilities P_i are equivalent, and hence $G_1 = \cdots = G_n$, we get again that the worst ratio $r^\star$ becomes a tight upper bound on MSE_2/MSE_1, following the same reasoning as explained in the MMSE

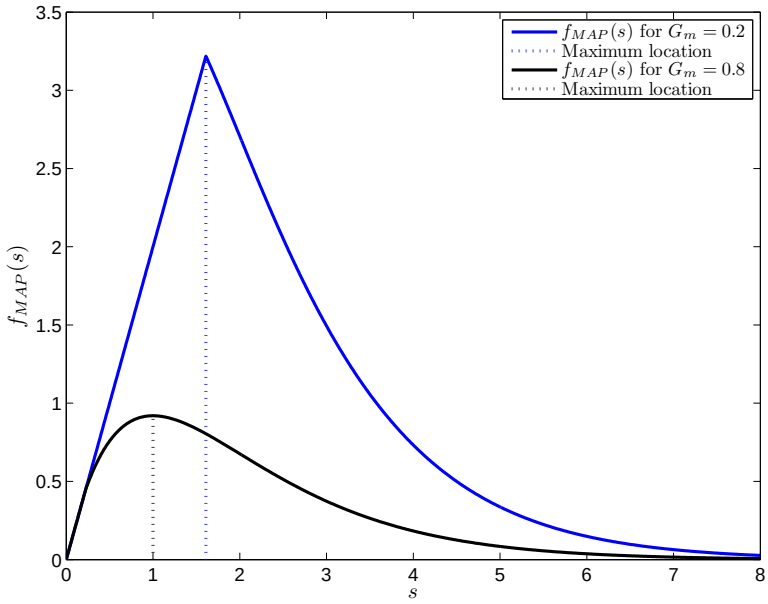


Figure 4: A plot of the function $f_{MAP}(s)$, exhibiting the two cases, where the maximum changes character according to the value G_m.

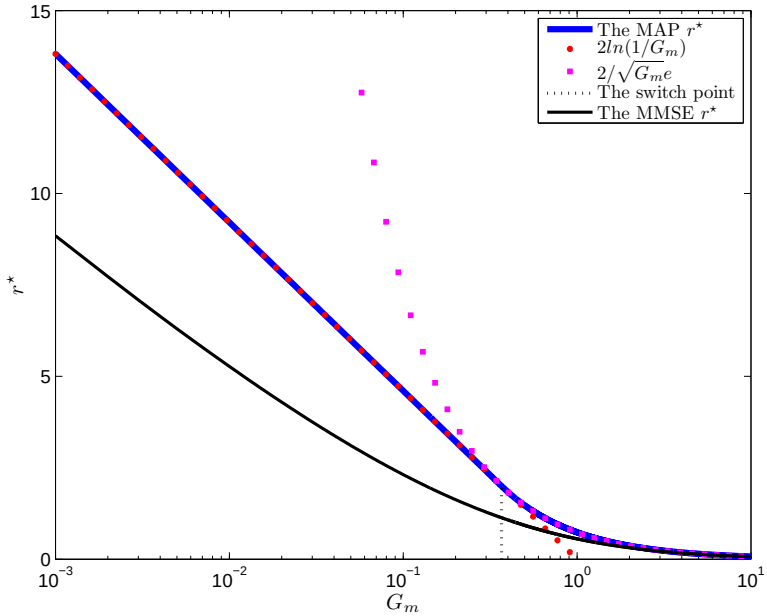


Figure 5: The worst ratio $r^\star$ for MAP, as given in Equation (63). This graph also shows the two portions of this function, the location of the switch between them, and the MMSE ratio $r^\star$.

case. The worst-case β_k's are all given by

$$\beta_k = \begin{cases} \pm\frac{2\sigma^2}{c^2}\sqrt{2\ln\left(\frac{1}{G_m}\right)} & G_m < e^{-1} \\ \pm\frac{2\sigma^2}{c^2} & G_m \geq e^{-1} \end{cases}.$$

4.4. MMSE and MAP Bounds – A Summary

The bounds developed above suggest that both the MMSE and the MAP estimators lead in the unitary case to a mean-squared error that is at worst a constant times the oracle MSE. The analysis given above provides exact expressions for these ratios.

We should note that the bounds developed above are based on a worst-case scenario. A more practical goal would be to bound the average case, as this should tell us more about the behavior of real-life signals. We leave this topic to future work.

As a last point in this section, we consider the following question: When are the MAP and MMSE nearly equivalent? Recall that the errors of these two estimators are given in Equations (47) and (49) as

$$\begin{aligned} \text{MSE}\left(\hat{\mathbf{x}}^{MMSE}\right) &= \sum_{k=1}^{n} c^2\sigma^2 g_k + c^4 \sum_{k=1}^{n} \beta_k^2 \left(g_k - g_k^2\right) \\ \text{MSE}\left(\hat{\mathbf{x}}^{MAP}\right) &= \sum_{k=1}^{n} c^2\sigma^2 g_k + c^4 \sum_{k=1}^{n} \beta_k^2 \left[g_k + \mathbf{I}_{\mathcal{MAP}}(k)(1-2g_k)\right]. \end{aligned}$$

In order for these two errors to be close, we should therefore impose for all k

$$g_k - g_k^2 \approx g_k + \mathbf{I}_{\mathcal{MAP}}(k)(1-2g_k) \quad \Rightarrow \quad g_k^2 - 2\mathbf{I}_{\mathcal{MAP}}(k)g_k + \mathbf{I}_{\mathcal{MAP}}(k) \approx 0. \tag{69}$$

If $P_k \to 0$, this leads to $g_k \to 0$, since $g_k = q_k/(1+q_k)$ and $q_k = \sqrt{1-c^2}P_k/(1-P_k) \cdot \exp c^2\beta_k^2/2\sigma^2$. From Equation (36) we also have that $\mathbf{I}_{\mathcal{MAP}}(k) = 0$, implying that this index is not part of the MAP support. Returning to the requirement posed in Equation (69), we obtain the condition $g_k^2 \approx 0$, which is readily satisfied. Thus, we conclude that one case where the two estimators, MAP and MMSE, align, is when $P_k \to 0$.

When $P_k \to 1$, this leads to $g_k \to 1$. Relying again on Equation (36) we also have that $\mathbf{I}_{\mathcal{MAP}}(k) = 1$ this time, implying that this index is now part of the MAP support. Returning to the requirement posed in Equation (69), we obtain the condition $g_k^2 - 2g_k + 1 = (g_k - 1)^2 \approx 0$, again satisfied (since g_k is close to 1. Thus, another case where the two estimators align is when $P_k \to 1$.

5. Experimental Results

Here we demonstrate the MAP and MMSE estimators for unitary dictionaries and provide both synthetic and real-signal experiments to illustrate these algorithms.

5.1. Synthetic Experiments

In the first experiment we use a 2D Wavelet dictionary $\mathbf{D}$ (Daubachies-5 filters) [10], with 3 levels of resolution. We choose all the atom probabilities P_i and all the variances σ_i to be the same in this test. We use $P = 0.1$ and $\sigma_x = 1$.

Generating a two-dimensional signal according to the proposed model is done by first randomly choosing whether each atom is part of the support or not with probability P. For the selected atoms, coefficients x_i are drawn independently from a normal distribution $\mathcal{N}(0, \sigma_x^2)$. The resulting sparse vector of coefficients is multiplied by the unitary dictionary to obtain the *ground-truth* two-dimensional signal. Each entry is independently contaminated by white Gaussian noise $\mathcal{N}(0, \sigma^2)$ to create the *input* signal $\mathbf{y}$. The values of the additive noise power, σ, are varied in the range $[0.1, 1]$ to demonstrate the effect of the noise level on the overall performance. Each of the (noisy) signals is then approximated using the following estimators:

1. **Empirical Oracle** estimation and its MSE. This estimator appears in Equation (27).
2. **Theoretical Oracle** estimation error, as given in Equation (42).
3. **Empirical MMSE** estimation and its MSE. We use Equation (40) in order to compute the estimation, and then assess its error empirically.
4. **Theoretical MMSE** estimation error, using Equation (45) directly.
5. **Empirical MAP** estimation and its MSE. We use the closed-form solution given in Equation (36).
6. **Theoretical MAP** estimation error, as given in Equation (49).

The above process is repeated for 1000 randomly generated signals of size 128×128, and the mean L_2 error is averaged over all signals to obtain an estimate of the expected quality of each estimator. Figure 6 shows the relative denoising effect (compared to the original noisy signal) achieved by each estimator. The improved performance of the MMSE estimator over the MAP is clearly seen, as well as a clear validation of the theoretical derivations.

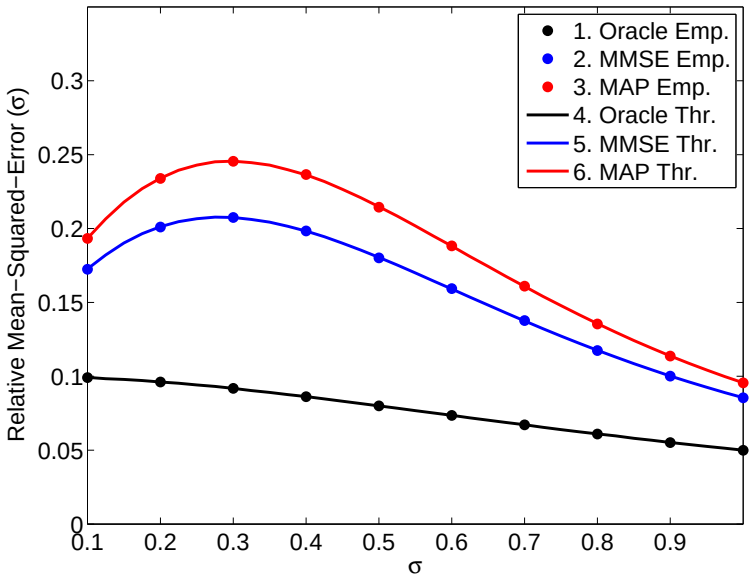


Figure 6: Empirical and theoretical evaluations of the MSE as a function of the input noise for synthetic signals ($P = 0.1$, $\sigma_x = 1$, and $n = 128 \times 128$).

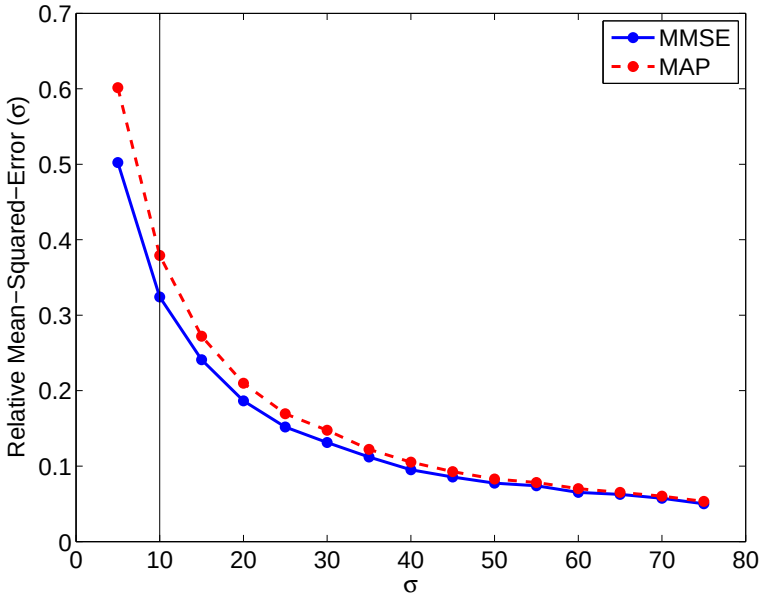


Figure 7: Relative denoising achieved by the MAP and the MMSE estimators, obtained for the image `Peppers`, with varying input noise power.

5.2. Real-World Signals

Next, we experiment with real-world signals – images. The unitary dictionary for this experiment is the same 2D Wavelet Transform dictionary used in the synthetic experiment. This dictionary is known to serve natural image content adequately (i.e., sparsify image content). There are two main obstacles when aiming to operate on non-synthetic signals:

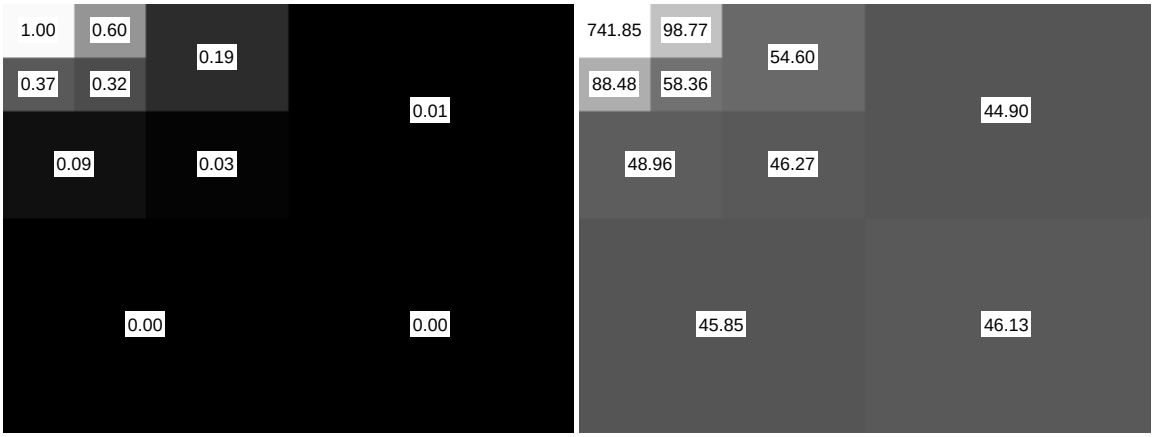
1. The assumption that all the non-zero entries in $\mathbf{x}$ share the same variance is inadequate, and we should generalize the above discussion to a heteroscedastic model.
2. The parameters that describe the signal model are unknown and need to be estimated from the corrupted signal.

Our handlng of these two issues is described in detail in Appendix A.

It is important to note that our main goal in this experiment is to demonstrate the power of the MMSE and the MAP estimators, and their comparison. We do not attempt to compare these results to state-of-the-art image denoising algorithms, as the current model is too limited for this comparison to be fair, due to the non-adaptiveness and the unitarity of the dictionary.

We experiment with the image `Peppers` shown in Figure 10. The noise levels considered are: 5, 10, 15, ... , 70, where the pixel values are in the range $[0, 255]$. The relative MSE of the cleaned image compared to the noisy one appears in Figure 7, as a function of the input noise power. Per each σ, the parameters are estimated, and then used within the MAP and the MMSE estimators.

Clearly, the MMSE outperforms the MAP for all the noise levels, the gap being bigger for high SNR levels. Nevertheless, it is also evident from this graph that the difference between the two is relatively small. Figure 8



(a) P_i (b) σ_i

Figure 8: Estimated parameters for the "Peppers" image with noise $\sigma = 10$.

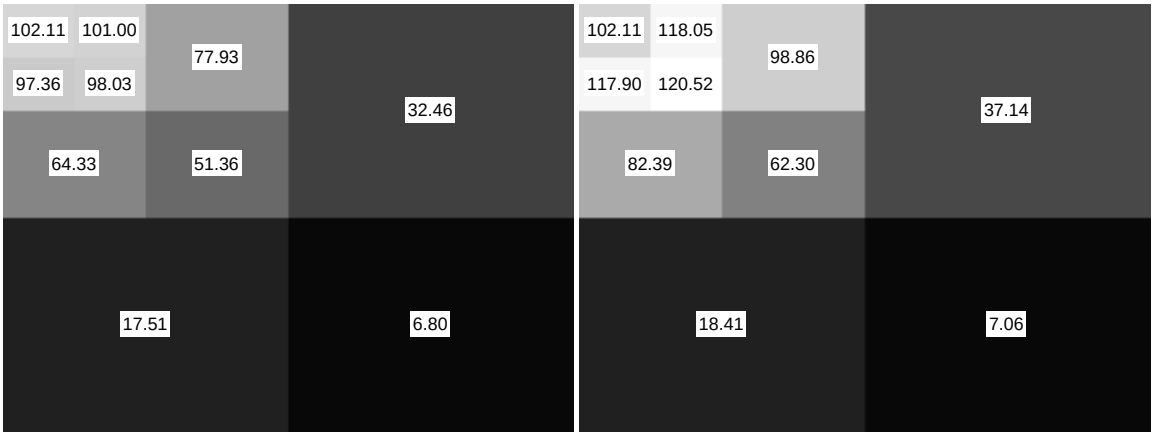
shows the estimated parameters learnt from the noisy frame for each band, and the values of these parameter may provide an explanation for this phenomenon.

As we have observed in the previous section, the gap between the MMSE and the MAP is expected to be negligible if P_k are nearly zeros or ones. This means that among the 10 bands in the wavelet transform, the three high-resolution and the single low-resolution bands are expected to give the same performance for both estimators. This suggests that the difference between the MAP and the MMSE is only due to the image energy that resides in the 6 middle-bands. Figure 9 shows the actual errors per band, as obtained by the MMSE and the MAP, and indeed, as expected, the difference in these errors exists mostly in the 6 middle bands.

Finally, a visual comparison of the results of the different estimators is presented in Figure 10 for the image `Peppers`, to which white Gaussian noise with $\sigma = 10$ is added. As expected, the MMSE result shows a small visual improvement over the MAP.

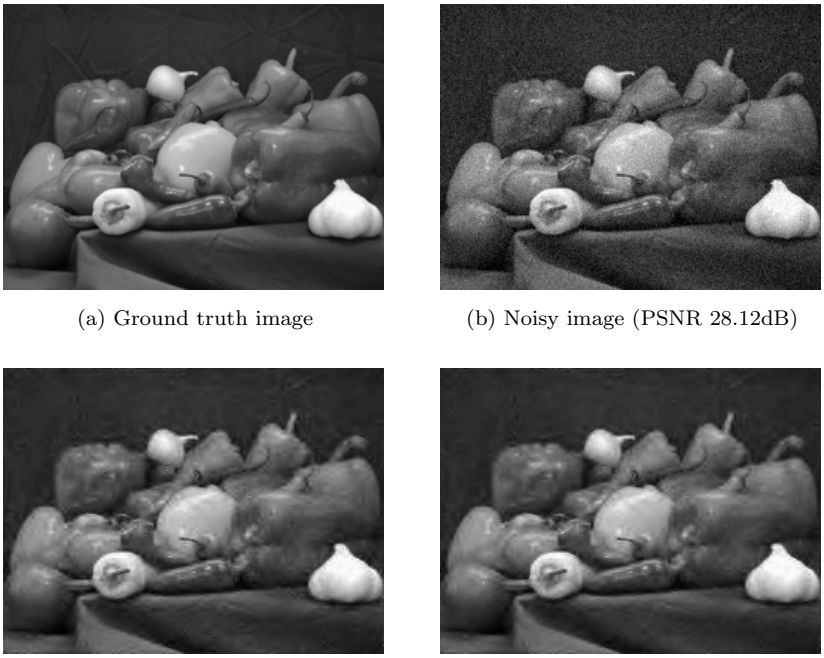
6. Summary and Conclusions

In this work we have studied a model where each atom has a given probability to be part of the support. This model assumes that all the supports are possible, thus avoiding assumptions on the (generally unknown) support size. We study MAP and MMSE estimators for the model with a general dictionary, including an overview of their performance. Then, we focus on unitary dictionaries, for which both estimators have simple and accurate closed formulas for their computation. After developing the closed-form MAP and MMSE estimators, it is shown how can they be interpreted in terms of shrinkage. We describe the relation of the MAP and MMSE estimators in this model to existing models appearing in the literature. This development is extended by looking at the theoretical performance of the estimators. Here, analytical bounds on the worst-case denoising performance is shown. Finally, synthetic and real-world experiments show the performance of the estimators, and the clear advantage of MMSE estimator over MAP estimator.



(a) MMSE (b) MAP

Figure 9: The error per band for the MMSE and the MAP estimators.



(a) Ground truth image (b) Noisy image (PSNR 28.12dB)

(c) MAP (PSNR 32.33dB) (d) MMSE (PSNR 33.01dB)

Figure 10: Visual comparison of the reconstructed image by the MAP and MMSE estimators ($\sigma = 10$).

Appendix A – Handling Images

As mentioned in Section 5, in order to handle a given noisy image, we should extend the model to allow for distinct variances for the different atoms, and we should also estimate the model parameters from the image. This appendix describes these two tasks.

A.1. Extension to Heteroscedastic Model

In the derivations in this paper we have assumed that all the non-zero entries in $\mathbf{x}$ have the same variance. As this is rarely the case for natural images, we treat now a more general problem, where this variance is atom-dependent. Such a model is known as heteroscedastic. Our goal is to show that most of the results remain of similar form, with modest changes. Thus, we shall keep the discussion in the section brief, and only state the main results.

We change the covariance matrix in Equation (2) to be a more general diagonal matrix $\mathbf{V}_{\mathcal{S}}$, given by

$$\mathbf{V}_{\mathcal{S}} = \text{diag}\left(\sigma^2_{\mathcal{S}_1}, \ldots, \sigma^2_{\mathcal{S}_k}\right), \tag{A-1}$$

where $k = |\mathcal{S}|$. For the general estimators developed in section 2.2, the changes due to this generalization are all absorbed in the matrices $\mathbf{Q}_{\mathcal{S}}$ and $\mathbf{C}_{\mathcal{S}}$, becoming

$$\begin{aligned} \mathbf{C}_{\mathcal{S}} &= \mathbf{D}_{\mathcal{S}}\mathbf{V}_{\mathcal{S}}\mathbf{D}_{\mathcal{S}}^T + \sigma^2\mathbf{I}_n, \\ \mathbf{Q}_{\mathcal{S}} &= \mathbf{V}_{\mathcal{S}}^{-1} + \frac{1}{\sigma^2}\mathbf{D}_{\mathcal{S}}^T\mathbf{D}_{\mathcal{S}}, \end{aligned}$$

and the relation between them in Equation (9) is still valid.

Moving to the unitary case, the matrix $\mathbf{Q}_{\mathcal{S}}$ is a diagonal matrix of the form

$$\mathbf{Q}_{\mathcal{S}} = \text{diag}\left(\frac{\sigma^2_{\mathcal{S}_1} + \sigma^2}{\sigma^2_{\mathcal{S}_1}\sigma^2}, \ldots, \frac{\sigma^2_{\mathcal{S}_k} + \sigma^2}{\sigma^2_{\mathcal{S}_k}\sigma^2}\right). \tag{A-2}$$

Its inversion, $\mathbf{Q}_{\mathcal{S}}^{-1}$, can easily be calculated, and the oracle solution becomes

$$\hat{\mathbf{x}}^{Oracle} = \text{diag}\left(c^2_{\mathcal{S}_1}, \ldots, c^2_{\mathcal{S}_k}\right) \cdot \beta_{\mathcal{S}}, \tag{A-3}$$

where $c^2_{\mathcal{S}_i} = \sigma^2_{\mathcal{S}_i}/(\sigma^2_{\mathcal{S}_i} + \sigma^2)$. The support of the MAP estimator is given by

$$\mathcal{S}^{\mathcal{MAP}} = \arg\max_{\mathcal{S}\in\Omega} \prod_{i\in\mathcal{S}} \sqrt{1-c_i^2} \cdot \frac{P_i}{1-P_i} \cdot \exp\left\{\frac{c_i^2}{2\sigma^2}\beta_i^2\right\} \prod_{j\notin\mathcal{S}} 1. \tag{A-4}$$

Lastly, the unitary MMSE estimate presented in Equation (40) becomes

$$\hat{\mathbf{x}}^{MMSE} = \sum_{k=1}^{n} c_k^2 \frac{q_k}{1+q_k}\beta_k\mathbf{e}_k, \tag{A-5}$$

where $q_k = \frac{P_k}{1-P_k}\sqrt{1-c_k^2}\exp\left\{\frac{c_k^2}{2\sigma^2}\beta_k^2\right\}$.

A.2. Parameter Estimation

The parameters of the image generation model are not known in advance and thus they should be estimated. We shall assume that each band in the wavelet transform is characterized by a pair of parameters σ_i, P_i, and there are r such bands overall (10 in the experiment reported in Section 5). We propose to estimate these parameters directly from the noisy image, by performing the following optimization task:

$$\arg \max_{\{P_i,\sigma_i\}_{i=1}^r} P\left(\mathbf{y} \,|\{P_i,\sigma_i\}_{i=1}^r\right). \tag{A-6}$$

Marginalization of this likelihood term with respect to the support of the image in the wavelet domain reads

$$P\left(\mathbf{y} \,|\{P_i,\sigma_i\}_{i=1}^r\right) = \sum_{\mathcal{S}} P\left(\mathbf{y} \,|\mathcal{S}, \{P_i,\sigma_i\}_{i=1}^r\right) \cdot P\left(\mathcal{S} \,|\{P_i,\sigma_i\}_{i=1}^r\right). \tag{A-7}$$

As maximization of this summation may be computationally difficult, we turn to approximate it by considering only one item – the dominant one within this sum. Thus, we propose to solve

$$\begin{aligned} \arg \max_{\{P_i,\sigma_i\}_{i=1}^r} \quad & P\left(\mathbf{y} \,|\{P_i,\sigma_i\}_{i=1}^r\right) \\ & \approx \arg \max_{\{P_i,\sigma_i\}_{i=1}^r,\mathcal{S}} P\left(\mathbf{y} \,|\mathcal{S}, \{P_i,\sigma_i\}_{i=1}^r\right) \cdot P\left(\mathcal{S} \,|\{P_i,\sigma_i\}_{i=1}^r\right) \cdot P\left(\mathcal{S}\right), \end{aligned} \tag{A-8}$$

where we maximize with respect to the support as well. Note that we have introduced a prior on the support size, $P\left(\mathcal{S}\right)$. We shall use the form

$$P\left(\mathcal{S}\right) = \prod_{i=1}^{r} \exp\left\{-\lambda_i |\mathcal{S}_i|\right\},$$

with $\mathcal{S}_i$ the support in the i-th band. This prior controls the support sparsity in each band, and as we show next, it stabilizes the estimation procedure. The values λ_i are set to be high for low-frequency bands, and decrease for the higher frequency bands.

We use the model definitions in Section 2.1 in order to develop an expression that depends only on the parameters of the r bands. Starting with $P\left(\mathcal{S} \,|\{P_i,\sigma_i\}_{i=1}^r\right)$, we get

$$P\left(\mathcal{S} \,|\{P_i,\sigma_i\}_{i=1}^r\right) = \prod_{i=1}^{r} P_i^{|\mathcal{S}_i|} \left(1-P_i\right)^{n_i-|\mathcal{S}_i|}, \tag{A-9}$$

where n_i is the size of the i-th band. Using the fact that the wavelet dictionary is unitary and exploiting Equation (A-2), we have

$$\det\left(\mathbf{C}_{\mathcal{S}}\right) = \left(\frac{\sigma_i^2+\sigma^2}{\sigma^2}\right)^{|\mathcal{S}|} \sigma^{2n} = \sigma^{2n} \prod_{i\in\mathcal{S}} \frac{\sigma_i^2+\sigma^2}{\sigma^2} = \sigma^{2n} \prod_{i=1}^{r} \left(\frac{\sigma_i^2+\sigma^2}{\sigma^2}\right)^{|\mathcal{S}_i|}. \tag{A-10}$$

Plugging Equation (5) and the above expressions into (A-8), the parameters estimation task becomes

$$\arg \max_{\mathcal{S},\{P_i,\sigma_i\}_{i=1}^r} \prod_{i=1}^{r} \left(\frac{\sigma_i^2+\sigma^2}{\sigma^2}\right)^{-\frac{|\mathcal{S}_i|}{2}} P_i^{|\mathcal{S}_i|} \left(1-P_i\right)^{n_i-|\mathcal{S}_i|} \exp\left\{\frac{1}{2\sigma^2}\frac{\sigma_i^2}{\sigma^2+\sigma_i^2}\left\|\beta_{\mathcal{S}_i}\right\|^2 - \lambda_i|\mathcal{S}_i|\right\}.$$

Two important features of this expression deserve our attention: First, rather than seeking the support $\mathcal{S}$, this expression reveals that all we need are the cardinalities $|\mathcal{S}|$ within each band. Second, this expression is

separable with respect to the r bands, implying that we can estimate P_i, σ_i for the i-th band by solving

$$\arg \max_{|\mathcal{S}_i|, P_i, \sigma_i} \left(\frac{\sigma_i^2 + \sigma^2}{\sigma^2} \right)^{-\frac{|\mathcal{S}_i|}{2}} P_i^{|\mathcal{S}_i|} (1 - P_i)^{n_i - |\mathcal{S}_i|} \exp \left\{ \frac{1}{2\sigma^2} \frac{\sigma_i^2}{\sigma^2 + \sigma_i^2} \left\| \beta_{\mathcal{S}_i} \right\|^2 - \lambda_i |\mathcal{S}_i| \right\}.$$

Taking the log of the above expression, we obtain an alternative function to maximize,

$$\begin{aligned} f\left(|\mathcal{S}_i|, P_i, \sigma_i\right) = \quad & -\frac{|\mathcal{S}_i|}{2} \log \left(\frac{\sigma_i^2 + \sigma^2}{\sigma^2} \right) + |\mathcal{S}_i| \log P_i \\ & + (n_i - |\mathcal{S}_i|) \log (1 - P_i) + \frac{1}{2\sigma^2} \frac{\sigma_i^2}{\sigma^2 + \sigma_i^2} \left\| \beta_{\mathcal{S}_i} \right\|^2 - \lambda_i |\mathcal{S}_i|. \end{aligned} \tag{A-11}$$

To obtain the estimates for σ_i and P_i we differentiate f with respect to these unknowns. The derivative with respect to P_i leads to

$$0 = \frac{\partial f\left(|\mathcal{S}_i|, P_i, \sigma_i\right)}{\partial P_i} = \frac{|\mathcal{S}_i|}{P_i} - \frac{(n_i - |\mathcal{S}_i|)}{1 - P_i} \quad \Longrightarrow \quad P_i = \frac{|\mathcal{S}_i|}{n_i}. \tag{A-12}$$

Similarly, the derivative with respect to σ_i gives

$$0 = \frac{\partial f\left(|\mathcal{S}_i|, P_i, \sigma_i\right)}{\partial \sigma_i} = -|\mathcal{S}_i| \frac{\sigma_i}{\sigma_i^2 + \sigma^2} + \frac{\sigma_i}{\left(\sigma^2 + \sigma_i^2\right)^2} \left\| \beta_{\mathcal{S}_i} \right\|^2 \quad \Longrightarrow \quad \sigma_i^2 = \frac{\left\| \beta_{\mathcal{S}_i} \right\|^2}{|\mathcal{S}_i|} - \sigma^2. \tag{A-13}$$

The last step in this estimation process is to discover the cardinality $|\mathcal{S}_i|$. Returning to the expression to be maximized in Equation (A-11), we can plug in the solutions obtained for P_i and σ_i, both being functions of $|\mathcal{S}_i|$. The overall expression is thus a function of the scalar $|\mathcal{S}_i|$, and the maximizer value can be found by a simple sweep of this unknown in the range $[0, n_i]$. We should note that for every value tested, we should also update the vector β_S to include only non-zero elements of $\mathcal{S}_i$. Since we are maximizing $f\left(|\mathcal{S}_i|, P_i, \sigma_i\right)$, we should choose the largest entries (in absolute value) within this vector. After this exhaustive process is done, we pick the support size and the respective calculated parameters that maximize the optimization task (A-11).

[1] F. Abramovich, T. Sapatinas and B.W. Silverman, Wavelet thresholding via a Bayesian approach, *J. R. Statist. Soc. B*, 60:725–749, 1998.

[2] A. Antoniadis, J. Bigot, and T. Sapatinas, Wavelet estimators in nonparametric regression: a comparative simulation study, *J. Stat. Software*, 6(6):1–83 , 2001.

[3] Z. Ben-Haim, Y.C. Eldar, and M. Elad, Coherence-based performance guarantees for estimating a sparse vector under random noise, submitted to *IEEE Transactions on Signal Processing.*

[4] A.M. Bruckstein, D. L. Donoho, and M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, *SIAM review*, 51(1):34–81, February 2009.

[5] E.J. Candes and T. Tao, The Danzig-Selector: Statistical estimation when p is much larger than n, *Annals. Statistics*, 35(6):2313–2351, 2007.

[6] S.S. Chen, D.L. Donoho, and M.A. Saunders, Atomic decomposition by basis pursuit, *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[7] M. Clyde and E.I. George, Empirical Bayes estimation in wavelet nonparametric regression. In Bayesian Inference in Wavelet Based Models, P. Muller and B. Vidakovic (Eds.), Lect. Notes Statist., 141:309–322, New York: Springer-Verlag, 1998.

[8] M. Clyde and E.I. George, Flexible empirical Bayes estimation for wavelets, *J. R. Statist. Soc. B*, 62:681–698, 2000.

[9] M. Clyde, G. Parmigiani and B. Vidakovic, Multiple shrinkage and subset selection in wavelets, *Biometrika*, 85:391–401, 1998.

[10] I. Daubechies, *Ten lectures on wavelets*, SIAM, 1992.

[11] D.L. Donoho and I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, 81(3):425–455, September 1994.

[12] M. Elad and I. Yavneh, A plurality of sparse representations is better than the sparsest one alone, *IEEE Trans. on Information Theory*, 55(10):4701–4714, October 2009.

[13] S.M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Volume I, Prentice Hall, 1993.

[14] E. Larsson and Y. Selen, Linear regression with a sparse parameter vector, *IEEE Transactions on Signal Processing*, 55:451–460, 2007.

[15] S. Mallat and Z. Zhang, Matching Pursuits with time-frequency dictionaries, *IEEE Trans. on Signal Processing*, 41(12):3397–3415, 1993.

[16] P. Moulin and J. Liu, Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors, *IEEE Trans. Inf. Theory*, 45(3):909–919, April 1999.

[17] B.K. Natarajan, Sparse approximate solutions to linear systems, *SIAM Journal on Computing*, 24:227–234, 1995.

[18] M. Protter, I. Yavneh, and M. Elad, Closed-form MMSE estimator for denoising signals under sparse reconstruction modeling, *Eleventh IEEEI conference*, Eilat, Israel, Dec. 2008.

[19] M. Protter, I. Yavneh, and M. Elad, Closed-form MMSE estimation for signal denoising under sparse representation modeling over a unitary dictionary, submitted to *IEEE Transactions on Signal Processing*.

[20] P. Schnitter, L. C. Potter, and J. Ziniel, Fast Bayesian matching pursuit, Proc. Workshop on Information Theory and Applications (ITA), (La Jolla, CA), Jan. 2008.

[21] E.P. Simoncelli and E.H. Adelson, Noise removal via Bayesian wavelet coring, in Proc. ICIP, Laussanne, Switzerland, pp. 379-382, September 1996.