# Learning Compressed Sensing

Yair Weiss[1,2]
[1] Hebrew University of Jerusalem
yweiss@cs.huji.ac.il

Hyun Sung Chang[2] and William T. Freeman[2]
[2] MIT CSAIL
{hyunsung,billf} @mit.edu

***Abstract*— Compressed sensing [7], [6] is a recent set of mathematical results showing that sparse signals can be exactly reconstructed from a small number of linear measurements. Interestingly, for ideal sparse signals with no measurement noise, *random* measurements allow perfect reconstruction while measurements based on principal component analysis (PCA) or independent component analysis (ICA) do not. At the same time, for other signal and noise distributions, PCA and ICA can significantly outperform random projections in terms of enabling reconstruction from a small number of measurements.**

**In this paper we ask: given a training set typical of the signals we wish to measure, what are the optimal set of linear projections for compressed sensing ? We show that the optimal projections are in general not the principal components nor the independent components of the data, but rather a seemingly novel set of projections that capture what is still uncertain about the signal, given the training set. We also show that the projections onto the learned *uncertain components* may far outperform random projections. This is particularly true in the case of natural images, where random projections have vanishingly small signal to noise ratio as the number of pixels becomes large.**

## I. Introduction

Compressed sensing [7], [6] is a set of recent mathematical results on a classic question: given a signal $x \in R^n$ and a set of $p$ linear measurements $y \in R^p$, $y = Wx$, how many measurements are required to allow reconstruction of $x$ ?

Obviously, if we knew nothing at all about $x$, i.e. $x$ can be any $n$ dimensional vector, we would need $n$ measurements. Alternatively, if we know our signal $x$ lies in a low-dimensional linear subspace, say of dimension $k$, then $k$ measurements are enough. But what if we know that $x$ lies in a low-dimensional *nonlinear* manifold ? Can we still get away with fewer than $n$ measurements ?

To motivate this question, consider the space of natural images. An image with $n$ pixels can be thought of as a vector in $R^n$ but natural images occupy a tiny fraction of the set of all signals in this space. If there was a way to exploit this fact, we could build cameras with a small number of sensors that would still enable us perfect, high resolution, reconstructions for natural images.

The basic mathematical results in compressed sensing deal with signals that are $k$ sparse. These are signals that can be represented with a small number, $k$ of active (non-zero) basis elements. For such signals, it was shown in [7], [5], that $ck \log n$ *generic* linear measurements are sufficient to recover the signal exactly (with $c$ a constant). Furthermore, the recovery can be done by a simple convex optimization or by a greedy optimization procedure [8].

These results have generated a tremendous amount of excitement in both the theoretical and practical communities. On the theoretical side, the performance of compressed sensing with random projections has been analyzed when the signals are not exactly $k$ sparse, but rather *compressible* (i.e. can be well approximated with a small number of active basis elements) [7], [5] as well as when the measurements are contaminated with noise [11], [19]. On the practical side, applications of compressed sensing have been explored in building "single-pixel" cameras [20], medical imaging [14] and geophysical data analysis [12].

Perhaps the most surprising result in compressed sensing is that perfect recovery is possible with *random projections*. This is surprising given the large amount of literature in machine learning and statistics devoted to finding projections that are optimal in some sense (e.g. [4]). In fact, as we review in the next section, for ideal sparse signals with no measurement noise, *random* measurements significantly outperform measurements based on principal component analysis (PCA) or independent component analysis (ICA). At the same time, for other signal and noise distributions, PCA and ICA can significantly outperform random projections.

In this paper we ask: given a training set typical of the signals we wish to measure, what are the optimal set of linear projections for compressed sensing ? We show that the optimal projections are in general not the principal components nor the independent components of the data, but rather a seemingly novel set of projections that capture what is still uncertain about the signal, given the training set. We also show that the projections onto the learned *uncertain components* may far outperform random projections. This is particularly true in the case of natural images, where random projections have vanishingly small signal to noise ratio as the number of pixels becomes large.

## II. Random Projections versus PCA and ICA

To compare random projections to PCA and ICA, consider the sparse signals illustrated by the image patches in figure 1. In this dataset, each signal $x$ has exactly one non-zero component, and this non-zero component is uniformly distributed in the range $[-U_i, U_i]$. We assume that all indices $i$ have approximately the same range, i.e. $U_i \approx 1$, but to break symmetries we set $U_i = 1 + \epsilon/i$.

We are interested in the probability of correct reconstruction, from a projected signal:

$$y = Wx \tag{1}$$

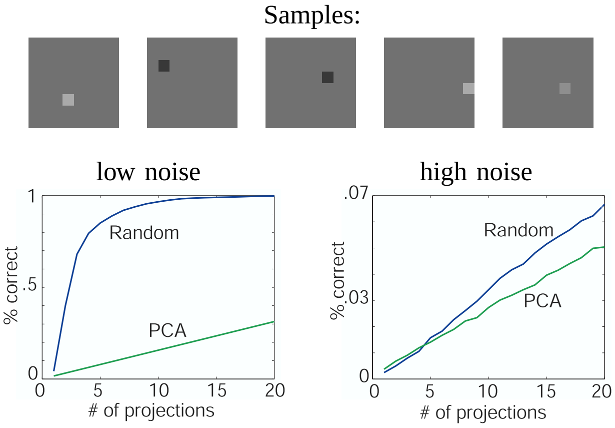

Fig. 1. Comparing PCA and random projections for ideal sparse signals. Each signal has exaclty one nonzero pixel. Random projections work much better than PCA

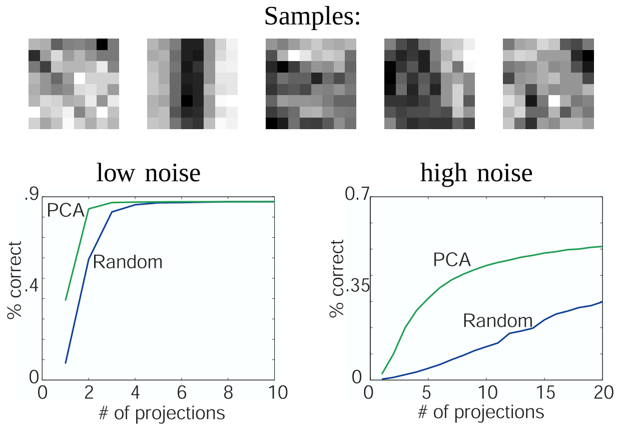

Fig. 2. Comparing PCA and random projections for small image patches. Each signal is an 8×8 patch, randomly cropped from a natural image. Random projections work much worse than PCA

where $W$ is a $p \times n$ matrix.

The probability of correct reconstruction is simply:

$$P_{corr}(W) = E(\delta(x^{MAP} - x)) \tag{2}$$

Where $x^{MAP}$ is the MAP decoding:

$$x^{MAP} = \arg\max_x \Pr(x|y;W) \tag{3}$$

We first calculate $P_{corr}(W)$ for PCA and ICA. The principal components are the eigenvectors of the covariance matrix with maximal eigenvalue. For this dataset, the covariance matrix is a diagonal matrix so the principal components are simply the unit vectors $e_i$. These unit vectors $e_i$ will also be recovered by most ICA algorithms [4].

When will projecting along $p$ unit vectors allow recovery of the original $x$ ? Obviously, this will happen only if the active coefficient in $x$ is one of the $p$ projection directions. This gives:

$$P_{corr}(W_{PCA}) = \frac{p}{n} \tag{4}$$

Thus for a fixed $p$ and large signal dimension $n$ the probability of correct recovery from compressed sensing goes to zero using PCA and ICA. The reason is is that for a large fraction of signals $x$, the projection $y = Wx$ is not unique. It turns out that for a *random* measurement matrix W, where every element of $W$ is chosen independently and randomly, the projections can be shown to be unique with probability one, as long as the number of projections $p$ is greater than or equal to two. This follows from the following lemma (see appendix for a short proof).

**Sparse Random Projection Lemma:** Let $W$ be a random $p \times n$ matrix. Define $y = Wx$. With probability one, if $p \geq 2k$ then any $k$ sparse signal has a unique projection .

This gives:

$$P_{corr}(W_{rand}) = 1, \quad p \geq 2 \tag{5}$$

Thus for this idealized setting, where the signals are highly sparse and there is no measurement noise, random projections are much better than PCA and ICA. Suppose our signal lies in a $10^6$ dimensional space, then two random projections will give perfect recovery while two PCA projections will only reconstruct correctly with probability $2/10^6$.

We emphasize that this advantage of random projections assumes no noise, a highly sparse signal and MAP decoding. Haupt and Nowak [10] have analytically compared random projection to traditonal, pixel-based sampling and shown that in the low SNR regime, pixel-based sampling may actually outperform random projections. Elad [9] has shown that when MAP decoding is replaced with LP decoding, one can improve on random measurement matrices.

To explore the performance under other signal and noise regimes, we conducted experiments using simulations. In these simulations, we assumed the signal $x$ came from a (possibly very large) set of discrete signals $X$. This assumption of discrete $X$ allowed us to perform MAP decoding using exhaustive search and allowed $P_{corr}$ to be nonzero even in the presence of noise.

We first used a discrete version of the sparse signal set, and assumed noisy measurements:

$$y = Wx + \eta \tag{6}$$

where $\eta$ is Gaussian noise with variance $\sigma^2$. To avoid a trivial way of overcoming the noise, the rows of $W$ were constrained to have unit norm.

As shown in figure 1, when the variance of $\eta$ is small ($\sigma^2 = 0.05$), the simulation results are similar to the ideal analytical results. The PCA correct decodings increase linearly with the number of projections, while random projections achieve good performance with a few projections. With larger noise ($\sigma^2 = 0.5$) variances, random measurements are still better, but the advantage is less dramatic.

However, when we change the signal distribution, the results are markedly different (figure 2). We randomly sampled 7200 $8 \times 8$ patches from natural images, and repeated the exact same protocol as used in the synthetic sparse signals. Here, PCA projections work better than random projections, both for small and large amounts of noise.

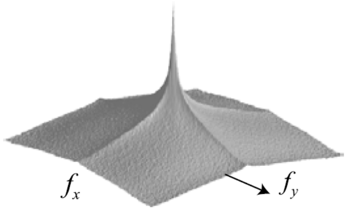

Fig. 3. The power spectrum of natural images falls off as $1/f^2$ (replotted from [17]). We use this fact to prove that the SNR of a random projection approaches zero as the number of pixels grows.

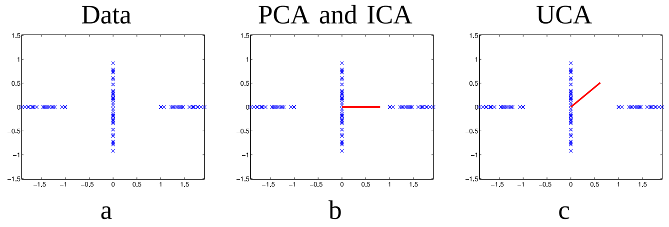

Fig. 4. Learning compressed sensing. Given the dataset $\{x_i\}$ shown on the left, we wish to find a single projection vector $w$ that will enable reconstructing the 2D signal $x$ from its noisy projection, if you are allowed to exploit the training data. Both PCA and ICA give bad projections (middle) while UCA (right) allows nearly perfect estimation of the 2D signal from the 1D projection.

### A. Random Projections and Natural Image Statistics

Can we attribute the poor performance of random projections in our simulations to the experimental protocol? While it is difficult to analytically predict the percent of correct decoding of random projections on natural images, we show in this section that the signal to noise ratio (SNR) of a random unit norm projection on a natural image approaches zero as the number of pixels grows. In contrast, a single unit norm PCA projection gives a constant, positive SNR, even as the number of pixels grows.

Analyzing performance on "natural images" would seem to require a precise definition of the statistics of such images. This is an active area of research (e.g. [16]). But it turns out that simply characterizing the second-order statistics is enough to prove our result. These second-order statistics are well understood: the eigenvectors of the covariance matrix are the Fourier basis elements (since images are spatially stationary) and their eigenvalues fall off with increasing spatial frequency $f$. Furthermore, these eigenvalues (which are just the power spectra of natural images) fall off as a power law [18]) - typically falling off as $1/f^2$. This is a remarkably consistent property - figure 3 shows the mean power of 6000 natural scenes (replotted from [17]) which obeys a power law with the exponent $2.02$.

**Theorem 1: Random Projections and Natural Image** Let $x$ be a natural image with $n$ pixels. Let $w$ be a random projection with (approximately) unit norm - each component $w(i)$ is sampled IID from a zero mean Gaussian with variance $1/n$. Define $y = w^T x + \eta$ with any nonzero noise variance $\sigma^2$. Then for large $n$ the signal to noise ratio SNR(w) is given by:

$$SNR(w) = \frac{1}{n}\frac{\pi^2}{6\sigma^2} \tag{7}$$

with probability one.

**Proof:** The SNR is by definition the ratio of the signal variance, $Var(w^T x)$ and the noise variance $\sigma^2$. Since $w$ is random, the signal variance is also random, but its expectation is given by:

$$\begin{aligned} E[Var(w^T x)] &= E_w[Var(\sum_f \hat{w}(f)\hat{x}(f))] && (8)\\ &= E_w[\sum_f |w\hat{(f)}|^2 Var(\hat{x}(f))] && (9)\\ &= E_w[\sum_f |w\hat{(f)}|^2 \frac{1}{f^2}] && (10)\\ &= \frac{1}{n}\sum_f \frac{1}{f^2} && (11)\\ &\to \frac{1}{n}\frac{\pi^2}{6} && (12) \end{aligned}$$

where we have used Parseval's theorem to rewrite $w^T x$ in terms of the Fourier transform $\hat{w}(f), \hat{x}(f)$ of the projection vector and the signal.

A similar calculation shows that the variance (with respect to w) of the signal variance goes to zero so that almost any random $w$ will have the expected signal variance. Since the signal variance approaches zero, while the noise variance is constant, the SNR of almost any random projection goes to zero as the number of pixel grows.

In contrast, the PCA SNR does not approach zero. In fact, by choosing $w_{PCA}$ to be a unit norm projection whose power spectrum is all in the lowest spatial frequency, we achive:

$$SNR(w_{PCA}) = \frac{1}{1^2\sigma^2} = \frac{1}{\sigma^2} \tag{13}$$

regardless of the number of pixels.

To illustrate this difference, assume the number of pixels, $n$, is a million. An imaging system with $500,000$ different random projections will capture less signal variance than a single PCA projection.

To summarize, neither random projections nor PCA and ICA are in general the best projections for compressed sensing. PCA and ICA work terribly for ideal sparse signals, while random projections work terribly for natural images. What is needed is a new component analysis that takes advantage of both signal and noise statistics.

## III. Uncertain component analysis

Figure 4a shows toy "cross" dataset in $R^2$. Suppose we are only allowed a single linear projection. We are looking for a projection $w$ for which a measurement $y = w^T x$ , plus knowledge of the datasets statistics, would allow us to recover the original signal $x$.

Figure 4b shows the first principal component of the data - the horizontal axis. While this direction of projection maximizes the variance of the projection, it is *not* a good

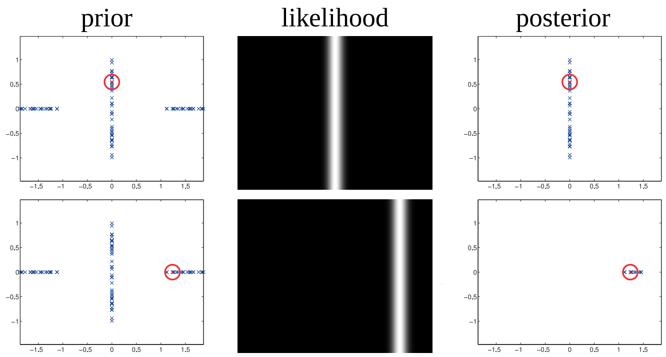

Fig. 5. The intuition behind our definition of uncertain component analysis. For two different signals, and the projection of Fig. 4b, we show the prior, likelihood, and posterior. Maximizing the posterior penalizes projections where many datapoints have nearby projections (top) and rewards projections where the reconstruction of many signals from the projection is easy (e.g. the signal shown on bottom).

measurement vector for compressed sensing. All the points along the vertical axis, will project to the same point, and therefore cannot be reconstructed.

This dataset has two independent components - one for the vertical axis and one for the horizontal axis. Again, both of these directions are bad for compressed sensing since all points on the orthogonal axis will be projected to the same point. So at least for this dataset, neither PCA nor ICA will give the best projection for compressed sensing. In fact, in the noiseless case, both PCA and ICA give the worst projections.

We define a new component analysis, Uncertain Component Analysis (UCA). The first uncertain component, $w^*$, is defined to be the projection direction that maximizes the probability of the data, given the projections and the training data.

$$w^* = \arg \max_{w, \|w\|=1} \prod_i \Pr(x_i | y_i; w) \tag{14}$$

with $y_i = w^T x_i$.

Figure 5 gives some intuition for this definition. The left panel shows the empirical prior probability of $x$, in this case it is simply uniform over all signals in the dataset, and zero for any $x$ not in the dataset. The middle panel shows the likelihood of a signal, given the horizontal projection of the signal shown in red in the left panel. From the generative model this is just:

$$\Pr(y_i | x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - w^T x)^2 / 2\sigma^2} \tag{15}$$

Finally, the right panel shows the empirical posterior probability which is obtained by multiplying the prior and the likelihood and normalizing. Signals whose projection is far from $y_i$ get vanishingly small probability and are not shown.

Note that we are maximizing the *posterior* probability of the input signal $x$ given its projection $y = w^T x$. Since the posterior probability is normalized - the sum of the posterior over all signals in the training set is one, when a datapoint $x_i$ has high posterior probability given its projection (e.g. bottom of figure 5) this means that there are few datapoints in the dataset that give rise to similar projections and successful recovery of $x_i$ from its noisy projection is likely. On the other hand, when a datapoint $x_i$ has low posterior probability given its projection (e.g. top of figure 5) this means that there are many datapoints which give rise to similar projections, and successful recovery of $x_i$ from its noisy projection is unlikely. The UCA definition is therefore trying to maximize the number of datapoints that can be accurately recovered from their noisy projections.

Figure 4c shows the UCA vector for the cross dataset (calculated with $\sigma = 0.05$). Unlike PCA and ICA which choose one of the coordinate axes and therefore will fail to reconstruct points on the orthogonal axis, UCA chooses a vector where most points can be robustly reconstructed from their noisy projection.

### A. Information Maximization

UCA is closely related to a classical approach to finding linear projections called information maximization (or InfoMax [2], [13], [1], [4]). In our setting, InfoMax would search for a matrix $W$ so that the mutual information between the signal $X$ and its noisy projection $Y$ is maximal. Since $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$, and $H(X), H(Y|X)$ are independent of $W$, InfoMax is equivalent to maximizing the entropy of the output $H(Y)$ or minimizing the entropy of the input given the output $H(X|Y)$.

To relate UCA to InfoMax, consider a *stochastic* version of UCA in which, for each signal $x_i$, we *sample* a noisy projection $y_i = W x_i + \eta$, and then, as in ordinary UCA, maximize the probability of the data, $\prod_i \Pr(x_i | y_i)$. In this stochastic version of UCA, the log likelihood of the data will converge to $-H(X|Y)$. Thus, this stochastic UCA is exactly the same as InfoMax. Ordinary UCA, in which $y_i = W x_i$, will be exactly the same as InfoMax for $\sigma \rightarrow 0$ and can be thought of as a deterministic approximation to InfoMax for the general case.

Interestingly, InfoMax has been shown to be equivalent to ICA when the matrix $W$ is invertible [4], [3]. But in the compressed sensing setting, where the number of projections $p$ is less than the dimension of the signal $x$, ICA and UCA can give very different projections.

## IV. Characterizing Optimal Projections

The uncertain component in figure 4c was calculated by searching over a dense sampling of unit norm vectors. It would be better to get an algorithmic solution. The following analytic characterization of $w^*$ allows doing so.

**Observation:** Let $w^*$ be the first uncertain component (equation 14). Then $w^*$ satisfies the following fixed-point equations, relating the data assignment probabilities $q_{ij}$ and the projection direction $w$.

$$\begin{aligned} q_{ij} &= \Pr(x_j | y_i; w) && (16) \\ w &= \operatorname{eigmax} \sum_{i,j} q_{ij} (x_i - x_j)(x_i - x_j)^T && (17) \end{aligned}$$

**Proof:** We first explicitly write the posterior probability:

$$\Pr(x_i|y_i) = \frac{\Pr(x_i)\Pr(y_i|x_i;w)}{\Pr(y_i;w)} \quad (18)$$

Note that the numerator is independent of $w$ since, by the likelihood equation (eq. 15), $\Pr(y_i|x_i) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-0}$. Thus we can alternatively rewrite the UCA criterion:

$$\begin{aligned} w^* &= \arg\max_w \prod_i \frac{1}{\Pr(y_i;w)} && (19) \\ &= \arg\max_w \sum_i -\log\Pr(y_i;w) && (20) \end{aligned}$$

The marginal log likelihood can be rewritten using the familiar, "free energy" functional (e.g. [15]):

$$\begin{aligned} \sum_i -\log\Pr(y_i;w) &= \min_{q:\sum_j q_{ij}=1} -\sum_{ij} q_{ij}\log\Pr(x_j,y_i;w) \\ &\quad + \sum_{ij} q_{ij}\log q_{ij} && (21) \end{aligned}$$

So that:

$$w^* = \arg\max_w \min_q F(w,q) \quad (22)$$

with:

$$F(w,q) = \sum_{ij} q_{ij}(w^T(x_i - x_j))^2 + \sum_{ij} q_{ij}\log q_{ij} \quad (23)$$

(where we have assumed that $\Pr(x_i)$ is uniform over the dataset).

The fixed point equations are simply saying that at the optimal $w^*$, minimizing $F(w,q)$ with respect to $q$ and then maximizing with respect to $w$ should leave us at the same $w$.

We can extend the UCA definition to $p$ vectors by defining the $p \times n$ matrix $W^*$ whose rows are the projection vectors.

$$W^* = \arg\max_{W,WW^T=I} \prod_i \Pr(x_i|y_i;W) \quad (24)$$

with $y_i = Wx_i$.

It is easy to show that the fixed-point equations still hold. The only difference is that the rows of $W^*$ should be the top $p$ eigenvectors of:

$$\sum_{i,j} q_{ij}\left(x_i - x_j\right)\left(x_i - x_j\right)^T$$

This characterization of the fixed-point allows us to understand the behavior of UCA in different special cases.

**Corollary 1: UCA $\Rightarrow$ PCA.** As $\sigma \to \infty$ UCA approaches PCA.

**Proof:** As $\sigma \Rightarrow \infty$ the likelihood (equation 15) approaches a uniform function of $x$, and assuming the prior is uniform over the dataset, the posteriors $q_{ij}$ will also be uniform. Thus the UCA matrix are simply the eigenvectors of $(x_i - x_j)(x_i - x_j)$ and these are the principal components of the data.

**Corollary 2: UCA=PCA for $p$ dimensional data.** If the data $\{x_i\}$ lie in a $p$ dimensional subspace, then the UCA vectors and the top $p$ PCA vectors span the same subspace.

**Proof:** We can define a new dataset whose elements are the difference vectors $d_{ij} = (x_i - x_j)$. The UCA vectors are the prinicpal components of the dataset $\{d_{ij}\}$ where each difference vector is weighted by $q_{ij}$. Since $x_i, x_j$ both lie in a $p$ dimensional subspace, so does $q_{ij}d_{ij}$ and hence UCA will recover an orthogonal basis of this $p$ dimensional subspace. On the other hand, if the data lie in a $p$ dimensional basis, PCA will also recover an orthogonal basis of this $p$ dimensional subspace.

**Corollary 3: UCA=Random for noiseless sparse data** If the data $\{x_i\}$ are $k$ sparse in any basis, and $p \geq 2k$ then for $\sigma \to 0$ a random $W$ matrix maximizes the UCA cost function with probability one.

**Proof:** This follows from the sparse random projection lemma - with probability one, no two $k$ sparse signals can have the same random projection. This means that the empirical posterior probability $\Pr(x_i|y_i;W)$ will approach one as $\sigma \to 0$ for all datapoints $x_i$.

While the fixed-point equations show that under certain conditions, PCA and UCA give the same projections, they also highlight the difference. PCA tries to maximize the variance of the projections, which can be thought of as maximizing the *average* distance between the projections of any two signals. UCA maximizes a *weighted average* distance between the projections of any two signals, weighted by the probability of assignment to each observation. The weighted average gives high weight to pairs of signals whose projections are similar (determined by the noise level $\sigma$). This makes sense in terms of robust reconstruction. For a given noise level $\sigma$ two signals whose projected distance is $10\sigma$ are almost as good as two signals whose projected distance is $100\sigma$.

### A. Algorithms

Direct calculation gives the gradient of the log likelihood with respect to $w$:

$$\frac{\partial \log P}{\partial w} = \left(\sum_{ij} q_{ij}\left(x_i - x_j\right)\left(x_i - x_j\right)^T\right) w \quad (25)$$

with $q_{ij}$ as in the fixed-point equations (eq. 16).

However, in our experiments, this gradient update can be very slow (especially since one needs to enforce the unit norm constraint). Often, better performance is achieved iterating a dampened version of the fixed-point equations (eq. 16,17) (moving only part-way from the old values to the new ones).

Note that unlike other uses of the free energy in machine learning (e.g. the EM algorithm), iterating the fixed-point equations is not guaranteed to improve the likelihood at every iteration. This is because the global optimum is a saddle point of $F(p,q)$ and not a minimum. Nevertheless, if we do happen to converge to a fixed-point, it is guaranteed to be a local constrained optimum of the UCA cost function.

## V. Results

We first estimated uncertain components for ideal sparse signals for different imaging noise values $\sigma^2$ and different

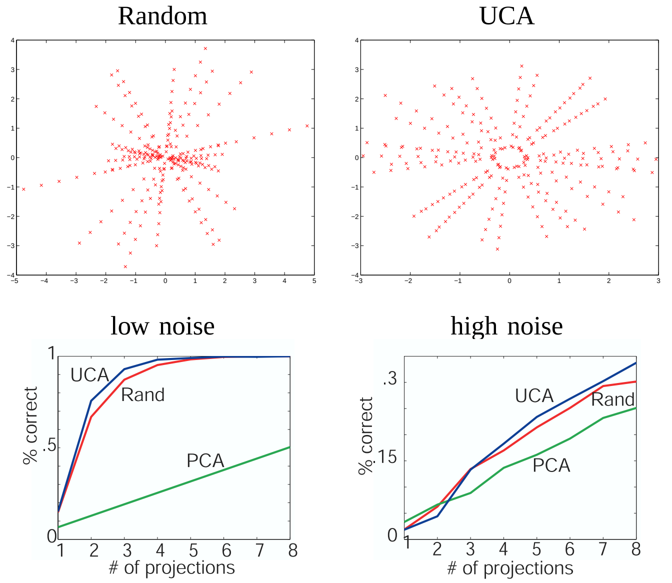

Fig. 6. UCA results on ideal sparse images. Each image has exactly one nonzero component. **Top:** Projection of the full dataset from 16 dimensions onto two dimensions using random projections and UCA. **Bottom:** Comparison of percentage of correct decodings as a function of number of projections, for different noise levels.

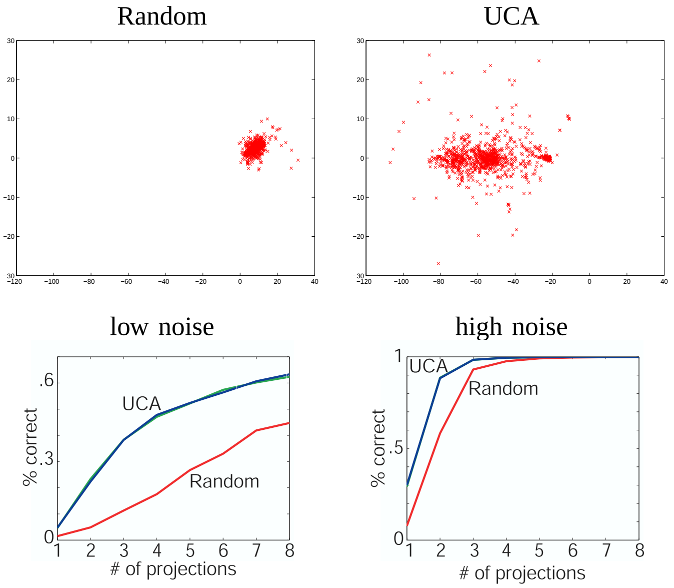

Fig. 7. UCA results on natural image patches. **Top:** Projection of the full dataset from 16 dimensions onto two dimensions using random projections and UCA. **Bottom:** Comparison of percentage of correct decodings as a function of number of projections, for different noise levels. The UCA and PCA results are almost identical so the green line is occluded by the blue line.

numbers of projections $p$. The signals were in $4 \times 4$ image patches, and each patch had one nonzero pixel. The value in that pixel was an integer distributed uniformly in the range $[-16, 16]$. Recall that for noiseless measurements, random projections are optimal for such signals (from the sparse random projections lemma).

As expected by corollary 3, when $\sigma^2$ is very small, any random projection is a fixed-point of the algorithm. But when $\sigma^2$ is large, UCA learns projections that are still incoherent (i.e. they have nonzero elements for all pixels) but nonrandom. To visualize the learnt UCA projections, we plot in figure 6 the projections of the sparse signals into two dimensions using random projections (left) and the UCA projections (right). Since all signals are $1$ sparse in the high dimensional space, the signal set defines a discrete set of rays in high dimensions, all starting at the origin. In both the random projections and the UCA projections, one can still observe the projected rays, but UCA finds a projection in which these rays are (approximately) emanating at regular angles. Thus UCA is finding a projection in which the number of signals with similar projections is smaller than in a random projection. Figure 6 compares the decoding performance of the different projections (again, using MAP decoding). As expected, UCA performs slightly better than random projections, and both UCA and random perform much better than PCA.

In our second experiment, we estimated uncertain components for a set of $1,000$ $4 \times 4$ image patches randomly sampled from natural images. For this dataset, we found that UCA learns projections that are nearly identical to PCA. This is to be expected from the $1/f^2$ power spectrum of natural images, which means that the image patches lie (approximately) in a low dimensional subspace. In fact, for this dataset, $99\%$ of the variance is captured by the first two principal components. Thus corollary 2 predicts that UCA and PCA should give very similar results for this data. Again, to visualize the UCA projections versus a random projection, we show projections of the image signals into two dimensions using random projections (figure 7 left) and the UCA projections (right). Note that the variance of the random projections is significantly smaller than that of the UCA projections, as predicted by theorem 1. We repeated the experiment with $10,000$ $15 \times 15$ image patches and (as predicted by theorem 1) found that random projections capture an even smaller amount of signal variance. Figure 7 compares the decoding performance of the different projections (again, using MAP decoding). As expected, UCA performs identically to PCA and much better than random projections.

## VI. Discussion

Suppose we are allowed to take a small number of linear projections of signals in a dataset, and then use the projections plus our knowledge of the dataset to reconstruct the signals. What are the best projections to use? We have shown that these projections are not necessarily the principal components nor the independent components of the data nor random projections, but rather a new set of projections which we call uncertain components. We formalized this notion by maximizing the probability of a signal given its projection, and derived fixed-point equations that need to be satisifed at the optimum. Our experiments show that learning projections can give much better performance compared to simply using random projections. This is particularly true for natural image signals, where random projections don't perform well and can be shown to have vanishingly small signal to noise ratio as the number of pixels increases.

### *Acknowledgements*

Funding for this research was provided by NGA NEGI-1582-04-0004, Shell Research, ONR-MURI Grant N00014-06-1-0734 and the Israeli Science Foundation.

## REFERENCES

[1] JJ Atick. Could information theory provide an ecological theory of sensory processing. *Network: computation in neural systems*, 3:213–251, 1992.

[2] H.B. Barlow. Possible principles underlying the transformations of sensory messages. In W.A. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, 1961.

[3] S. Becker. Modelling the mind: From circuits to systems. In Simon Haykin, Jose C. Principe, Terrence J. Sejnowski, and John McWhirter, editors, *New Directions in Statistical Signal Processing: From sytems to brain*. MIT Press, 2005.

[4] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, (37):3327–3338, 1997.

[5] E. Candes and T. Tao. Near optimal signal recovery from random projections and universal encoding strategies. *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[6] Emmanuel Candes and Justin Romberg. Practical signal recovery from random projections. In *SPIE Symposium on Electronic Imaging*, 2005.

[7] D. Donoho. Compressed sensing. *IEEE Transactions Info Theory*, 52(4):1289–1306, 2006.

[8] Marco Duarte, Michael Wakin, Dror Baron, and Richard Baraniuk. Universal distributed sensing via random projections. In *Proc. International Conference on Information Processing in Sensor Networks*, 2006.

[9] Michael Elad. Optimized projections for compressed sensing. 2006. submitted.

[10] Jarvis Haupt and Robert Nowak. Compressive sampling vs. conventional imaging. In *ICIP*, pages 1269–1272, 2006.

[11] Jarvis Haupt and Robert D. Nowak. Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52(9):4036–4048, 2006.

[12] Tim Lin and Felix. J. Herrmann. Compressed wavefield extrapolation. *Geophysics*, 2007. to apear.

[13] R. Linsker. Perceptual neural organization: some approaches based on network models and information theory. *Annual Rev Neurosci.*, 13:257–81, 1990.

[14] M. Lustig, J. M. Santos, D. L. Donoho, and J. M. Pauly. k-t sparse:high frame rate dynamic mri exploiting spatio-temporal sparsity. In *Proc. 14th. Annual Meeting of ISMRM*, 2006.

[15] R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental and other variants. In M.I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. MIT Press, 1999.

[16] E.P. Simoncelli. Statistical models for images:compression restoration and synthesis. In *Proc Asilomar Conference on Signals, Systems and Computers*, pages 673–678, 1997.

[17] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14:391–412, 2003.

[18] A. van der Schaaf and J.H. van Hateren. Modelling the power spectra of natural images: statistics and information. *Vision Research*, 36(17):2759–70, 1996.

[19] Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. In *Proc. Allerton Conference on Communication, Control, and Computing*, 2006.

[20] Michael B. Wakin, Jason N. Laska, Marco F. Duarte, Dror Baron, Shriram Sarvotham, Dharmpal Takhar, Kevin F. Kelly, and Richard G. Baraniuk. An architecture for compressive imaging. In *ICIP*, pages 1273–1276. IEEE, 2006.

## APPENDIX: PROOF OF SPARSE RANDOM PROJECTION LEMMA

**Sparse Random Projection Lemma:** Let $W$ be a random $p \times n$ matrix. Define $y = Wx$. With probability one, if $p \geq 2k$ then any $k$ sparse signal has a unique projection .

**Proof:** Suppose, by way of contradiction, that there exists a second $k$ sparse vector $z$ exists, so that $Wx = Wz$. Let $I$ be a set of $2k$ indices that includes all the indices on which both $x$ and $z$ are nonzero. Note that since both $x$ and $z$ are $k$ sparse, their set of nonzero indices cannot be of size greater than $2k$. Define $W_I$ to be a $p \times |I|$ submatrix of $W$ obtained by taking all columns in $I$ and all rows. By the defintion of matrix multiplication $Wx = W_I x_I$ and $Wz = W_I z_I$ (since the zero elements can be ignored in the matrix multiply). This means that $W_I z_I = W_I x_I$ with $x_I \neq z_I$ which implies that the $|I|$ columns of $W$ are linearly dependent. But since these columns of $W$ are $|I|$ *random* $p$ dimensional vectors and $|I| \leq p$ this happens with probability zero.