Sparse Approximation, Denoising, and Large Random Frames

Alyson K. Fletcher*, Sundeep Rangan† and Vivek K Goyal‡

alyson@eecs.berkeley.edu, s.rangan@flarion.com, vgoyal@mit.edu

*University of California, Berkeley, Berkeley, CA 94720 USA
†Flarion Technologies, Bedminster, NJ 07921 USA
‡Massachusetts Institute of Technology, Cambridge, MA 02139 USA

ABSTRACT

If a signal x is known to have a sparse representation with respect to a frame, the signal can be estimated from a noise-corrupted observation y by finding the best sparse approximation to y. The ability to remove noise in this manner depends on the frame being designed to efficiently represent the signal while it *inefficiently* represents the noise. This paper analyzes the mean squared error (MSE) of this denoising scheme and the probability that the estimate has the same sparsity pattern as the original signal. Analyses are for dictionaries generated randomly according to a spherically-symmetric distribution. Easily-computed approximations for the probability of selecting the correct dictionary element and the MSE are given. In the limit of large dimension, these approximations have simple forms. The asymptotic expressions reveal a critical input signal-to-noise ratio (SNR) for signal recovery.

Keywords: dictionary-based representations, estimation, isotropic random matrices, nonlinear approximation, stable signal recovery, subspace fitting

1. INTRODUCTION

Estimating a signal from a noise-corrupted observation of the signal is a recurring task in science and engineering. This paper explores the limits of estimation performance in the case where the only *a priori* structure on the signal $x \in \mathbb{R}^N$ is that it has known sparsity K with respect to a given set of vectors $\Phi = \{\varphi_i\}_{i=1}^M \subset \mathbb{R}^N$. The set Φ is called a *dictionary* and is generally a *frame*.[1,2] The sparsity of K with respect to Φ means that the signal x lies in the set

$$\Phi_K \;=\; \left\{ v \in \mathbb{R}^N \;\middle|\; v = \sum_{i=1}^{M} \alpha_i \varphi_i \quad \text{with at most } K \text{ nonzero } \alpha_i\text{'s} \right\}. \tag{1}$$

In many areas of computation, exploiting sparsity is motivated by reduction in complexity; if $K \ll N$ then certain computations may be more efficiently made on α than on x. In compression, representing a signal exactly or approximately by a member of Φ_K is a common first step in efficiently representing the signal, though much more is known when Φ is a basis or union of wavelet bases than is known in the general case.[3] Of more direct interest here is that sparsity models are becoming prevalent in estimation problems.[4,5]

1.1. Denoising by Sparse Approximation with a Frame

Consider the problem of estimating a signal $x \in \mathbb{R}^N$ from the noisy observation $y = x + d$ where $d \in \mathbb{R}^N$ has the i.i.d. Gaussian $\mathcal{N}(0, \sigma^2 I_N)$ distribution. Suppose we know that x lies in given K-dimensional subspace of $\mathbb{R}^N$. Then projecting y to the given subspace would remove a fraction of the noise without affecting the signal component. Denoting the projection operator by P, we would have

$$\hat{x} \;=\; Py \;=\; P(x+d) \;=\; Px + Pd \;=\; x + Pd,$$

and Pd has only K/N fraction of the power of d.

In this paper we consider the more general signal model $x \in \Phi_K$. The set Φ_K defined in (1) is the union of at most $J = \binom{M}{K}$ subspaces of dimension K. We henceforth assume $M > K$ (thus $J > 1$); if not, the model reduces to the classical case of knowing a single subspace that contains x.

With the addition of the noise d, the observed vector y will (almost surely) not be represented sparsely, *i.e.*, not be in Φ_K. Intuitively, a good estimate for x is the point from Φ_K that is closest to y in Euclidean distance. Formally, because the probability density function of d is a strictly decreasing function of $\|d\|_2$, this is the maximum likelihood estimate of x given y. The estimate is obtained by applying an optimal sparse approximation procedure to y. We will write

$$\hat{x}_{\mathrm{SA}} = \operatorname*{argmin}_{x \in \Phi_K} \|y - x\|_2 \tag{2}$$

for this estimate and call it the optimal K-term approximation of y. Henceforth we omit the subscript 2 indicating the Euclidean norm.

The results of this paper are analyses of the per-component mean-squared estimation error $\frac{1}{N}\mathbf{E}\left[\|x - \hat{x}_{\mathrm{SA}}\|^2\right]$ for denoising via sparse approximation and of the probability that x and $\hat{x}_{\mathrm{SA}}$ lie in the same K-dimensional subspace of Φ_K. Though the estimator does not use the distribution of x, the analysis includes an expectation over the randomly generated dictionary Φ. Our first results on denoising by sparse approximation, which were for arbitrary dictionaries rather than averaging over random dictionaries, were presented in Ref. 6. The results in the present paper are excerpted from Refs. 7–9. In particular, the reader is referred to Ref. 9 for all proofs.

1.2. Related Work

Computing optimal K-term approximations is generally a difficult problem. Given $\epsilon \in \mathbb{R}^+$ and $K \in \mathbb{Z}^+$, to determine if there exists a K-term approximation $\hat{x}$ such that $\|x - \hat{x}\| \leq \epsilon$ is an NP-complete problem.[10, 11] This computational intractability of optimal sparse approximation has prompted study of heuristics. A greedy heuristic has been known as *matching pursuit* in the signal processing literature since the work of Mallat and Zhang.[12] Also, Chen, Donoho and Saunders proposed a convex relaxation of the approximation problem (2) called *basis pursuit*.[13]

Two related discoveries have touched off a flurry of recent research:

(a) *Stability of sparsity*—Under certain conditions, the positions of the nonzero entries in a sparse representation of a signal are stable: applying optimal sparse approximation to a noisy observation of the signal will give a coefficient vector with the original support. Typical results are upper bounds (functions of the norm of the signal and the coherence of the dictionary) on the norm of the noise that allows a guarantee of stability.[14–18]

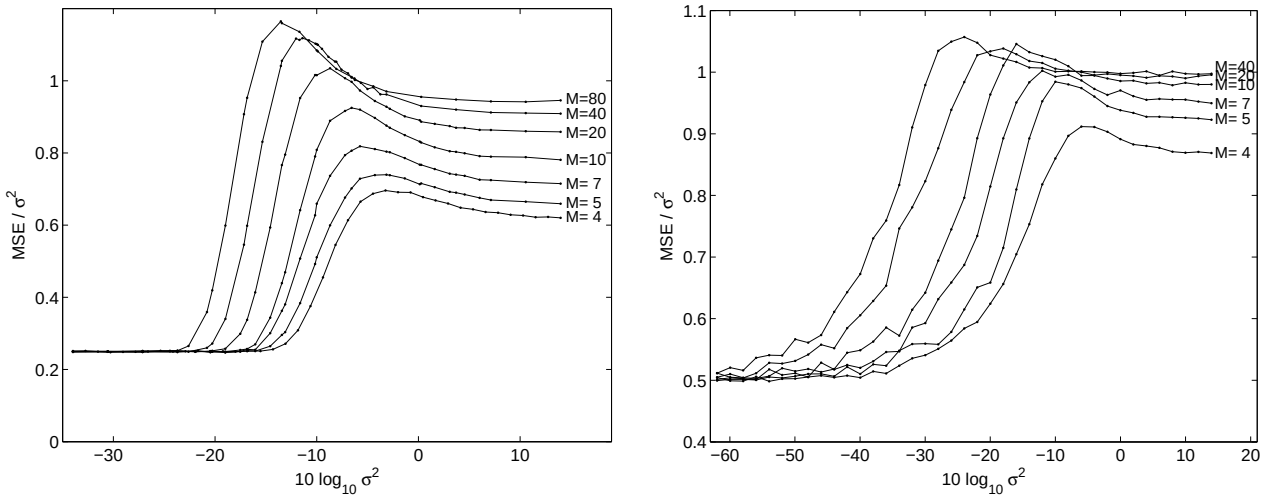
(b) *Effectiveness of heuristics*—Both basis pursuit and matching pursuit are able to find optimal sparse approximations, under certain conditions on the dictionary and the sparsity of signal.[17–22]

To contrast: in this paper we consider noise with unbounded support and thus a positive probability of failing to satisfy a sufficient condition for stability as in (a) above; and we do not address algorithmic issues in finding sparse approximations. It bears repeating that finding optimal sparse approximations is presumably computationally intractable except in the cases where a greedy algorithm or convex relaxation happens to succeed. Our results are thus bounds on the performance of the algorithms that one would probably use in practice.

Denoising by finding a sparse approximation is similar to the concept of denoising by compression popularized by Saito[23] and Natarajan.[24] More recent works in this area include those by Krim *et al.*,[25] Chang *et al.*[26] and Liu and Moulin.[27] All of these works use bases rather than frames. To put the present work into a similar framework would require a "rate" penalty for redundancy. Instead, the only penalty for redundancy comes from choosing a subspace that does not contain the true signal ("overfitting" or "fitting the noise").

1.3. Preview of Results and Outline

To motivate the paper, we present a set of numerical results from Monte Carlo simulations. In these experiments, N, M, and K are small because of the high complexity of computing optimal approximations and because a large number of independent trials is needed to get adequate precision. Each data point shown is the average of 100 000 trials.



(a) $K = 1$ (b) $K = 2$

Figure 1. Performance of denoising by sparse approximation when the true signal $x \in \mathbb{R}^4$ has an exact K-term representation with respect to a dictionary that is an optimal M-element Grassmannian packing.

Consider a true signal $x \in \mathbb{R}^4$ ($N = 4$) that has an exact 1-term representation ($K = 1$) with respect to M-element dictionary Φ. We observe $y = x + d$ with $d \sim \mathcal{N}(0, \sigma^2 I_4)$ and compute estimate $\hat{x}_{\mathrm{SA}}$ from (2). The signal is generated with unit norm so that the signal-to-noise ratio (SNR) is $1/\sigma^2$ or $-10 \log_{10} \sigma^2$ dB. Throughout we use the following definition for mean-squared error:

$$\mathrm{MSE} = \frac{1}{N} \mathbf{E} \left[\|x - \hat{x}_{\mathrm{SA}}\|^2 \right].$$

To have tunable M, we used dictionaries that are M maximally separated unit vectors in $\mathbb{R}^N$, where separation is measured by the minimum pairwise angle among the vectors and their negations. These are cases of Grassmannian packings[28, 29] in the simplest case of packing one-dimensional subspaces (lines). We used packings tabulated by Sloane with Hardin, Smith and others.[30]

Figure 1(a) shows the MSE as a function of σ for several values of M. Note that for visual clarity, $\mathrm{MSE} \,/\, \sigma^2$ is plotted, and all of the same properties are illustrated for $K = 2$ in Figure 1(b). For small values of σ, the MSE is $(1/4)\sigma^2$. This is an example of the general statement that

$$\mathrm{MSE} = \frac{K}{N} \sigma^2 \qquad \text{for small } \sigma,$$

as described in detail in Section 2. For large values of σ, the scaled MSE approaches a constant value:

$$\lim_{\sigma \to \infty} \frac{\mathrm{MSE}}{\sigma^2} = g_{K,M},$$

where $g_{K,M}$ is a slowly increasing function of M and $\lim_{M \to \infty} g_{K,M} = 1$. This limiting value makes sense because in the limit $\hat{x}_{\mathrm{SA}} \approx y = x + d$ and each component of d has variance σ^2; the denoising does not do anything. The characterization of the dependence of $g_{K,M}$ on K and M was the focus of our earlier work, reported in preliminary form in Ref. 6 and finally in Ref. 9.

Another apparent pattern in Figure 1 that we would like to explain is the transition between low and high SNR behavior. The transition occurs at smaller values of σ for larger values of M. Also, $\mathrm{MSE} \,/\, \sigma^2$ can exceed 1, so in fact the sparse approximation procedure can *increase* the noise. We are not able to characterize the transition well for arbitrary sequences of frames; in Section 3 we obtain results for large frames that are generated by choosing vectors uniformly at random from the unit sphere in $\mathbb{R}^N$. There we get a sharp transition between low and high SNR behavior.

2. PRELIMINARY COMPUTATIONS

Recall from the introduction that we are estimating a signal $x \in \Phi_K \subset \mathbb{R}^N$ from an observation $y = x + d$ where $d \sim \mathcal{N}(0, \sigma^2 I_N)$. Φ_K was defined in (1) as the set of vectors that can be represented as a linear combination of K vectors from $\Phi = \{\varphi_m\}_{m=1}^M$. We are studying the performance of the estimator

$$\hat{x}_{\mathrm{SA}} = \operatorname*{argmin}_{x \in \Phi_K} \|y - x\|.$$

This estimator is the maximum likelihood estimator of x in this scenario in which d has a Gaussian density and the estimator has no probabilistic prior information on x. The subscript SA denotes "sparse approximation" because the estimate is obtained by finding the optimal sparse approximation of y. There are values of y such that $\hat{x}_{\mathrm{SA}}$ is not uniquely defined. These collectively have probability zero and we ignore them.

Finding $\hat{x}_{\mathrm{SA}}$ can be viewed as a two-step procedure: first, find the subspace spanned by K elements of Φ that contains $\hat{x}_{\mathrm{SA}}$; then, project y to that subspace. The identification of a subspace and the orthogonality of $y - \hat{x}_{\mathrm{SA}}$ to that subspace will be used in our analyses. Let $\mathcal{P}_K = \{P_i\}_i$ be the set of the projections onto subspaces spanned by K of the M vectors in Φ. Then $\mathcal{P}_K$ has at most $J = \binom{M}{K}$ elements,* and the estimate of interest is given by

$$\hat{x}_{\mathrm{SA}} \;=\; P_T y, \qquad \text{where} \qquad T \;=\; \operatorname*{argmax}_i \|P_i y\|. \tag{3}$$

The distribution of the error $x - \hat{x}_{\mathrm{SA}}$ and the average performance of the estimator both depend on the true signal x. To remove this dependence, the performance measure analyzed here is the conditional MSE

$$e(x) \;=\; \frac{1}{N}\mathbf{E}\left[\|x - \hat{x}_{\mathrm{SA}}\|^2 \;\middle|\; x\right]. \tag{4}$$

In the case that T is independent of d, the projection in (3) is to a fixed K-dimensional subspace, so

$$e(x) \;=\; \frac{K}{N}\,\sigma^2. \tag{5}$$

This occurs when $M = K$ (there is just one element in $\mathcal{P}_K$) or in the limit of high SNR (small σ^2). In the latter case, the subspace selection is determined by x, unperturbed by d.

3. ANALYSIS FOR ISOTROPIC RANDOM FRAMES

In general, the performance of sparse approximation denoising is given by

$$e(x) = \frac{1}{N}\mathbf{E}\left[\|x - \hat{x}_{\mathrm{SA}}\|^2 \mid x\right] = \frac{1}{N}\int_{R^{\mathbb{N}}} \left\| x - \left(\operatorname*{argmin}_{\hat{x} \in \Phi_K} \|x + \eta - \hat{x}\|_2 \right) \right\|^2 f(\eta)\, d\eta$$

where $f(\cdot)$ is the density of the noise d. While this expression does not give any fresh insight, it does remind us that the performance depends on every element of Φ. In this section, we attack this dependence by assuming that each dictionary element is an independent random vector and that the dictionary is large. The results are expectations over both the noise d and the dictionary itself. In addition to analyzing the MSE, we also analyze the probability of error in the subspace selection, *i.e.*, the probability that x and $\hat{x}_{\mathrm{SA}}$ lie in different subspaces.

Section 3.1 delineates the additional assumptions made in this section. The probability of error and MSE analyses are then given in Section 3.2. Estimates of the probability of error and MSE are numerically validated in Section 3.3, and finally limits as $N \to \infty$ are studied in Section 3.4

*It is possible for distinct subsets of Φ to span the same subspace.

3.1. Modeling Assumptions

This section specifies the precise modeling assumptions in analyzing denoising performance with large, isotropic random frames. Though the results are limited to the case of $K = 1$, the model is described for general K. Difficulties in extending the results to general K are described in the concluding comments of the paper. While many practical problems involve $K > 1$, the analysis of the $K = 1$ case presented here illustrates a number of unexpected qualitative phenomena, some of which we have observed for higher values of K.

The model is unchanged from earlier in the paper except that the dictionary Φ and signal x are random:

(a) *Dictionary generation:* The dictionary Φ consists of M i.i.d. random vectors uniformly distributed on the unit sphere in $\mathbb{R}^N$.

(b) *Signal generation:* The true signal x is a linear combination of the first K dictionary elements so that

$$x = \sum_{i=1}^{K} \alpha_i \varphi_i,$$

for some random coefficients $\{\alpha_i\}$. The coefficients $\{\alpha_i\}$ are independent of the dictionary except in that x is normalized to have $\|x\|^2 = N$ for all realizations of the dictionary and coefficients.

(c) *Noise:* The noisy signal y is given by $y = x + d$ where, as before, $d \sim \mathcal{N}(0, \sigma^2 I_N)$. d is independent of Φ and x. We will let

$$\gamma = 1/\sigma^2,$$

which is the input SNR because of the scaling of x.

For the special case when M and N are large and $K = 1$, we will estimate two quantities:

DEFINITION 3.1. *The* subspace selection error probability p_{err} *is defined as*

$$p_{\text{err}} = \Pr\left(T \neq j_{\text{true}}\right), \tag{6}$$

where T is the subspace selection index and j_{true} is the index of the subspace containing the true signal x, i.e.*, j_{true} is the index of the subset $\{1, 2, \ldots, K\}$.*

DEFINITION 3.2. *The* normalized expected MSE *is defined as*

$$E_{\text{MSE}} = \frac{1}{N\sigma^2} \boldsymbol{E}\left[\|x - \hat{x}_{\text{SA}}\|^2\right] = \frac{\gamma}{N} \boldsymbol{E}\left[\|x - \hat{x}_{\text{SA}}\|^2\right]. \tag{7}$$

Normalized expected MSE is the per-component MSE normalized by the per-component noise variance $\frac{1}{N}\mathbf{E}\left[\|d\|^2\right] = \sigma^2$. The term "expected MSE" emphasizes that the expectation in (7) is over not just the noise d, but also the dictionary Φ and signal x.

3.2. Analyses of Subspace Selection Error and MSE

The first result shows that the subspace selection error probability can be bounded by a double integral and approximately computed as a single integral. The integrands are simple functions of the problem parameters M, N, K and γ. While the result is only proven for the case of $K = 1$, K is left in the expressions to indicate the precise role of this parameter.

THEOREM 3.3 (9). *Consider the model described in Section 3.1. When $K = 1$ and M and N are large, the subspace selection error probability defined in (6) is bounded above by*

$$p_{\text{err}} \;<\; 1 - \int_0^\infty \int_0^\infty f_r(u) f_s(v) \exp\left(-\frac{(CG(u,v))^r}{1 - G(u,v)}\right) 1_{\{G(u,v) \leq G_{\max}\}}\, dv\, du, \tag{8}$$

and p_{err} is approximated well by

$$\begin{aligned}\widehat{p}_{\text{err}}(N,M,K,\gamma) &= 1-\int_0^\infty f_r(u)\exp\left(-\left(\frac{C(N-K)\sigma^2u}{N+(N-K)\sigma^2u}\right)^r\right)du \\ &= 1-\int_0^\infty f_r(u)\exp\left(-\left(\frac{Cau}{1+au}\right)^r\right)du, \end{aligned} \tag{9}$$

where

$$G(u,v) = \frac{au}{au+\left(1-\sigma\sqrt{\frac{Kv}{N}}\right)^2} \tag{10}$$

$$G_{\max} = (r\beta(r,s))^{1/(r-1)},$$

$$C = \left(\frac{J-1}{r\beta(r,s)}\right)^{1/r}, \quad J=\binom{M}{K} \tag{11}$$

$$r = \frac{N-K}{2}, \quad s=\frac{K}{2}, \tag{12}$$

$$a = \frac{(N-K)\sigma^2}{N} = \frac{N-K}{N\gamma}, \tag{13}$$

$f_r(u)$ is the probability distribution

$$f_r(u) = r^r\Gamma(r)u^{r-1}e^{-ru}, \quad u\in[0,\infty), \tag{14}$$

$\beta(r,s)$ is the beta function, and $\Gamma(r)$ is the Gamma function.

It is interesting to evaluate $\widehat{p}_{\text{err}}$ in two limiting cases. First, suppose that $J=1$. This corresponds to the situation where there is only one subspace. In this case, $C=0$ and (9) gives $\widehat{p}_{\text{err}}=0$. This is expected, since with one subspace there is no chance of a subspace selection error.

At the other extreme, suppose that N, K, and γ are fixed and $M\to\infty$. Then $C\to\infty$ and $\widehat{p}_{\text{err}}\to 1$. Again, this is expected since as the size of the frame increases, the number of possible subspaces increases and the probability of error increases.

The next result approximates the normalized expected MSE with a double integral. The integrand is relatively simple to evaluate and decays quickly as $\rho\to\infty$ and $u\to\infty$ so numerically approximating the double integral is not difficult.

THEOREM 3.4 (9). *Consider the model described in Section 3.1. When $K=1$ and M and N are large, the normalized expected MSE defined in (7) is given approximately by*

$$\widehat{E}_{\text{MSE}}(N,M,K,\gamma) = \frac{K}{N}+\int_0^\infty\int_0^\infty f_r(u)g_r(\rho)F(\rho,u)\,d\rho\,du, \tag{15}$$

where $f_r(u)$ is given in (14), $g_r(\rho)$ is the probability distribution

$$g_r(\rho) = rC^rr^{r-1}\exp\left(-(C\rho)^r\right), \tag{16}$$

$$F(\rho,u) = \begin{cases} \gamma(au(1-\rho)+\rho), & \text{if } \rho(1+au)<au; \\ 0, & \text{otherwise,}\end{cases} \tag{17}$$

and C, r, and a are defined in (11)–(13).

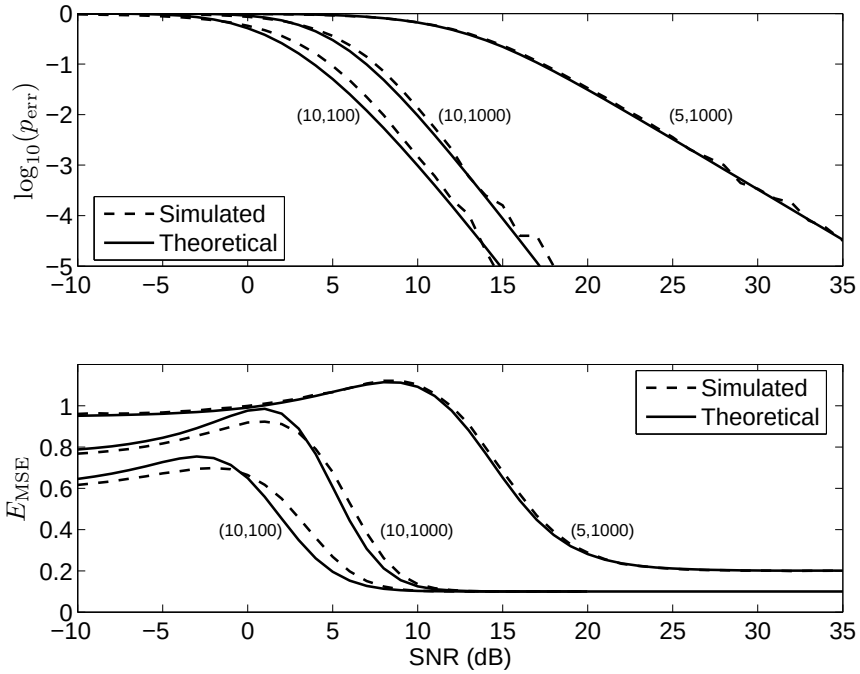


Figure 2. Simulation of subspace selection error probability and normalized expected MSE for isotropic random dictionaries. Calculations were made for integer SNRs (in dB), with 5×10^5 independent simulations per data point. In all cases $K = 1$. The curve pairs are labeled by (N, M). Simulation results are compared to the estimates from Theorems 3.3 and 3.4.

3.3. Numerical Examples

We now present simulation results to examine the accuracy of the approximations in Theorems 3.3 and 3.4. Three pairs of (N, M) values were used: (5,1000), (10,100), and (10,1000). For each integer SNR from -10 dB to 35 dB, the subspace selection and normalized MSE were measured for 5×10^5 independent experiments. The resulting empirical probabilities of subspace selection error and normalized expected MSEs are shown in Figure 2. Plotted alongside the empirical results are the estimates $\widehat{p}_{\rm err}$ and $\widehat{E}_{\rm MSE}$ from (9) and (15).

Comparing the theoretical and measured values in Figure 2, we see that the theoretical values match the simulation closely over the entire SNR range. Also note that the bottom panel of Figure 2 shows qualitatively the same behavior as Figure 1 (note that the direction of the horizontal axis is reversed). In particular, $E_{\rm MSE} \approx \frac{K}{N}$ for high SNR and the low SNR performance approaches a constant that depends on M and N.

3.4. Asymptotic Analysis

The estimates $\widehat{p}_{\rm err}$ and $\widehat{E}_{\rm MSE}$ are not difficult to compute numerically, but the expressions (9) and (15) provide little direct insight. It is thus interesting to examine the asymptotic behavior of $\widehat{p}_{\rm err}$ and $\widehat{E}_{\rm MSE}$ as N and M grow. The following theorem gives an asymptotic expression for the limiting value of the error probability function.

THEOREM 3.5 (9). *Consider the function $\widehat{p}_{\rm err}(N, M, K, \gamma)$ defined in (9). Define the* critical SNR *as a function of M, N, and K as*

$$\gamma_{\rm crit} \;=\; C - 1 \;=\; \left(\frac{J-1}{r\beta(r,s)}\right)^{1/r} - 1. \tag{18}$$

where C, r, s and J are defined in (11) and (12). For $K = 1$ and any fixed γ and $\gamma_{\rm crit}$,

$$\lim_{\substack{N,M\to\infty \\ \gamma_{\rm crit}\ constant}} \widehat{p}_{\rm err}(N, M, K, \gamma) \;=\; \begin{cases} 1, & \text{if } \gamma < \gamma_{\rm crit}; \\ 0, & \text{if } \gamma > \gamma_{\rm crit}, \end{cases} \tag{19}$$

where the limit is on any sequence of M and N with γ_{crit} constant.

The theorem shows that, asymptotically, there is a critical SNR γ_{crit} above which the error probability goes to one and below which the probability is zero. Thus, even though the frame is random, the error event asymptotically becomes deterministic.

A similar result holds for the asymptotic MSE.

THEOREM 3.6 (9). *Consider the function $\widehat{E}_{\text{MSE}}(M, N, K, \gamma)$ defined in (15) and the critical SNR γ_{crit} defined in (18). For $K = 1$ and any fixed γ and γ_{crit},*

$$\lim_{\substack{N, M \to \infty \\ \gamma_{\text{crit}} \text{ constant}}} \widehat{E}_{\text{MSE}}(M, N, K, \gamma) = \begin{cases} \widehat{E}_{\text{lim}}(\gamma), & \text{if } \gamma < \gamma_{\text{crit}}; \\ 0, & \text{if } \gamma > \gamma_{\text{crit}}, \end{cases} \tag{20}$$

where the limit is on any sequence of M and N with γ_{crit} constant, and

$$\widehat{E}_{\text{lim}}(\gamma) = \frac{\gamma + \gamma_{\text{crit}}}{1 + \gamma_{\text{crit}}}.$$

Remarks:

(i) Theorems 3.5 and 3.6 hold for any values of K. They are stated for $K = 1$ because the significance of $\widehat{p}_{\text{err}}(N, M, K, \gamma)$ and $\widehat{E}_{\text{MSE}}(M, N, K, \gamma)$ is proven only for $K = 1$.

(ii) Both Theorems 3.5 and 3.6 involve limits with γ_{crit} constant. It is useful to examine how M, N and K must be related asymptotically for this condition to hold. One can use the definition of the beta function, $\beta(r, s) = \Gamma(r)\Gamma(s)/\Gamma(r + s)$, along with Stirling's approximation, to show that when $K \ll N$,

$$(r\beta(r, s))^{1/r} \approx 1. \tag{21}$$

Substituting (21) into (18), we see that $\gamma_{\text{crit}} \approx J^{1/r} - 1$. Also, for $K \ll N$ and $K \ll M$,

$$J^{1/r} = \binom{M}{K}^{2/(N-K)} \approx (M/K)^{2K/N},$$

so that

$$\gamma_{\text{crit}} \approx (M/K)^{2K/N} - 1$$

for small K and large M and N. Therefore, for γ_{crit} to be constant, $(M/K)^{2K/N}$ must be constant. Equivalently, the dictionary size M must grow as $K(1 + \gamma_{\text{crit}})^{N/(2K)}$, which is exponential in the inverse sparsity N/K.

The asymptotic normalized MSE is plotted in Figure 3 for various values of the critical SNR γ_{crit}. When $\gamma > \gamma_{\text{crit}}$, the normalized MSE is zero. This is expected: from Theorem 3.5, when $\gamma > \gamma_{\text{crit}}$, the estimator will always pick the correct subspace. We know that for a fixed subspace estimator, the normalized MSE is K/N. Thus, as $N \to \infty$, the normalized MSE approaches zero.

What is perhaps surprising is the behavior for $\gamma < \gamma_{\text{crit}}$. In this regime, the normalized MSE actually *increases* with increasing SNR. At the critical level, $\gamma = \gamma_{\text{crit}}$, the normalized MSE approaches its maximum value

$$\max \widehat{E}_{\text{lim}} = \frac{2\gamma_{\text{crit}}}{1 + \gamma_{\text{crit}}}.$$

When $\gamma_{\text{crit}} > 1$, the limit of the normalized MSE $\widehat{E}_{\text{lim}}(\gamma)$ satisfies $\widehat{E}_{\text{lim}}(\gamma) > 1$. Consequently, the sparse approximation results in noise *amplification* instead of noise reduction. In the worst case, as $\gamma_{\text{crit}} \to \infty$, $\widehat{E}_{\text{lim}}(\gamma) \to 2$. Thus, sparse approximation can result in a noise amplification by a factor as large as 2.

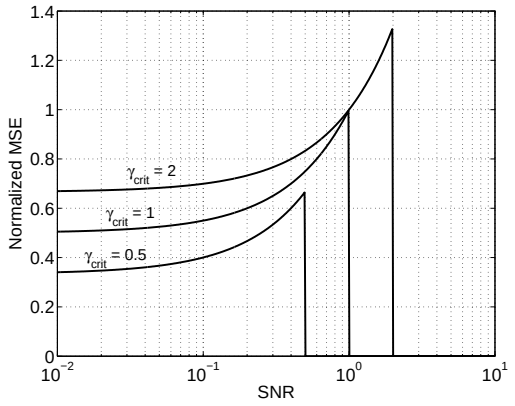


Figure 3. Asymptotic normalized MSE as $N \to \infty$ (from Theorem 3.6) for various critical SNRs γ_{crit}.

4. COMMENTS AND CONCLUSIONS

This paper has addressed properties of denoising by sparse approximation that are geometric in that the signal model is membership in a specified union of subspaces, without a probability density on that set. The denoised estimate is the feasible signal closest to the noisy observed signal.

The main results apply to the expected performance when the dictionary itself is random with i.i.d. entries selected according to an isotropic distribution. Easy-to-compute estimates for the probability that the subspace containing the true signal is not selected and for the MSE are given (Theorems 3.3 and 3.4). The accuracy of these estimates is verified through simulations. Unfortunately, these results are proven only for the case of $K = 1$. The main technical difficulty in extending these results to general K is that the distances to the various subspaces are not mutually independent. The analysis extends to larger K if the signal model and estimation procedure are changed somewhat. Instead of having M vectors and $J = \left(\begin{smallmatrix} M \\ K \end{smallmatrix} \right)$ possible subspaces (because vectors can be chosen in any combination), have J independently generated K-dimensional subspaces. By construction, this eliminates the previous technical difficulty. Furthermore, the modified model has applications in multi-user wireless communications.

Asymptotic analysis reveals a critical value of the SNR (Theorems 3.5 and 3.6). Below the critical SNR, the probability of selecting the subspace containing the true signal approaches zero and the expected MSE approaches a constant with a simple, closed form; above the critical SNR, the probability of selecting the subspace containing the true signal approaches one and the expected MSE approaches zero. The asymptotic analysis is quantitatively consistent with the information-theoretic capacity of an additive white Gaussian noise channel.

REFERENCES

1. R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series," *Trans. Amer. Math. Soc.* **72**, pp. 341–366, 1952.
2. I. Daubechies, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.
3. D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inform. Th.* **44**, pp. 2435–2476, Oct. 1998.
4. I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Proc.* **45**, pp. 600–616, Mar. 1997.
5. D. M. Malioutov, M. Çetin, and A. S. Willsky, "Sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Proc.*, to appear.
6. A. K. Fletcher and K. Ramchandran, "Estimation error bounds for frame denoising," in *Proc. Wavelets: Appl. in Sig. & Image Proc. X, part of SPIE Int. Symp. on Optical Sci. & Tech.*, **5207**, pp. 40–46, (San Diego, CA), Aug. 2003.

7. A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Denoising by sparse approximation: Error bounds based on rate–distortion theory," Electron. Res. Lab. Memo. M05/5, Univ. California, Berkeley, Sept. 2004.
8. A. K. Fletcher, "Estimation via sparse approximation: Error bounds and random frame analysis," Master's thesis, Univ. California, Berkeley, May 2005.
9. A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Denoising by sparse approximation: Error bounds based on rate–distortion theory," *EURASIP J. Appl. Sig. Proc.*, to appear 2005.
10. G. Davis, *Adaptive Nonlinear Approximations*. PhD thesis, New York Univ., Sept. 1994.
11. B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Computing* **24**, pp. 227–234, Apr. 1995.
12. S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Proc.* **41**, pp. 3397–3415, Dec. 1993.
13. S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.* **43**(1), pp. 129–159, 2001.
14. M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Inform. Th.* **48**, pp. 2558–2567, Sept. 2002.
15. R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inform. Th.* **49**, pp. 3320–3325, Dec. 2003.
16. R. Gribonval and M. Nielsen, "Highly sparse representations from dictionaries are unique and independent of the sparseness measure," Tech. Rep. R-2003-16, Dept. Mathematical Sciences, Aalborg University, Oct. 2003.
17. D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Acad. Sci.* **100**, pp. 2197–2202, Mar. 2003.
18. D. L. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inform. Th.*, submitted Feb. 2004.
19. J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Th.* **50**, pp. 2231–2242, Oct. 2004.
20. J. A. Tropp, "Just relax: Convex programming methods for subset selection and sparse approximation," ICES Report 0404, Univ. of Texas at Austin, Feb. 2004.
21. R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," Tech. Rep. 1619, IRISA, Rennes, France, Apr. 2004.
22. R. Gribonval and M. Nielsen, "Beyond sparsity: Recovering structured representations by ℓ_1 minimization and greedy algorithms—Application to the analysis of sparse underdetermined ICA," Tech. Rep. 1684, IRISA, Rennes, France, Jan. 2005.
23. N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion," in *Wavelets in Geophysics*, E. Foufoula-Georgiou and P. Kumar, eds., pp. 299–324, Academic Press, San Diego, CA, 1994.
24. B. K. Natarajan, "Filtering random noise from deterministic signals via data compression," *IEEE Trans. Signal Proc.* **43**, pp. 2595–2605, Nov. 1995.
25. H. Krim, D. Tucker, S. Mallat, and D. Donoho, "On denoising and best signal representation," *IEEE Trans. Inform. Th.* **45**, pp. 2225–2238, Nov. 1999.
26. S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Proc.* **9**, pp. 1532–1546, Sept. 2000.
27. J. Liu and P. Moulin, "Complexity-regularized image denoising," *IEEE Trans. Image Proc.* **10**, pp. 841–851, June 2001.
28. J. H. Conway, R. H. Hardin, and N. J. A. Sloane, "Packing lines, planes, etc.: Packings in Grassmannian spaces," *Experimental Mathematics* **5**(2), pp. 139–159, 1996.
29. T. Strohmer and R. W. Heath Jr., "Grassmannian frames with applications to coding and communication," *Appl. Comput. Harm. Anal.* **14**, pp. 257–275, May 2003.
30. N. J. A. Sloane, R. H. Hardin, and W. D. Smith, "A library of putatively optimal spherical codes, together with other arrangements which may not be optimal but are especially interesting for some reason.." URL: `http://www.research.att.com/~njas/packings`.