Robust Coding Over Noisy Overcomplete Channels

Eizaburo Doi, Doru C. Balcan, and Michael S. Lewicki

***Abstract*—We address the problem of robust coding in which the signal information should be preserved in spite of intrinsic noise in the representation. We present a theoretical analysis for 1- and 2-D cases and characterize the optimal linear encoder and decoder in the mean-squared error sense. Our analysis allows for an arbitrary number of coding units, thus including both under- and over-complete representations, and provides insights into optimal coding strategies. In particular, we show how the form of the code adapts to the number of coding units and to different data and noise conditions in order to achieve robustness. We also present numerical solutions of robust coding for high-dimensional image data, demonstrating that these codes are substantially more robust than other linear image coding methods such as PCA, ICA, and wavelets.**

***Index Terms*—Channel capacity constraint, channel noise, mean-squared error (MSE) bounds, overcomplete representations, robust coding.**

I. Introduction

MANY approaches to optimal coding focus on representing information with minimum entropy codes derived by approximating the underlying statistical density of the data, such as principal or independent component analysis (PCA or ICA), or by developing encoding/decoding algorithms with desirable computational and representational properties, such as Fourier and wavelet-based codes. Another important, but less commonly addressed, aspect of coding is *robustness*: How much information about the signal is retained when the representation is subject to noise, i.e., when the representation has limited precision?

Standard approaches to coding often fail tests of robustness. Although a code may achieve maximum dimensionality reduction with minimal error or may be statistically optimal in terms of minimal entropy, the representation is often assumed to be real valued and noise free, which implicitly assumes a representation whose coefficients have infinite precision. If the coefficients are subject to noise or their precision is limited, optimality of the representation cannot be guaranteed. Optimality under limited precision is a common practical concern. It would be useful if the data can be represented with small error and with low-bit precision. This issue is also relevant to biological neural representations where the coding precision of individual neurons has been reported to be as low as a few bits per spike (for a review, see [1]).

Manuscript received January 31, 2005; revised August 8, 2006. This work was supported in part by the National Science Foundation under Awards 0238351 and 04131152 and in part by the National Geospatial-Intelligence Agency under Award HM1582-04-0004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Amir Said.

E. Doi is with the Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: edoi@cnbc.cmu.edu).

D. C. Balcan is with the Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: dbalcan@cs.cmu.edu).

M. S. Lewicki is with the Center for the Neural Basis of Cognition and the Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: lewicki@cnbc.cmu.edu).

Digital Object Identifier 10.1109/TIP.2006.888352

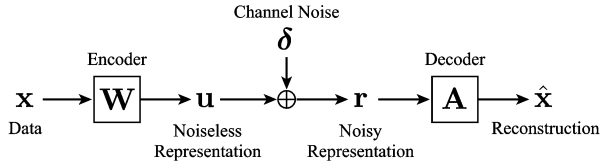


Fig. 1. Diagram of the model.

In this paper, we present a new coding scheme called *robust coding* that makes use of an arbitrary number of coding units to minimize the reconstruction error. One characteristic of robust coding is that it can *introduce* redundancy in the code in order to compensate for channel noise, unlike PCA or ICA that aim to *reduce* redundancy. Because noisy, low-precision codes can be interpreted as a representational bottleneck, the problem might appear similar to dimensionality reduction or compression. However, it is a fundamentally different problem. To take an example, if a great number of coding elements were available while their coding precision were significantly restricted, the apparent dimensionality would increase while the total representational capacity would still be limited.

This paper is organized as follows. First, in Section II, we formulate the problem. In Section III, we analyze the solutions in the general case, and then derive the optimal solutions for the 1- and 2-D cases. In Section IV, we demonstrate robustness of the proposed coding method in the context of image coding. Finally, in Section V, we summarize our results and discuss related studies.

II. Problem Formulation

To define our model, we assume that the data are N-dimensional with zero mean and covariance matrix $\mathbf{\Sigma_x}$. For each data point $\mathbf{x}$, its representation $\mathbf{r}$ in the model is the linear transform of $\mathbf{x}$ through a matrix $\mathbf{W} \in \mathbb{R}^{M\times N}$, perturbed by the additive noise (i.e., channel noise) $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \sigma_\delta^2 \mathbf{I}_M)$

$$\mathbf{r} = \mathbf{W}\mathbf{x} + \boldsymbol{\delta} = \mathbf{u} + \boldsymbol{\delta}. \tag{1}$$

We refer to $\mathbf{W}$ as the *encoding matrix* and to its rows as *encoding vectors*. The reconstruction of a data point is the linear transform of the noisy representation using matrix $\mathbf{A} \in \mathbb{R}^{N\times M}$

$$\hat{\mathbf{x}} = \mathbf{A}\mathbf{r} = \mathbf{A}\mathbf{W}\mathbf{x} + \mathbf{A}\boldsymbol{\delta}. \tag{2}$$

We refer to $\mathbf{A}$ as the *decoding matrix* and to its columns as *decoding vectors*. The term $\mathbf{AWx}$ in (2) determines how the reconstruction depends on the data, and $\mathbf{A}\boldsymbol{\delta}$ expresses the influence of the channel noise on the reconstruction. If there is no channel noise ($\boldsymbol{\delta} = \mathbf{0}$), then $\mathbf{AW} = \mathbf{I}$ is equivalent to perfect reconstruction. A graphical description of this system is shown in Fig. 1.

The goal is to form an accurate representation of the data that is robust to the presence of channel noise. More precisely, we seek an optimal pair of linear encoder and decoder, and quantify the accuracy of a representation by the mean-squared error (MSE) of the reconstruction. The error of each sample point is

$$\boldsymbol{\epsilon} = \mathbf{x} - \hat{\mathbf{x}} = (\mathbf{I}_N - \mathbf{AW})\mathbf{x} - \mathbf{A}\boldsymbol{\delta} \tag{3}$$

and the MSE $\mathcal{E} = \langle \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \rangle = \mathrm{tr}(\langle \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \rangle)$ in matrix form is

$$\mathcal{E}(\mathbf{A}, \mathbf{W}) = \mathrm{tr}\{(\mathbf{I}_N - \mathbf{AW})\boldsymbol{\Sigma}_{\mathbf{x}}(\mathbf{I}_N - \mathbf{AW})^T\} + \sigma_\delta^2 \mathrm{tr}\{\mathbf{AA}^T\}. \tag{4}$$

Note that the optimal values of $\mathbf{A}$ and $\mathbf{W}$ depend solely on second-order statistics, i.e., the covariance matrix of the data $\boldsymbol{\Sigma}_{\mathbf{x}}$ and the chanel noise variance σ_δ^2.

We are interested in the system in which the precision of the code is limited, i.e., the representation $\mathbf{r}$ has a limited signal-to-noise ratio (SNR). In order to limit the SNR, we fix the variance of each coding unit

$$\langle u_j^2 \rangle = \sigma_u^2 \tag{5}$$

which yields the SNR

$$\gamma^2 = \frac{\sigma_u^2}{\sigma_\delta^2}. \tag{6}$$

Here, we assume that the SNR is the same for each unit. As the channel capacity of information is defined by $C = (1/2)\log_2(\gamma^2 + 1)$ [2], limiting the SNR is equivalent to limiting the capacity for each unit. We will refer to this constraint as the *channel capacity constraint*.

III. Optimal Solutions and Their Characteristics

The goal is to minimize the MSE (4) subject to the channel capacity constraint (5). In this section, we first analyze the problem in the general case as far as possible, and then present the optimal solutions for 1-D and 2-D data.

First, let us consider how to incorporate the channel capacity constraint in the analysis. Equation (5) is expressed in terms of $\mathbf{W}$ as

$$\mathrm{diag}(\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{W}^T) = \sigma_u^2 \mathbf{1}_M \tag{7}$$

where $\mathbf{1}_M = (1, \ldots, 1)^T \in \mathbb{R}^M$. It takes a convenient form

$$\mathrm{diag}(\mathbf{VV}^T) = \mathbf{1}_M \tag{8}$$

when we define $\mathbf{V}$ by a linear transform of $\mathbf{W}$

$$\mathbf{V} \equiv \mathbf{WES}/\sigma_u \tag{9}$$

where $\boldsymbol{\Sigma}_{\mathbf{x}} = \mathbf{EDE}^T$ is the eigenvalue decomposition of the data covariance matrix $\mathbf{S} = \mathbf{D}^{(1/2)} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_M})$, and $\lambda_k \equiv \mathbf{D}_{kk}$ are the eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{x}}$. To summarize, the channel capacity constraint is now written in terms of $\mathbf{W}$ so that its linear transform $\mathbf{V}$ should have row vectors of unit length.

Next, let us consider a necessary condition for the minimum $\mathcal{E}$, i.e., the first derivative of $\mathcal{E}$ should be zero with respect to all free parameters. $\partial\mathcal{E}/\partial\mathbf{A} = \mathbf{O}$ yields

$$\mathbf{A} = \frac{1}{\sigma_u}\gamma^2 \mathbf{ES}(\mathbf{I}_N + \gamma^2 \mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T \tag{10}$$

(see Appendix A for the derivation).

Using the channel capacity constraint (9) and the necessary condition with respect to $\mathbf{A}$ (10), the MSE (4) can be simplified as (Appendix B)

$$\mathcal{E} = \mathrm{tr}\{\mathbf{D} \cdot (\mathbf{I}_N + \gamma^2 \mathbf{V}^T\mathbf{V})^{-1}\}. \tag{11}$$

As a result, the problem has been reformulated as finding $\mathbf{V}$ that minimizes (11) where $\mathbf{V}$ should satisfy the channel capacity constraint (8).

A. One-Dimensional Data

In the 1-D case, the encoding (1) and the decoding (2) become

$$\mathbf{r} = \mathbf{w}x + \boldsymbol{\delta} \tag{12}$$

$$\hat{x} = \mathbf{a}^T\mathbf{r} \tag{13}$$

where $\sigma_x^2 = \Sigma_{\mathbf{x}} \in \mathbb{R}$, $\mathbf{a} = \mathbf{A}^T \in \mathbb{R}^{1\times M}$, and $\mathbf{w} = \mathbf{W} \in \mathbb{R}^{M\times 1}$. From (9), $\mathbf{V} = \mathbf{v} \in \mathbb{R}^{M\times 1}$ is by definition

$$\mathbf{v} = \mathbf{w} \cdot 1 \cdot \sigma_x/\sigma_u. \tag{14}$$

Accordingly, the problem is to minimize

$$\mathcal{E} = \mathrm{tr}\left\{\sigma_x^2 \cdot (1 + \gamma^2 \mathbf{v}^T\mathbf{v})^{-1}\right\} \tag{15}$$

subject to

$$\mathrm{diag}(\mathbf{vv}^T) = \mathbf{1}_M \Leftrightarrow v_j^2 = 1. \tag{16}$$

The optimal $\mathbf{w}$ is determined solely from the channel capacity constraint (16)

$$w_j = \pm\frac{\sigma_u}{\sigma_x}. \tag{17}$$

Plugging it into (10) and (15), we obtain the optimal $\mathbf{a}$ and the minimum MSE

$$a_j = \frac{1}{w_j} \cdot \frac{\gamma^2}{M\gamma^2 + 1} \tag{18}$$

$$\mathcal{E} = \frac{\sigma_x^2}{M\gamma^2 + 1}. \tag{19}$$

This simplest case already exhibits characteristics of the more general cases. The encoding (17) is just repeating the same measurement of the data (up to sign) M times (with the appropriate scaling in order to satisfy the channel capacity constraint), while the decoding (18) depends on the SNR (γ^2) in such a way that the reconstruction becomes less representation-dependent if the SNR is small ($a_j \to 0$ as $\gamma^2 \to 0$). This dependency is counteracted by the increase of the number of units (M). Accordingly, the minimum MSE depends on the SNR and the number

of units, monotonically decreasing with respect to both, and a decrease in SNR can be compensated by an increase of the number of units to maintain the reconstruction accuracy. Another aspect of the minimum MSE is that the second term in (4), $\sigma_\delta^2 \mathrm{tr}(\mathbf{A}\mathbf{A}^T) = \|\mathbf{a}\|_2^2$ leads the optimal $\mathbf{a}$ into having as small length as possible, while the first term prevents it from being arbitrarily small; the optimum is given by the best tradeoff between them. It implies that $\mathbf{AW} = \mathbf{I}_N$ is not optimal when $\sigma_\delta^2 > 0$ [i.e., the identity in the data-dependent term in (2) is not optimal under the presence of channel noise].

B. Two-Dimensional Data

The constraint (8) implies that the row vectors of $\mathbf{V}$ should be on the unit circle. We can parametrize $\mathbf{V}$ as

$$\mathbf{V} = \begin{pmatrix} \cos\theta_1 & \sin\theta_1 \\ \vdots & \vdots \\ \cos\theta_M & \sin\theta_M \end{pmatrix} \tag{20}$$

where $\theta_j \in [0, 2\pi)$ is the angle between the jth row of $\mathbf{V}$ and the principal eigenvector of the data $\mathbf{e}_1$ ($\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2], \lambda_1 \geq \lambda_2 > 0$).

Using this parametrization, (11) is further simplified as (see Appendix C)

$$\mathcal{E} = \frac{(\lambda_1 + \lambda_2)\left(\frac{M}{2}\gamma^2 + 1\right) - \frac{\gamma^2}{2}(\lambda_1 - \lambda_2)\mathrm{Re}(Z)}{\left(\frac{M}{2}\gamma^2 + 1\right)^2 - \frac{1}{4}\gamma^4|Z|^2} \tag{21}$$

where, by definition

$$Z = \sum_{j=1}^{M} z_j = \sum_{j=1}^{M} (\cos 2\theta_j + i \sin 2\theta_j). \tag{22}$$

Now, the problem is reduced to finding a complex number Z that minimizes $\mathcal{E}$ [note that once Z is determined, we can obtain feasible $\{\theta_j\}_{j=1}^M$ in $\mathbf{V}$, which, in turn, determines $\mathbf{W}$ and $\mathbf{A}$ by (9) and (10)].

In the following, we analyze the problem in two cases: when the data variance is isotropic (i.e., $\lambda_1 = \lambda_2$ in terms of the data variance along the principal axes), and when it is anisotropic ($\lambda_1 > \lambda_2$). As we will see, the solutions are qualitatively different in these two cases.

C. Isotropic Case

Due to the isotropy of the data variance, $\lambda_1 = \lambda_2 \equiv \sigma_x^2$, and without loss of generality, $\mathbf{E} = \mathbf{I}$. These simplify (21) as

$$\mathcal{E} = \frac{2\sigma_x^2\left(\frac{M}{2}\gamma^2 + 1\right)}{\left(\frac{M}{2}\gamma^2 + 1\right)^2 - \frac{1}{4}\gamma^4|Z|^2}. \tag{23}$$

In this case, $\mathcal{E}$ is minimized whenever $|Z|^2$ is minimized.

1) If $M = 1$: By definition (22), $|Z|^2 = |z_1|^2 = 1$, yielding the optimal solutions as

$$\mathbf{W} = \frac{\sigma_u}{\sigma_x}\mathbf{V} \tag{24}$$

$$\mathbf{A} = \frac{\sigma_x}{\sigma_u} \cdot \frac{\gamma^2}{\gamma^2 + 1}\mathbf{V}^T \tag{25}$$

where $\mathbf{V} = \mathbf{V}(\theta_1), \forall \theta_1 \in [0, 2\pi)$. Equations (24) and (25) mean that the orientation of the optimal encoding and decoding vectors is arbitrary (Fig. 2, $M = 1$) and that the length of those vectors is adjusted exactly as in the 1-D case [(17) and (18) with $M = 1$]. Accordingly, the minimum MSE is given by

$$\mathcal{E} = \frac{\sigma_x^2}{\gamma^2 + 1} + \sigma_x^2. \tag{26}$$

The first term takes the same form as in the 1-D case [(18) with $M = 1$], corresponding to the error component along the axis that the encoding/decoding vectors represent, while the second term is the whole data variance along its orthogonal direction (along which no reconstruction is made), as depicted in Fig. 2 with $M = 1$.

2) If $M \geq 2$: There always exists a set of angles θ_j for which $|Z|^2$ is 0. This can be verified by representing Z in the complex plane (the Z diagram in Fig. 2): There is always a configuration of connected, unit-length bars that starts from, and ends up at the origin, indicating $Z = |Z|^2 = 0$. Accordingly, the optimal solution is

$$\mathbf{W} = \frac{\sigma_u}{\sigma_x}\mathbf{V} \tag{27}$$

$$\mathbf{A} = \frac{\sigma_x}{\sigma_u} \cdot \frac{\gamma^2}{\frac{M}{2}\gamma^2 + 1}\mathbf{V}^T \tag{28}$$

where the optimal $\mathbf{V} = \mathbf{V}(\theta_1, \ldots, \theta_M)$ is given by such $\{\theta_j\}_{j=1}^M$ for which $Z = 0$. Namely, as illustrated in Fig. 2, if $M = 2$, then z_1 and z_2 must be antiparallel but are not otherwise constrained, making the two decoding vectors (and also the two encoding vectors) orthogonal, yet free to rotate.[1] Likewise, if $M = 3$, the decoding vectors should be evenly distributed yet still free to rotate, due to the equilateral triangle of Z configuration. If $M = 4$, the four vectors should be two pairs of orthogonal vectors since the Z configuration should be a rhomboid, consisting of two pairs of antiparallel bars. If $M \geq 5$, there is no obvious regularity anymore. In all cases, the norm of the decoding vectors gets smaller by increasing M or decreasing γ^2, while the norm of the encoding vectors are always constant [i.e., σ_u/σ_x from (27)] as observed in Fig. 2. With $Z = 0$, (23) is minimized as

$$\mathcal{E} = \frac{2\sigma_x^2}{\frac{M}{2}\gamma^2 + 1}. \tag{29}$$

It takes the same form as in the 1-D case (18) considering that in both cases the numerator is the data variance, $\mathrm{tr}(\mathbf{\Sigma_x})$, and that the factor of γ^2 is the overcompleteness ratio, M/N.

D. Anisotropic Case

Considering the anisotropic condition ($\lambda_1 > \lambda_2$), (21) is minimized when $Z = \mathrm{Re}(Z) \geq 0$ for a fixed value of $|Z|^2$. There-

[1]Note that both the encoding (27) and the decoding (28) vectors are parallel to the rows of $\mathbf{V}$. Also, from (20) and (22), the angle of z_j from the real axis is twice as large as that of $\mathbf{a_j}$ and of $\mathbf{w_j}$.

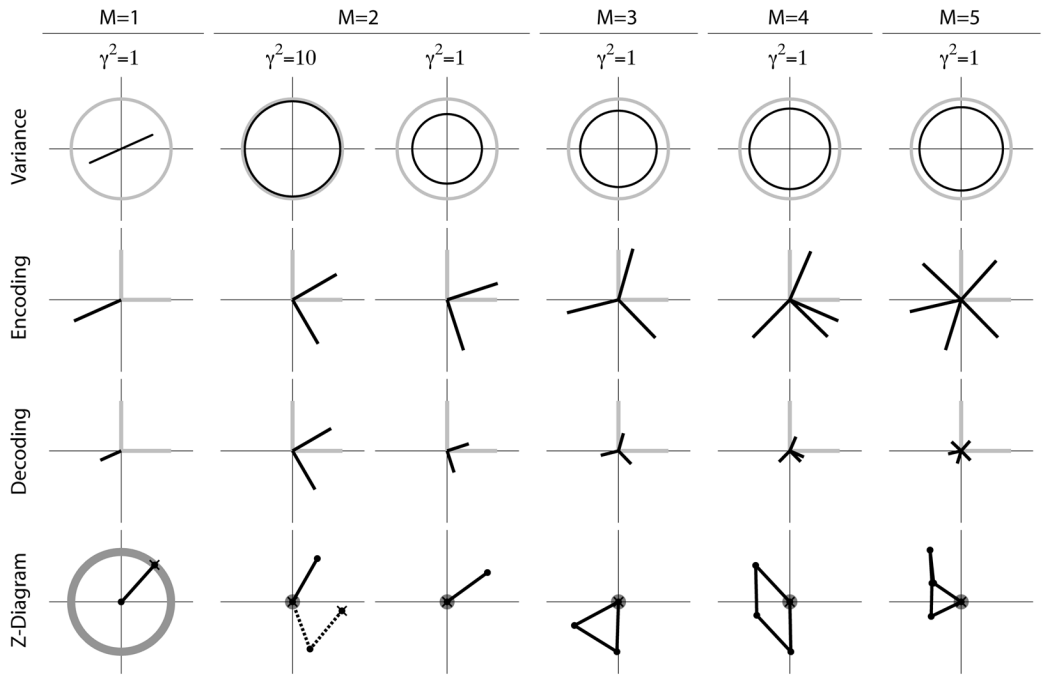


Fig. 2. Optimal solutions for isotropic data. M is the number of units and γ^2 is the SNR in the representation. "Variance" shows the variance ellipses for the data (gray) and the reconstruction (solid). For perfect reconstruction, the two ellipses should overlap. "Encoding" and "Decoding" show encoding vectors and decoding vectors (solid bars), respectively. The gray bars show the principal axes of the data, $\mathbf{e}_1$ and $\mathbf{e}_2$. "Z diagram" represents $Z = \Sigma_{j=1}^{M} z_j$ (22) in the complex plane, where each unit length bar corresponds to a z_j, and the end point indicated by "$\times$" represents the coordinates of Z. The set of dark-gray dots in a plot corresponds to optimal values of Z; when this set reduces to a single dot, the optimal Z is unique. In general, there could be multiple configurations of bars for a single Z, implying multiple equivalent solutions of $\mathbf{V}$ (and, therefore, those of $\mathbf{A}$ and $\mathbf{W}$). At $M = 2$ and $\gamma^2 = 10$, we show with dotted bars an example of Z that is not optimal (corresponding encoding and decoding vectors not shown).

fore, the problem is reduced to seeking a real value $Z = y \in [0, M]$ that minimizes

$$\mathcal{E} = \frac{(\lambda_1 + \lambda_2)\left(\frac{M}{2}\gamma^2 + 1\right) - \frac{\gamma^2}{2}(\lambda_1 - \lambda_2)y}{\left(\frac{M}{2}\gamma^2 + 1\right)^2 - \frac{1}{4}\gamma^4 y^2}. \tag{30}$$

1) If $M = 1$: In this case, from (22), $Z = \mathrm{Re}(Z) \geq 0$ iff $\theta_1 = 0$, which determines the optimal solutions

$$\mathbf{W} = t\frac{\sigma_u}{\sqrt{\lambda_1}}\mathbf{e}_1^T, \tag{31}$$

$$\mathbf{A} = t\frac{\sqrt{\lambda_1}}{\sigma_u} \cdot \frac{\gamma^2}{\gamma^2 + 1}\mathbf{e}_1 \tag{32}$$

where $t = \pm 1$. As illustrated in Fig. 3, with $M = 1$, the encoding and decoding vectors are fixed along the first principal axis $(\mathbf{e}_1)$, which contrasts to the isotropic case where the angle is arbitrary [Fig. 2, $M = 1$]. Accordingly, the minimum MSE is

$$\mathcal{E} = \frac{\lambda_1}{\gamma^2 + 1} + \lambda_2. \tag{33}$$

This has the same form as in the isotropic case (26) except that the first term is now specified to the variance along the first principal axis, λ_1, by which the encoding/decoding vectors can most effectively be exploited for representing the data, while the second term is specified as the data variance along the second principal (or minor) axis, λ_2, by which the total misreconstruction is mostly minimized.

2) If $M \geq 2$: We can derive the optimal y from the necessary condition for the minimum, $d\mathcal{E}/dy = 0$, yielding

$$\left[\frac{\sqrt{\lambda_1} - \sqrt{\lambda_2}}{\sqrt{\lambda_1} + \sqrt{\lambda_2}}\left(M + \frac{2}{\gamma^2}\right) - y\right] \times \left[\frac{\sqrt{\lambda_1} + \sqrt{\lambda_2}}{\sqrt{\lambda_1} - \sqrt{\lambda_2}}\left(M + \frac{2}{\gamma^2}\right) - y\right] = 0. \tag{34}$$

The existence of a root y in the domain $[0, M]$ depends on how γ^2 compares to the following quantity:

$$\gamma_c^2 = \frac{\sqrt{\lambda_1/\lambda_2} - 1}{M} \tag{35}$$

which we shall call the *critical point*.

If $\gamma^2 \geq \gamma_c^2$, then (34) has a root within $[0, M]$

$$y = \frac{\sqrt{\lambda_1} - \sqrt{\lambda_2}}{\sqrt{\lambda_1} + \sqrt{\lambda_2}}\left(\frac{2}{\gamma^2} + M\right) \tag{36}$$

with $y = M$ if $\gamma^2 = \gamma_c^2$. Accordingly, the optimal solution is given by

$$\mathbf{W} = \mathbf{V}\begin{pmatrix} \sigma_u/\sqrt{\lambda_1} & 0 \\ 0 & \sigma_u/\sqrt{\lambda_2} \end{pmatrix}\mathbf{E}^T \tag{37}$$

$$\mathbf{A} = \frac{\sqrt{\lambda_1} + \sqrt{\lambda_2}}{2\sigma_u} \cdot \frac{\gamma^2}{\frac{M}{2}\gamma^2 + 1}\mathbf{E}\mathbf{V}^T \tag{38}$$

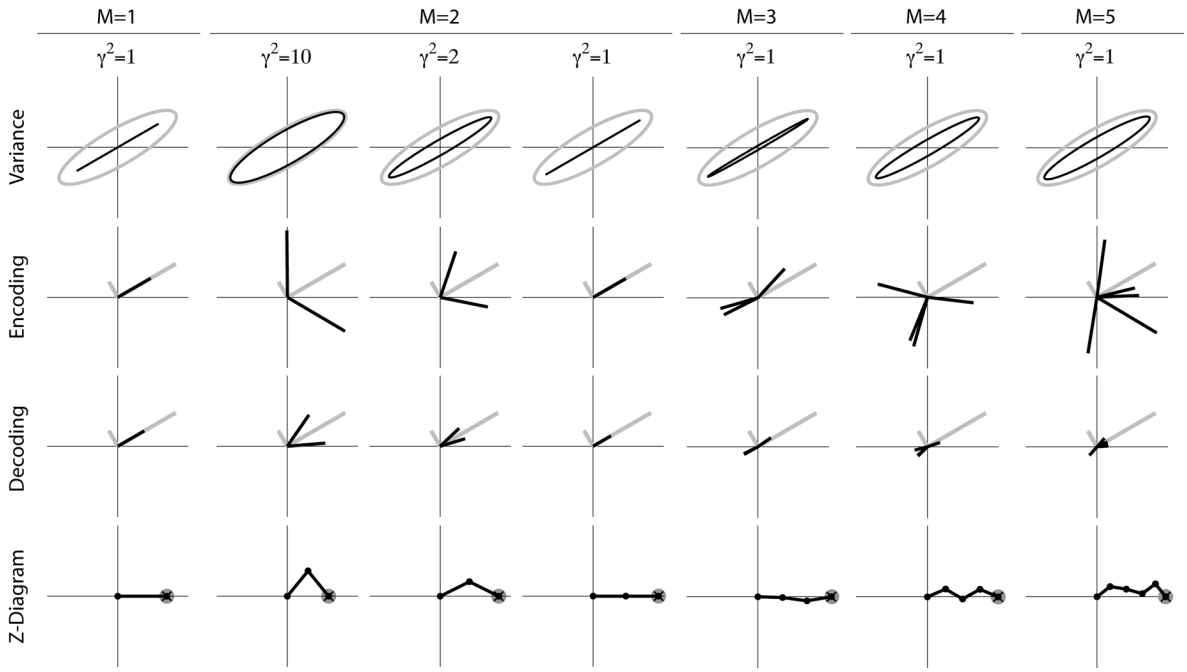


Fig. 3. Optimal solutions for anisotropic data. Notations are as in Fig. 2. We arbitrarily set $\lambda_1 = 1.87$ and $\lambda_2 = 0.13$. $\gamma^2 > \gamma_c^2$ holds for all $M \geq 2$ but the one with $M = 2$ and $\gamma^2 = 1$.

where the optimal $\mathbf{V} = \mathbf{V}(\theta_1, \ldots, \theta_M)$ is characterized by the Z diagram as illustrated in Fig. 3, which we will describe shortly.[2] Accordingly, the minimum MSE is

$$\mathcal{E} = \frac{1}{\frac{M}{2}\gamma^2 + 1} \frac{(\sqrt{\lambda_1} + \sqrt{\lambda_2})^2}{2}. \tag{39}$$

Note that (37)–(39) are reduced to (27)–(29) if $\lambda_1 = \lambda_2$.

If $\gamma^2 < \gamma_c^2$, then $d\mathcal{E}/dy = 0$ does not have a root within the domain. However, $d\mathcal{E}/dy$ is always negative, and hence $\mathcal{E}$ decreases monotonically in $[0, M]$. Therefore, the minimum is obtained for $y = M$, yielding the optimal solution

$$\mathbf{W} = \frac{\sigma_u}{\sqrt{\lambda_1}} \mathbf{t} \mathbf{e}_1^T \tag{40}$$

$$\mathbf{A} = \frac{\sqrt{\lambda_1}}{\sigma_u} \cdot \frac{\gamma^2}{M\gamma^2 + 1} \mathbf{e}_1 \mathbf{t}^T \tag{41}$$

where $\mathbf{t} \in \mathbb{R}^M, t_k = \pm 1$, and the minimum MSE is given by

$$\mathcal{E} = \frac{\lambda_1}{M\gamma^2 + 1} + \lambda_2. \tag{42}$$

This shares the same form as in $M = 1$ (33) except that we can now decrease the error by increasing the number of units. It implies that the best strategy is to devote all the representational resources solely along the first principal axis if the representational power is too limited, either by M (the number of coding units) or γ^2 (the SNR), so that $\gamma^2 \leq \gamma_c^2$.

[2]One can see that the optimal encoding and decoding vectors are restricted on an ellipse and a circle, respectively. The decoding vectors are given by rotation of $\mathbf{V}$ using $\mathbf{E}$ followed by the uniform scaling (38). The encoding vectors are on the ellipse whose principal axes are the eigenvectors of the data with the scaling that flatten along the first data principal axis (37).

Let us describe the characteristics of the optimal solutions using the Z diagram (Fig. 3). First, the solution depends on the SNR relative to the critical point. If $\gamma^2 \geq \gamma_c^2$, the optimal Z corresponds to a certain point between 0 and M on the real axis. Specifically, for $M = 2$ the optimal configuration of the bars is unique (up to flipping about the real axis), meaning that the encoding/decoding vectors are symmetric about the first principal axis; for $M \geq 3$, on the other hand, there are infinitely many configurations of unit-length connected bars starting from the origin and ending at the optimal Z, and nothing can be added about their regularity. If $\gamma^2 \leq \gamma_c^2$, the optimal Z is M, and, therefore, all the bars must line up along the real axis (recall each bar has unit length). In this case, encoding/decoding vectors are all parallel to the principal axis $(\mathbf{e}_1)$, as described by (40) and (41). Such a degenerate code is characteristic of the anisotropic case.

Second, the optimal solutions for the overcomplete representation are not trivial in the sense that they are not a replication of the optimal code for the lower number of units. For instance, under $\gamma^2 = 1$ in Fig. 3, the optimal solution for $M = 4$ is not identical to the replication of the optimal solution for $M = 2$. We can prove it in the general case using (36): The optimal y for $M = m_1 + m_2$ is not equal to the sum of ys for $M = m_1$ and for $M = m_2$, implying that combining two optimal solutions for $M = m_1$ and $M = m_2$ will not be optimal for $M = m_1 + m_2$.

Finally, robust coding represents the major (first principal component) axis more accurately than the minor axis. This is obvious for the degenerate case (e.g., Fig. 3, $M = 2, \gamma^2 = 1$), where, as in $M = 1$, the optimal strategy is to preserve information along the major axis at the cost of losing all information along the minor axis. Such a bias exists for nondegenerate code as well, where the data along the major axis is more accurately reconstructed than that along the minor axis. More precisely,

TABLE I
MINIMUM MEAN-SQUARED ERROR

	$M=1$	$M\geq 2$	
1-D	$\frac{\sigma_x^2}{\gamma^2+1}$	$\frac{\sigma_x^2}{M\cdot\gamma^2+1}$	
2-D Isotropic	$\frac{\sigma_x^2}{\gamma^2+1}+\sigma_x^2$	$\frac{2\sigma_x^2}{\frac{M}{2}\cdot\gamma^2+1}$	
2-D Anisotropic	$\frac{\lambda_1}{\gamma^2+1}+\lambda_2$	$\frac{\lambda_1}{M\cdot\gamma^2+1}+\lambda_2$	if $\gamma^2\leq\gamma_c^2$
		$\frac{1}{\frac{M}{2}\cdot\gamma^2+1}\frac{(\sqrt{\lambda_1}+\sqrt{\lambda_2})^2}{2}$	if $\gamma^2\geq\gamma_c^2$

the error along $\mathbf{e}_1$ and along $\mathbf{e}_2$ with respect to the data variance has the ratio $\sqrt{\lambda_2}:\sqrt{\lambda_1}$ (note the switch of the subscripts), implying that the percentage of the error is smaller for the major axis (see Appendix D for the derivation). This is illustrated in Fig. 3, where the reconstruction ellipse is thinner than the data ellipse; if there were no bias, the reconstruction ellipse should be similar to the data ellipse. The biased reconstruction also implies that the variance of reconstruction error should be proportional to the standard deviation of the data for the optimal solution, not proportional to the data variance nor a constant irrespective of the frequencies (i.e., white noise), as is often assumed in the literature [3]–[5].

E. Summary of the Minimum MSE

We summarize the formulas of the minimum MSE in Table I. They define the error bound that the linear encoder and decoder can achieve under the presence of channel noise. There are similarities between them, as we have emphasized throughout this section. For $M=1$, the form of 1-D solution appears in the 2-D case, and for the anisotropic case the first term is fixed for the large eigenvalues in order to minimize the MSE. For $M\geq 2$, the increase of M reduces the MSE by virtually increasing the SNR γ^2. Specifically, for 1-D data the solution shares the same form as in $M=1$. For the 2-D isotropic case, the solution is not the same as the $M=1$ solution but shares the same form as in 1-D solution, considering that the numerator is the data variance and the factor of γ^2 is the overcompleteness ratio M/N. For the 2-D anisotropic case, the degenerate solution has the same form as in $M=1$; the nondegenerate solution does not have the same form as in $M=1$, but it is reduced to the solution for the isotropic data with $M\geq 2$ if $\lambda_1=\lambda_2$.

Based on these observations, we also conjectured that the minimum MSE for N-D data should be

$$\mathcal{E}=\frac{1}{\frac{M}{N}\cdot\gamma^2+1}\frac{\left(\sum_{k=1}^{N}\sqrt{\lambda_k}\right)^2}{N} \tag{43}$$

where we assume that the code is nondegenerate but the data can be either isotropic or anisotropic.

IV. APPLICATION TO IMAGE CODING

In this section, we examine robustness of the proposed model to channel noise in the application to image coding. Since we have not had analytic characterization in the high-dimensional case, we numerically derive the optimal code. Note that robust coding results for 2-D data can be interpreted in the context of image coding by translating the first and second principal axis into lower and higher spatial-frequency dimensions, considering the general tendency of $1/f$ amplitude spectra of natural images [6]. For instance, it is shown that robust coding for image data will preserve the lower spatial-frequency components more accurately than the higher spatial-frequency components (for more details, see [7]).

We only need to derive the optimal $\mathbf{W}$ because $\mathbf{A}$ can be determined through (10) once $\mathbf{W}$ (or equivalently $\mathbf{V}$) is given. $\mathbf{W}$ should minimize $\mathcal{E}$ and satisfy the channel capacity constraint (7). This is a constrained optimization problem [8] and the optimal $\mathbf{W}$ can be derived by minimizing

$$\mathcal{C}=\mathcal{E}+\kappa\sum_{j=1}^{M}\ln\frac{\langle u_j^2\rangle}{\sigma_u^2} \tag{44}$$

where $\kappa>0$ is the so-called penalty parameter that controls the influence of the second term. This penalizes such $\mathbf{W}$ that does not satisfy the channel capacity constraint [7] (the second term becomes 0 if all the coefficients' variances are equal to the target value, σ_u^2; in this study, κ is selected so that the largest deviation of the actual variance from the target variance should be within 0.5%). The update rule of $\mathbf{W}$ is given by the gradient descent of (44)

$$\Delta\mathbf{W}\propto\mathbf{A}^T(\mathbf{I}_N-\mathbf{AW})\mathbf{\Sigma_x}-\kappa\mathrm{diag}\left(\frac{\ln\left[\langle\mathbf{u}^2\rangle/\sigma_u^2\right]}{\langle\mathbf{u}^2\rangle}\right)\mathbf{W}\langle\mathbf{x}\mathbf{x}^T\rangle. \tag{45}$$

The data consists of 8×8 pixel image patches (therefore, $N=64$), and we use 65 536 samples as a training set. These are randomly sampled from the 512×512 pixel test image. For comparison, we also examine the robustness of traditional linear image coding methods under the same channel capacity condition, namely, we set the capacity for all coding methods as 1 bit by adding Gaussian noise to the representation (Fig. 1).

First, let us demonstrate the limitation of PCA when it is applied under noisy conditions. PCA represents the data according to the data variance, and, hence, it is the most effective linear method for dimensionality reduction. In Fig. 4(a), we show the original image data, and, in Fig. 4(b), its reconstruction using PCA with 32 coding units, which utilizes only a half of the

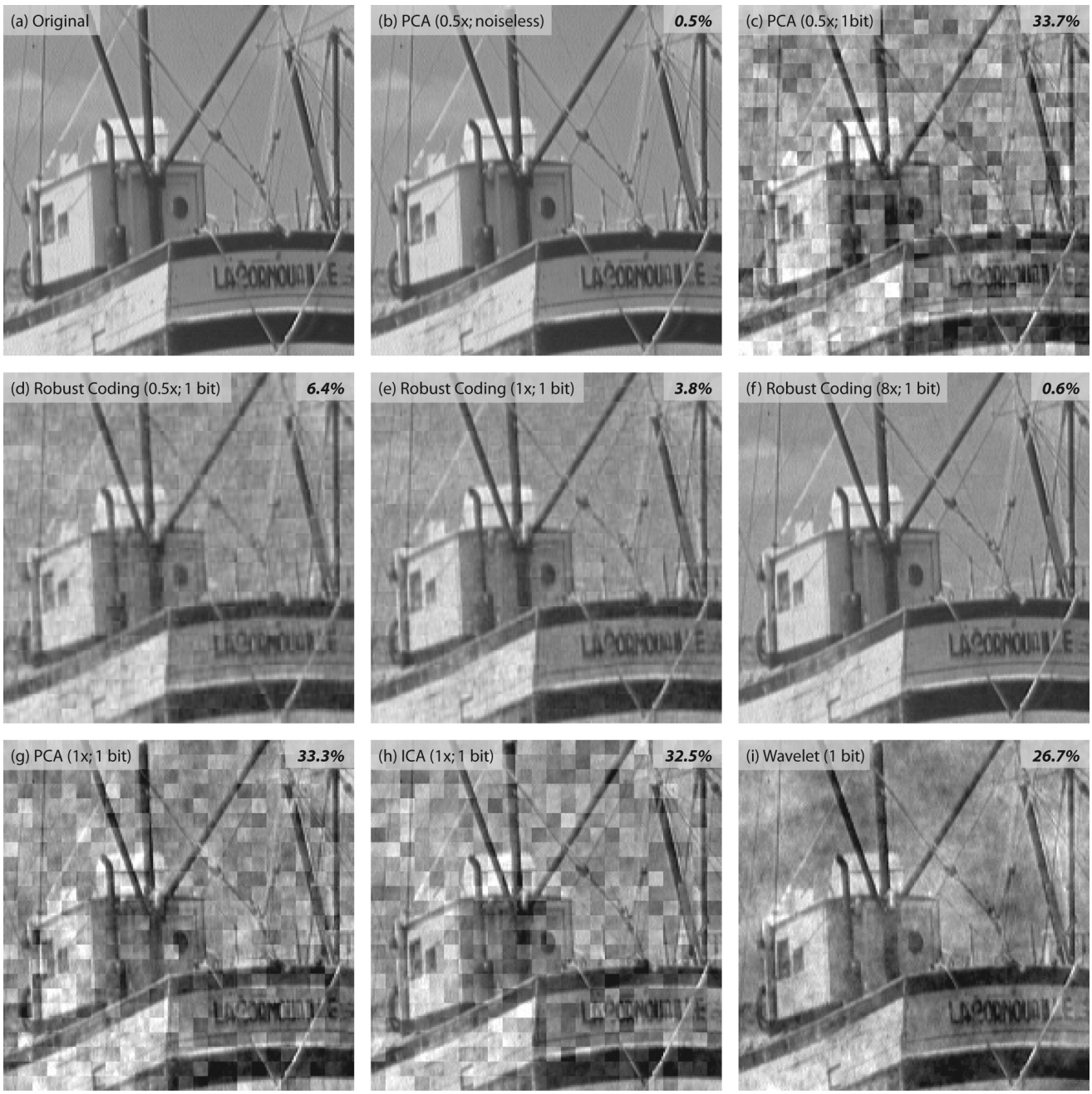


Fig. 4. Image coding under the presence of channel noise. For each reconstruction its percent error is indicated. (a) Original image. (b) PCA ($M = 32$) with noiseless representation. (c) PCA ($M = 32$) with 1-bit precision code. (d) Robust coding ($M = 32$) with 1-bit precision code. (e) Robust coding ($M = 64$) with 1-bit precision code. (f) Robust coding ($M = 512$) with 1-bit precision code. (g) PCA ($M = 64$) with 1-bit precision code. (h) ICA ($M = 64$) with 1-bit precision code. (i) Daubechies 9/7 wavelet with 1-bit precision code. The reconstruction errors in SNR [dB] are [from (b) to (i)]: 22.96, 4.75, 11.97, 14.17, 22.17, 4.79, 4.79, and 5.64. PSNR [dB] can be obtained by SNR $- 3.34$.

input dimension (its overcompleteness ratio is depicted by $0.5\times$). Its reconstruction error is small (0.5%),[3] and it is hard to see a difference from the original. However, when the code is noisy, the reconstruction becomes significantly poor (c; 33.7% error). This is somewhat obvious because PCA is not designed for the presence of channel noise.

The proposed robust coding is optimized for the noisy representation, and it can reduce the error considerably using the same number of the same noisy units [Fig. 4(d), 6.4% error]. Moreover, robust coding can further reduce the error by utilizing a greater number of coding units. Fig. 4(e) and (f) shows the results using 64 $(1\times)$ and 512 $(8\times)$ coding units, reducing the errors to 3.8% and 0.6%, respectively.[4]

For comparison, we also examined the reconstructions using all PCA coefficients in Fig. 4(g), using ICA in Fig. 4(h)—the most efficient representation in terms of coding cost, and using

[3]The percent error is defined by $(\mathrm{MSE})/(\mathrm{data\ variance}) \times 100$, indicating the unexplained portion of the data variance in the reconstruction. Using the notation in this paper, it is given by $\mathcal{E}/\mathrm{tr}(\Sigma_{\mathbf{x}}) \times 100$.

[4]The errors predicted by the conjecture (43) were consistent to the errors of the numerical solutions, i.e., 6.1%, 3.8%, and 0.6% for $0.5\times$, $1\times$, and $8\times$, respectively.

the "Daubechies 9/7" wavelet in Fig. 4(i)—one of the most widely used image codes.[5] All of these traditional linear coding methods yielded a large amount of the error, confirming a significant robustness of the proposed robust coding.

Finally, we examined some intuitive methods to construct a robust code. One reasonable method would be to allocate the limited representation resources according to the data variance. This can be implemented by repeating in the encoding and decoding matrices the eigenvector of the data covariance matrix so that the repetition is proportional to the associated eigenvalue. Such a code, with 1-bit representation and 64 coding units $(1\times)$, yielded 8.1% error, which is better than the traditional linear codes such as PCA (33.3%) but not as good as robust coding under the same conditions (3.8%). The proposed method outperforms the replication method because it provides a rigorous way to reach the error bound. Another approach for compensating for channel noise is to introduce redundancy by overcomplete representations. For example, simply replicating an existing $1\times$ code should be able to reduce the error. To compare with $8\times$ robust coding, we examined $8\times$ replications of PCA and $8\times$ replications of an optimal $1\times$ robust code. The resulting errors are 4.2% and 1.3%, respectively, demonstrating that the optimal $8\times$ robust code (which yielded 0.6% error) can reduce the error more than what is possible simply by replicating an existing code.

V. Discussion

A. Mean-Squared Error Versus Mutual Information

In this study, we measured the accuracy of the representation by the MSE of the reconstruction, which is one of the most commonly used measures (e.g., [3] and [4]). Alternatively, we can use as an accuracy measure the mutual information between the data and its reconstruction $I(\mathbf{x}, \hat{\mathbf{x}})$ and try to maximize it, which is known as the *infomax* principle [5], [9]. The optimal solutions for the infomax differs from those for the MSE objective. For example, assuming that the data are Gaussian and the representation is complete (i.e., $M = N$), the mutual information is upper bounded (Appendix E)

$$I(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{2} \ln \det(\gamma^2 \mathbf{V}\mathbf{V}^T + \mathbf{I}_N) \leq \frac{N}{2} \ln(\gamma^2 + 1) \quad (46)$$

with equality iff $\mathbf{V}\mathbf{V}^T = \mathbf{I}$. Because $\mathbf{V}\mathbf{V}^T$ is the correlation matrix of the representation $\mathbf{u}$ [note $\mathbf{V}\mathbf{V}^T = \mathbf{W}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{W}^T/\sigma_u^2$ from (7)], the optimal coding strategy for the infomax is to uncorrelate (more precisely, to whiten) the data, irrespective of the data distribution being isotropic or anisotropic.

Such a decorrelation strategy is an important difference to robust coding. Decorrelation allows us to send the maximum amount of information through channels with a limited information capacity, in which the coefficients are uncorrelated. To minimize the MSE, however, correlation among coefficients could be advantageous in order to compensate for channel noise, even if there is no redundancy in terms of the number of coding units (i.e., a complete representation). The proposed robust coding provides a rigorous way to achieve this goal. Considering that ICA is one form of whitening, the reconstruction results in Fig. 4 can be seen as demonstrating the suboptimality of the infomax solution in the MSE sense.

[5]The wavelet transform is applied to the whole image instead of image patches, and the ratio of the number of coefficients to the number of pixels is about 1.06, therefore, it is approximately $1\times$ representation. In order to compare a wavelet code under strictly the same conditions to the other codes (i.e., using 8×8 pixel blocks and $1\times$ number of units), we also examined a three-level Haar wavelet which yielded 33.2% error.

B. Relations to Previous Studies

The optimal MSE code over noisy channels was examined previously in [10] and [11]. However, there are two important differences between these and our analysis. One is that in the previous works the capacity constraint was defined for a population, not for each coding unit as in our study. The other is that their analysis is limited to the undercomplete case (i.e., $M < N$), while the approach described here can have an arbitrary number of units, which allows for the arbitrary improvement of robustness—and the arbitrary reduction of error—even for highly noisy units.

The so-called *frame expansion* is a linear encoding and decoding method that also employs overcomplete representations to compensate for channel noise [12], [13]. One significant difference from our approach is that the frame expansion is defined in a data independent manner; in other words, it is not adaptive to the data covariance matrix. Consequently, robust coding outperforms the frame expansions when the MSE is evaluated over the joint probability of channel noise and the data.

For example, the MSEs for the 2-D isotropic case are

$$\mathcal{E} = 2\sigma_x^2\sigma_\delta^2 \cdot \frac{1}{\frac{M}{2}\sigma_x^2 + \sigma_\delta^2} \quad (47)$$

for robust coding and

$$\mathcal{E}^{\text{frame}} = 2\sigma_x^2\sigma_\delta^2 \cdot \frac{1}{\frac{M}{2}\sigma_x^2} \quad (48)$$

for the frame expansion with a uniform tight frame [13], where we set $\sigma_u = \sigma_x$ for a fair comparison between the two methods. These show that the MSE is smaller by using robust coding if there is channel noise $(\sigma_\delta^2 > 0)$.

Frame expansions and robust coding are highly related. For example, the optimal encoding vectors of robust coding for 2-D isotropic data forms a tight frame, and our characterization of such optimal representations turned out to be a rediscovery of that for the tight frame [13]. However, the optimal robust encoders for 2-D anisotropic data are not tight frames, and, furthermore, they are not even frames in the degenerate case. Also, robust coding may be resilient to an erasure of some representation components even if it is not explicitly optimized for this type of noise. In the 2-D isotropic case (and, again, with $\sigma_u = \sigma_x$), an erasure of one representation component changes the MSE as

$$\mathcal{E}_1 = \frac{\sigma_\delta^2/\sigma_x^2 + M - 1}{\sigma_\delta^2/\sigma_x^2 + M - 2} \cdot \mathcal{E} \quad (49)$$

which gives the smaller effect of the erasure than that in the frame expansions with the tight frame

$$\mathcal{E}_1^{\text{frame}} = \frac{M-1}{M-2} \cdot \mathcal{E}^{\text{frame}} \quad (50)$$

where $\mathcal{E}_1$ and $\mathcal{E}_1^{\text{frame}}$ are the MSEs with one component erasure, respectively [13].

C. Concluding Remarks

In this paper, we have proposed and characterized *robust coding*: The optimal linear encoder and decoder in the MSE sense subject to the presence of channel noise. Our results are summarized as follows: we derived the error bound for the linear encoder and decoder subject to channel noise, and described the optimal configurations of linear encoder and decoder for 1- and 2-D data—particularly, we demonstrated that a completely redundant degenerate code can be optimal. The proposed coding scheme allows for a large number of coding units to arbitrarily reduce the residual error, and it is carried out by more than just replicating some code so as to cancel channel noise. Namely, the robustness is achieved by more accurately representing the major data axis than the other, by the optimal scaling of decoding vectors, and by employing overcomplete representations. Finally, numerical solutions of robust coding for high-dimensional image data significantly outperform traditional linear image coding methods such as PCA, ICA, and wavelets when channel noise exists.

APPENDIX

Here, we derive some of the formulas given in the main text.

A. Expression of the Decoding Matrix $\mathbf{A}$ in the General Case (10)

From (4)

$$\frac{\partial \mathcal{E}}{\partial \mathbf{A}} = \frac{\partial}{\partial A}\Big\{\mathrm{tr}(\mathbf{\Sigma_x}) - \mathbf{2}\cdot\mathrm{tr}(\mathbf{AW\Sigma_x}) + \mathrm{tr}(\mathbf{AW\Sigma_x W}^T\mathbf{A}^T) + \sigma_\delta^2\mathrm{tr}(\mathbf{AA}^T)\Big\} \quad (51)$$

$$= \mathbf{-2\Sigma_x W^T + 2AW\Sigma_x W^T + 2}\sigma_\delta^2\mathbf{A}. \quad (52)$$

Then, $\partial\mathcal{E}/\partial A = \mathbf{O}$ yields

$$\mathbf{A}\left(\sigma_\delta^2\mathbf{I}_M + \mathbf{W\Sigma_x W}^T\right) = \mathbf{\Sigma_x W}^T. \quad (53)$$

By replacing $\mathbf{\Sigma_x}$ with $\mathbf{ES}^2\mathbf{E}^T$ and $\mathbf{W}$ with $\sigma_u\mathbf{VS}^{-1}\mathbf{E}^T$ (9)

$$\mathbf{A} = \mathbf{\Sigma_x W}^T\left(\sigma_\delta^2\mathbf{I}_M + \mathbf{W\Sigma_x W}^T\right)^{-1} \quad (54)$$

$$= \sigma_u\mathbf{ESV}^T\left(\sigma_\delta^2\mathbf{I}_M + \sigma_u^2\mathbf{VV}^T\right)^{-1}. \quad (55)$$

This involves inverting an $M \times M$ matrix (it exists when $\sigma_\delta^2 > 0$), which could be computationally expensive if the representation is highly overcomplete. Using the Sherman–Morrison–Woodbury formula [14, p. 50]

$$\left(\sigma_\delta^2\mathbf{I}_M + \sigma_u^2\mathbf{VV}^T\right)^{-1} = \frac{1}{\sigma_\delta^2}\mathbf{I}_M - \frac{1}{\sigma_\delta^2}\sigma_u^2\mathbf{V}\left(\mathbf{I}_N + \mathbf{V}^T\frac{1}{\sigma_\delta^2}\sigma_u^2\mathbf{V}\right)^{-1}\mathbf{V}^T\frac{1}{\sigma_\delta^2} \quad (56)$$

$$= \frac{1}{\sigma_\delta^2}\left\{\mathbf{I}_M - \gamma^2\mathbf{V}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\right\}. \quad (57)$$

Therefore

$$\mathbf{A} = \sigma_u\mathbf{ESV}^T\frac{1}{\sigma_\delta^2}\{\mathbf{I}_M - \gamma^2\mathbf{V}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\} \quad (58)$$

$$= \frac{1}{\sigma_u}\gamma^2\mathbf{ESV}^T - \frac{1}{\sigma_u}\gamma^2\mathbf{ES} \times [\{\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V}\} - \mathbf{I}_N](\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T \quad (59)$$

$$= \frac{1}{\sigma_u}\gamma^2\mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T. \quad (60)$$

□

B. Simplified Expression of MSE $\mathcal{E}$ in the General Case (11)

The term $\mathbf{AW}$ is simplified using $\mathbf{A}$ (10) and $\mathbf{V}$ (9) as

$$\mathbf{AW} = \gamma^2\mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\mathbf{V}(\mathbf{ES})^{-1} \quad (61)$$

$$= \mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1} \times \{(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V}) - \mathbf{I}_N\}(\mathbf{ES})^{-1} \quad (62)$$

$$= \mathbf{I}_N - \mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}(\mathbf{ES})^{-1} \quad (63)$$

which yields

$$(\mathbf{I}_N - \mathbf{AW})\mathbf{\Sigma_x}(\mathbf{I}_N - \mathbf{AW})^T = \{\mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}(\mathbf{ES})^{-1}\}\mathbf{ESSE}^T \times \{\mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}(\mathbf{ES})^{-1}\}^T \quad (64)$$

$$= \mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-T}(\mathbf{ES})^T. \quad (65)$$

The second term of $\mathcal{E}$ is also simplified by replacing $\mathbf{A}$ (10)

$$\mathbf{AA}^T = \frac{\gamma^4}{\sigma_u^2}\mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\mathbf{V} \times (\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-T}\mathbf{SE}^T \quad (66)$$

$$= \frac{\gamma^2}{\sigma_u^2}[\mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-T}\mathbf{SE}^T - \mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1} \times (\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-T}\mathbf{SE}^T] \quad (67)$$

and

$$\sigma_\delta^2\mathrm{tr}(\mathbf{AA}^T) = \sigma_\delta^2\frac{\gamma^2}{\sigma_u^2}\cdot\mathrm{tr}[\mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-T}\mathbf{SE}^T - \mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1} \times (\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-T}\mathbf{SE}^T] \quad (68)$$

$$= \mathrm{tr}[\mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-T}\mathbf{SE}^T] - \mathrm{tr}[(\mathbf{I}_N - \mathbf{AW})\mathbf{\Sigma_x}(\mathbf{I}_N - \mathbf{AW})^T]. \quad (69)$$

Therefore

$$\mathcal{E} = \mathrm{tr}[(\mathbf{I}_N - \mathbf{AW})\mathbf{\Sigma_x}(\mathbf{I}_N - \mathbf{AW})^T] + \mathrm{tr}[\mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-T}\mathbf{SE}^T] - \mathrm{tr}[(\mathbf{I}_N - \mathbf{AW})\mathbf{\Sigma_x}(\mathbf{I}_N - \mathbf{AW})^T] \quad (70)$$

$$= \mathrm{tr}[\mathbf{SE}^T\mathbf{ES}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-T}] = \mathrm{tr}[\mathbf{D}(\mathbf{I}_N + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}]. \quad (71)$$

□

C. Simplified Expression of MSE $\mathcal{E}$ for 2-D Data (21)

If $N = 2$, (11) is $\mathcal{E} = \mathrm{tr}\{\mathbf{D}\cdot(\mathbf{I}_2 + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}\}$. Using (20), we get (72)–(75), shown at the bottom of the next page, where

Z is defined in (22). Therefore, we get (76)–(78), shown at the bottom of the page, where we used $\mathrm{Re}(Z) = \sum_{j=1}^{M} \cos 2\theta_j$. □

D. Error Variance of 2-D Data Along the Data Principal Axes

Reconstruction error of each sample point along principal axes is given by $\mathbf{E}^T\boldsymbol{\epsilon}$, and its variance over a set of samples is

$$\mathrm{diag}(\langle \mathbf{E}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\mathbf{E}\rangle) = \mathrm{diag}(\mathbf{E}^T\langle\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\rangle\mathbf{E}) \tag{79}$$

$$= \mathrm{diag}[\mathbf{S}(\mathbf{I}_2 + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}\mathbf{S}] \tag{80}$$

$$= \frac{1}{2}\frac{\sqrt{\lambda_1}+\sqrt{\lambda_2}}{1+\frac{M}{2}\gamma^2}\begin{pmatrix}\sqrt{\lambda_1}\\ \sqrt{\lambda_2}\end{pmatrix} \tag{81}$$

where we used $\langle\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\rangle = \mathbf{E}\mathbf{S}(\mathbf{I}_2 + \gamma^2\mathbf{V}^T\mathbf{V})^{-1}\mathbf{S}\mathbf{E}^T$ [see (70)]. It implies that the error along kth axis, whose data variance is given by λ_k, is proportional to $\sqrt{\lambda_k}$, or, equivalently, the ratio of the error to the data variance is

$$\frac{\sqrt{\lambda_1}}{\lambda_1} : \frac{\sqrt{\lambda_2}}{\lambda_2} = \sqrt{\lambda_2} : \sqrt{\lambda_1}. \tag{82}$$

□

E. Maximum Mutual Information Solution for N-Dimensional Complete Case (46)

The mutual information between the data $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$ is defined by $I(\mathbf{x},\hat{\mathbf{x}}) = H(\mathbf{x}) - H(\mathbf{x}\,|\,\hat{\mathbf{x}})$ where $H(\,\cdot\,)$ denotes the (conditional) entropy of a random variable [2]. If the data distribution is Gaussian, it is straightforward to show that the mutual information is given by

$$I(\mathbf{x},\hat{\mathbf{x}}) = \frac{1}{2}\ln\frac{\det[\mathbf{A}(\gamma^2\mathbf{V}\mathbf{V}^T + \mathbf{I}_M)\mathbf{A}^T]}{\det(\mathbf{A}\mathbf{A}^T)}. \tag{83}$$

When $M = N$ (complete representation), it can further be simplified as

$$I(\mathbf{x},\hat{\mathbf{x}}) = \frac{1}{2}\ln\frac{\det(\mathbf{A})\det(\gamma^2\mathbf{V}\mathbf{V}^T + \mathbf{I}_N)\det(\mathbf{A}^T)}{\det(\mathbf{A})\det(\mathbf{A}^T)}$$

$$= \frac{1}{2}\ln\det(\gamma^2\mathbf{V}\mathbf{V}^T + \mathbf{I}_N) \tag{84}$$

where $\det(\mathbf{A})$ need to be nonzero. Using the eigenvalue decomposition $\mathbf{V}\mathbf{V}^T \equiv \mathbf{P}\boldsymbol{\Omega}\mathbf{P}^T$

$$\det(\gamma^2\mathbf{V}\mathbf{V}^T + \mathbf{I}_N) = \det(\gamma^2\boldsymbol{\Omega} + \mathbf{I}_N)$$

$$= \prod_{k=1}^{N}(\gamma^2\omega_k + 1) \tag{85}$$

where $\mathbf{P}$ is orthogonal and $\mathrm{diag}(\boldsymbol{\Omega}) = (\omega_1,\ldots,\omega_N)$. From the inequality of arithmetic and geometric means

$$\left\{\prod_{k=1}^{N}(\gamma^2\omega_k + 1)\right\}^{1/N} \le \frac{1}{N}\sum_{k=1}^{N}(\gamma^2\omega_k + 1)$$

$$= \gamma^2 + 1 \tag{86}$$

where we used $\sum_{k=1}^{N}\omega_k = \mathrm{tr}(\mathbf{V}\mathbf{V}^T) = N$ [from (8)], and the equality holds iff $\omega_k = 1, \forall k$. □

ACKNOWLEDGMENT

The authors would like to thank Dr. J. Kovačević for helpful comments on the manuscript.

REFERENCES

[1] A. Borst and F. E. Theunissen, "Information theory and neural coding," *Nature Neurosci.*, vol. 2, pp. 947–957, 1999.

$$\mathbf{I}_2 + \gamma^2\mathbf{V}^T\mathbf{V} = \begin{pmatrix}1+\gamma^2\sum_{j=1}^{M}\cos^2\theta_j & \gamma^2\sum_{j=1}^{M}\cos\theta_j\sin\theta_j\\ \gamma^2\sum_{j=1}^{M}\cos\theta_j\sin\theta_j & 1+\gamma^2\sum_{j=1}^{M}\sin^2\theta_j\end{pmatrix} \tag{72}$$

$$= \begin{pmatrix}1+\frac{M}{2}\gamma^2+\frac{1}{2}\gamma^2\sum_{j=1}^{M}\cos 2\theta_j & \frac{1}{2}\gamma^2\sum_{j=1}^{M}\sin 2\theta_j\\ \frac{1}{2}\gamma^2\sum_{j=1}^{M}\sin 2\theta_j & 1+\frac{M}{2}\gamma^2-\frac{1}{2}\gamma^2\sum_{j=1}^{M}\cos 2\theta_j\end{pmatrix} \tag{73}$$

$$|\mathbf{I}_2 + \gamma^2\mathbf{V}^T\mathbf{V}| = \left(1+\frac{M}{2}\gamma^2\right)^2 - \frac{\gamma^4}{4}\left\{\left(\sum_{j=1}^{M}\cos 2\theta_j\right)^2 + \left(\sum_{j=1}^{M}\sin 2\theta_j\right)^2\right\} \tag{74}$$

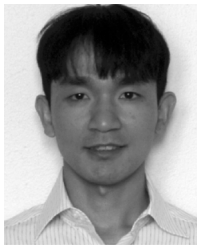
$$= \left(1+\frac{M}{2}\gamma^2\right)^2 - \frac{\gamma^4}{4}|Z|^2 \tag{75}$$

$$\mathcal{E} = \mathrm{tr}\left\{\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}\cdot\frac{1}{|\mathbf{I}_2+\gamma^2\mathbf{V}^T\mathbf{V}|}\begin{pmatrix}1+\frac{M}{2}\gamma^2-\frac{1}{2}\gamma^2\sum_{j=1}^{M}\cos 2\theta_j & -\frac{1}{2}\gamma^2\sum_{j=1}^{M}\sin 2\theta_j\\ -\frac{1}{2}\gamma^2\sum_{j=1}^{M}\sin 2\theta_j & 1+\frac{M}{2}\gamma^2+\frac{1}{2}\gamma^2\sum_{j=1}^{M}\cos 2\theta_j\end{pmatrix}\right\} \tag{76}$$

$$= \frac{1}{|\mathbf{I}_2+\gamma^2\mathbf{V}^T\mathbf{V}|}\left\{\lambda_1\left(1+\frac{M}{2}\gamma^2-\frac{1}{2}\gamma^2\sum_{j=1}^{M}\cos 2\theta_j\right)+\lambda_2\left(1+\frac{M}{2}\gamma^2+\frac{1}{2}\gamma^2\sum_{j=1}^{M}\cos 2\theta_j\right)\right\} \tag{77}$$

$$= \frac{(\lambda_1+\lambda_2)\left(\frac{M}{2}\gamma^2+1\right)-\frac{\gamma^2}{2}(\lambda_1-\lambda_2)\,\mathrm{Re}(Z)}{\left(\frac{M}{2}\gamma^2+1\right)^2-\frac{1}{4}\gamma^4|Z|^2} \tag{78}$$

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[3] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.
[4] M. S. Lewicki and B. A. Olshausen, "Probabilistic framework for the adaptation and comparison of image codes," *J. Opt. Soc. Amer. A*, vol. 16, pp. 1587–1601, 1999.
[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.
[6] D. J. Field, "Relations between the statistics of natural images and the response properteis of cortical cells," *J. Opt. Soc. Amer. A*, vol. 4, pp. 2379–2394, 1987.
[7] E. Doi and M. S. Lewicki, "Sparse coding of natural images using an overcomplete set of limited capacity units," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2005, vol. 17, pp. 377–384.
[8] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*. San Diego, CA: Academic, 1981.
[9] J. J. Atick and A. N. Redlich, "What does the retina know about natural scenes?," *Neural Comput.*, vol. 4, pp. 196–210, 1992.
[10] K. I. Diamantaras and S. Y. Kung, "Channel noise and hidden units," in *Principal Component Neural Networks: Theory and Applications*. New York: Wiley, 1996, pp. 122–145.
[11] K. I. Diamantaras, K. Hornik, and M. G. Strintzis, "Optimal linear compression under unreliable representation and robust PCA neural models," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1186–1195, Oct. 1999.
[12] I. Daubechies, "Discrete wavelet transforms: Frames," in *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992, pp. 53–105.
[13] V. K. Goyal, J. Kovačević, and J. A. Kelner, "Quantized frame expansions with erasures," *Appl. Comput. Harmon. Anal.*, vol. 10, pp. 203–233, 2001.
[14] G. H. Golub and C. F. van Loan, *Matrix Computation*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.



Eizaburo Doi received the B.S. degree in biology in 1996, the M.A. degree in psychology in 1999, and the Ph.D. degree in informatics in 2003, from Kyoto University, Kyoto, Japan.

From 2001 to 2003, he was a Visiting Scholar at the Institute for Neural Computation, University of California, San Diego. He is currently a Postdoctoral Research Associate at the Center of the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA. His research involves theories of computation in neural systems and biologically informed information processing, specifically image, color, and video signal processing.

Doru C. Balcan received the B.S. degree in computer science, in 2000, and the M.S. degree in applied computer science, in 2002, from the Faculty of Mathematics, University of Bucharest, Bucharest, Romania. He is currently pursuing the Ph.D. degree in computer science at Carnegie Mellon University, Pittsburgh, PA.

His research is focused on developing algorithms for efficient and robust signal processing and coding.

Michael S. Lewicki received the B.S. degree in mathematics and cognitive science in 1989 from Carnegie Mellon University (CMU), Pittsburgh, PA, and the Ph.D. degree in computation and neural systems from the California Institute of Technology, Pasadena, in 1996.

From 1996 to 1998, he was a Postdoctoral Fellow with the Computational Neurobiology Laboratory, The Salk Institute, University of California at San Diego, La Jolla. He is currently an Associate Professor with the Computer Science Department, CMU, and in the CMU/University of Pittsburgh Center for the Neural Basis of Cognition. His research involves the study and development of computational approaches to the representation, processing, and learning of structures in natural images and sounds.