DENSE ERROR CORRECTION VIA L1-MINIMIZATION

John Wright and Yi Ma

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
1308 West Main Street, Urbana, Illinois 61801, USA

ABSTRACT

We study the problem of recovering a non-negative sparse signal $\boldsymbol{x} \in \mathbb{R}^n$ from highly corrupted linear measurements $\boldsymbol{y} = A\boldsymbol{x}+\boldsymbol{e} \in \mathbb{R}^m$, where $\boldsymbol{e}$ is an unknown (and unbounded) error. Motivated by an observation from computer vision, we prove that for highly correlated dictionaries A, any non-negative, sufficiently sparse signal $\boldsymbol{x}$ can be recovered by solving an ℓ^1-minimization problem:

$$\min \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subject to} \quad \boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}.$$

If the fraction ρ of errors is bounded away from one and the support of $\boldsymbol{x}$ grows sublinearly in the dimension m of the observation, for large m, the above ℓ^1-minimization recovers all sparse signals $\boldsymbol{x}$ from almost all sign-and-support patterns of $\boldsymbol{e}$. This suggests that accurate and efficient recovery of sparse signals is possible even with nearly 100% of the observations corrupted.

Index Terms— Error correction, Signal representation, Signal reconstruction

1. INTRODUCTION

Recovery of high-dimensional sparse signals or errors has been one of the fastest growing research areas in signal processing in the past few years. A lot of excitement has been generated by remarkable successes in application areas such as image and speech processing, bioinformatics, communications, as well as computer vision and pattern recognition.[1]

One notable, and somewhat non-traditional, application of sparse representation is in automatic face recognition [3]. For each person, a set of training images are taken under different illuminations. Stack the images as columns of a matrix $A \in \mathbb{R}^{m\times n}$, where m is the number of pixels in an image and n is the total number of images for all the subjects of interest. We can try to represent a new query image, stacked as a vector $\boldsymbol{y} \in \mathbb{R}^m$ as a linear combination of all the images, i.e., $\boldsymbol{y} = A\boldsymbol{x}$ for some $\boldsymbol{x} \in \mathbb{R}^n$. Since in practice n can potentially be larger than m, the equations can be underdetermined and the solution $\boldsymbol{x}$ may not be unique. In this context, it is natural to seek a sparse solution for $\boldsymbol{x}$ whose large non-zero coefficients provide information about the subject's true identity. This can be done by solving an ℓ^1-minimization problem:

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{y} = A\boldsymbol{x}. \tag{1}$$

The problem becomes more interesting if the query image $\boldsymbol{y}$ is severely occluded or corrupted, as in Figure 1 (inset). In this case, one needs to solve a corrupted set of linear equations $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{e} \in \mathbb{R}^m$ is an unknown (and possibly unbounded) error vector. For sparse errors $\boldsymbol{e}$ and tall matrices A ($m > n$), Candes and Tao [4] proposed to multiply the equation $\boldsymbol{y} = A\boldsymbol{x}+\boldsymbol{e}$ with a matrix B such that $BA = 0$, and then use ℓ^1-minimization to recover the error vector $\boldsymbol{e}$ from the underdetermined linear equation $B\boldsymbol{y} = B\boldsymbol{e}$.

In face recognition (and many other applications), n can be larger than m and A can be full rank. One cannot directly apply the above technique even if the error $\boldsymbol{e}$ is known to be very sparse. To resolve this difficulty, in [3], the authors proposed to instead seek $[\boldsymbol{x}, \boldsymbol{e}]$ together as the sparsest solution to the extended equation $\boldsymbol{y} = [A\ \mathtt{I}]\,\boldsymbol{w}$ with $\boldsymbol{w} = \left[\begin{smallmatrix}\boldsymbol{x}\\ \boldsymbol{e}\end{smallmatrix}\right] \in \mathbb{R}^{m+n}$, by solving the extended ℓ^1-minimization problem:

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_1 \quad \text{subject to} \quad \boldsymbol{y} = [A\ \ \mathtt{I}]\,\boldsymbol{w}. \tag{2}$$

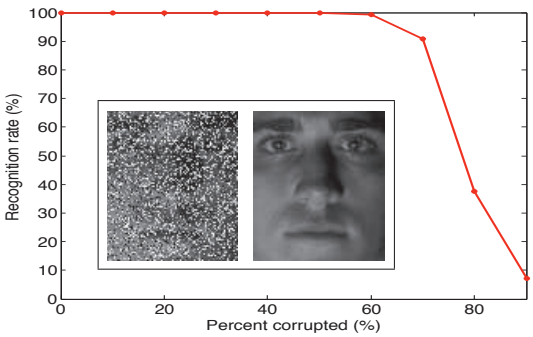


Fig. 1. **Face recognition under random corruption.** Inset left: face with 60% of pixels randomly corrupted. Inset right: face recovered by sparse representation within a database of face images [3]. Red curve: recognition rate across the entire range of corruption. It remains almost perfect up to 60% random corruption.

This work is partially supported by grants NSF CRS-EHS-0509151, NSF CCF-TF-0514955, ONR YIP N00014-05-1-0633, and NSF IIS 07-03756. John Wright is also supported by a Microsoft Fellowship (sponsored by Microsoft Live Labs, Redmond).

[1] For a more thorough survey of this rapidly expanding literature, see [1, 2].

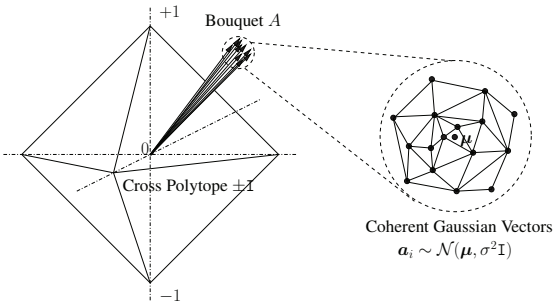


Fig. 2. **The "cross-and-bouquet" model.** Left: the bouquet A and the crosspolytope spanned by the matrix $\pm\mathtt{I}$. Right: tip of the bouquet magnified; it is a collection of iid Gaussian vectors with small variance σ^2 and common mean vector $\boldsymbol{\mu}$. The cross-and-bouquet polytope is spanned by vertices from both the bouquet A and the cross $\pm\mathtt{I}$.

This seemingly minor modification to the previous error correction approach has dramatic consequences on the performance of robust face recognition. Solving the modified ℓ^1-minimization enables almost perfect recognition even with more than 60% pixels of the query image are arbitrarily corrupted (see Figure 1 for an example), far beyond the amount of error that can theoretically be corrected by the previous error correction method [4].

Although ℓ^1-minimization is expected to recover sufficiently sparse solutions with overwhelming probability for general systems of linear equations (see [5]), it is rather surprising that it works for the equation $\boldsymbol{y} = [A\ \mathtt{I}]\,\boldsymbol{w}$ at all. In the application described above, the columns of A are highly correlated. As m becomes large (i.e. the resolution of the image becomes high), the convex hull spanned by all face images of all subjects is only an extremely tiny portion of the unit sphere $\mathbb{S}^{m-1}$.[2] For example, the images in Figure 1 lie on $\mathbb{S}^{8,063}$. The smallest inner product with their normalized mean is 0.723; they are contained within a spherical cap of volume $\leq 1.47 \times 10^{-229}$. These vectors are tightly bundled together as a "bouquet," whereas the vectors associated with the identity matrix and its negative $\pm\mathtt{I}$ together[3] form a standard "cross" in $\mathbb{R}^m$, as illustrated in Figure 2. Notice that such a "cross-and-bouquet" matrix $[A\ \mathtt{I}]$ is neither incoherent nor (restrictedly) isometric, at least not uniformly. Also, the density of the desired solution $\boldsymbol{w}$ is not uniform either. The $\boldsymbol{x}$ part of $\boldsymbol{w}$ is usually a very sparse non-negative vector, but the $\boldsymbol{e}$ part can be very dense and have arbitrary signs. Existing results for recovering sparse signals suggest that ℓ^1-minimization may have difficulty in dealing with such signals, contrary to its empirical success in face recognition.

We have experimented with similar cross-and-bouquet type models where the matrix A is a random matrix with highly correlated column vectors. The simulations reveal something even more striking and puzzling phenomenon: As the dimension m increases (and the sample size n grows in proportion), the percentage of errors that the ℓ^1-minimization (2) can correct seems to approach 100%! This may seem surprising, but this paper explains why this should be expected.

2. PROBLEM SETTING AND MAIN RESULT

Motivated by the face recognition example introduced above, we consider the problem of recovering a non-negative[4] sparse signal $\boldsymbol{x}_0 \in \mathbb{R}^n$ from highly corrupted observations $\boldsymbol{y} \in \mathbb{R}^m$:

$$\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{e}_0,$$

where $\boldsymbol{e}_0 \in \mathbb{R}^m$ is a sparse vector of errors of arbitrary magnitude. The model for $A \in \mathbb{R}^{m\times n}$ should capture the idea that it consists of small deviations about a mean, hence a "bouquet." In this paper, we consider the case where the columns of A are iid samples from a Gaussian distribution:

$$A = [\boldsymbol{a}_1 \dots \boldsymbol{a}_n] \in \mathbb{R}^{m\times n}, \quad \boldsymbol{a}_i \sim_{iid} \mathcal{N}\left(\boldsymbol{\mu}, \tfrac{\nu^2}{m}\mathtt{I}_m\right), \qquad \|\boldsymbol{\mu}\|_2 = 1, \qquad \|\boldsymbol{\mu}\|_\infty \leq C_\mu m^{-1/2}. \tag{3}$$

Together, the two assumptions on the mean force it to remain incoherent with the standard basis (or "cross") as $m \to \infty$.

We study the behavior of the solution to the ℓ^1-minimization (2) for this model, in the following asymptotic scenario:

Assumption 1 (Weak Proportional Growth)**.** *A sequence of signal-error problems exhibits weak proportional growth with parameters $\delta > 0, \rho \in (0,1), C_0 > 0, \eta_0 > 0$, denoted $WPG_{\delta,\rho,C_0,\eta_0}$, if as $m \to \infty$,*

$$\frac{n}{m} \to \delta, \quad \frac{\|\boldsymbol{e}_0\|_0}{m} \to \rho, \quad \|\boldsymbol{x}_0\|_0 \leq C_0\, m^{1-\eta_0}. \tag{4}$$

This should be contrasted with the "total proportional growth" (TPG) setting of, e.g., [6], in which the number of nonzero entries k_1 in the signal $\boldsymbol{x}_0$ also grows as a fixed fraction of the dimension. In that setting, one might expect a sharp phase transition in the combined sparsity of $(\boldsymbol{x}_0, \boldsymbol{e}_0)$ that can be recovered by ℓ^1-minimization. In WPG, on the other hand, we observe a striking phenomenon not seen in TPG: the correction of arbitrary fractions of errors. This comes at the expense of the stronger assumption that $k_1 \doteq \|\boldsymbol{x}_0\|_0$ is sublinear, an assumption that is valid in some real applications such as the face recognition example above.

In the following, we say the cross-and-bouquet model is ℓ^1*-recoverable at* $(I, J, \boldsymbol{\sigma})$ if for all $\boldsymbol{x}_0 \geq 0$ with support I and $\boldsymbol{e}_0$ with support J and signs $\boldsymbol{\sigma}$,

$$(\boldsymbol{x}_0, \boldsymbol{e}_0) \;=\; \arg\min \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subject to} \quad A\boldsymbol{x} + \boldsymbol{e} = A\boldsymbol{x}_0 + \boldsymbol{e}_0, \tag{5}$$

and the minimizer is uniquely defined. From the geometry of ℓ^1-minimization, if (5) does not hold for some pair $(\boldsymbol{x}_0, \boldsymbol{e}_0)$,

[2] At first sight, this seems somewhat surprising as faces of different people look so different to human eyes. That is probably because human brain has adapted to distinguish highly correlated visual signals such as faces or voices.

[3] We allow the error $\boldsymbol{e}$ to have both positive and negative signs.

[4] The non-negativity assumption is important: in the highly coherent systems considered here, ℓ^1-minimization does not recover signals $\boldsymbol{x}_0$ with arbitrary signs. Geometrically, this would require vectors from the "bouquet" to "see" through the crosspolytope to vectors that are nearly antipodal to them.

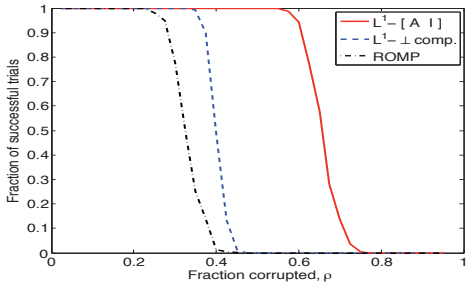


Fig. 3. **Comparison with alternative approaches.** Fraction of correct successes, as a function of the corruption level, ρ. Here, $\nu = 0.05, \delta = 0.25$. The extended ℓ^1-minimization "$L^1 - [A\ \mathtt{I}]$" outperforms the approach of [4] ("$L^1 - \perp$ comp") and ROMP [8].

then it does not hold for any $(\boldsymbol{x}, \boldsymbol{e})$ with the same signs and support as $(\boldsymbol{x}_0, \boldsymbol{e}_0)$ [7]. Understanding ℓ^1-recoverability at each $(I, J, \boldsymbol{\sigma})$ completely characterizes which solutions to $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$ can be correctly recovered. In this language, our main result can be stated more precisely as:

Theorem 1 (Error Correction with the Cross-and-Bouquet)**.** *For any $\delta > 0$, $\exists\, \nu_0(\delta) > 0$ such that if $\nu < \nu_0$ and $\rho < 1$, in $WPG_{\delta,\rho,C_0,\eta_0}$ with A distributed according to* (3)*, if the error support J and signs $\boldsymbol{\sigma}$ are chosen uniformly at random, then as $m \to \infty$,*

$$\mathbb{P}_{A,J,\sigma}\left[\ell^1\text{-recoverability at } (I, J, \boldsymbol{\sigma})\ \ \forall\, I \in \binom{[n]}{k_1}\right] \to 1.$$

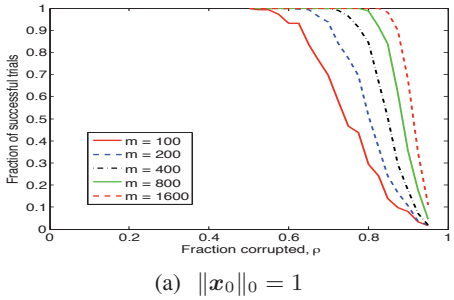
In other words, as long as the bouquet is sufficiently tight, asymptotically ℓ^1-minimization recovers any non-negative sparse signal from almost any error with support size less than 100%. The proof of the above result relies on a careful characterization of the faces of the polytope spanned by the cross and bouquet. While it requires only standard ideas from geometry, linear algebra and measure concentration, the details are far beyond the scope of this paper. We refer the interested reader to [2].

3. SIMULATIONS AND EXPERIMENTS

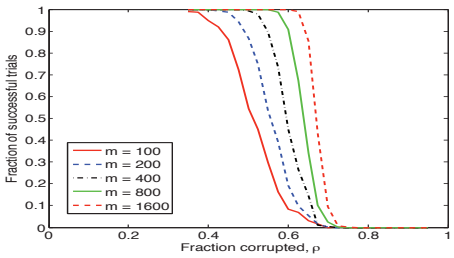
a) Comparison with alternative approaches. We first compare the performance of the extended ℓ^1-minimization (2) to two alternative approaches. The first is the error correction approach of [4], which multiplies by a full rank matrix B such that $BA = 0$,[5] solves $\min \|\boldsymbol{e}\|_1$ subj $B\boldsymbol{e} = B\boldsymbol{y}$, and then subsequently recovers $\boldsymbol{x}$ from the clean system of equations $A\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{e}$. The second is the Regularized Orthogonal Matching Pursuit (ROMP) algorithm [8], a state-of-the-art greedy method for recovering sparse signals.

For this experiment, the ambient dimension is $m = 500$; the parameters of the CAB model are $\nu = 0.05, \delta = 0.25$. We fix the signal support $k_1 = 15$, and vary the fraction of errors from 0 to 0.95. For each error fraction, we generate 500 independent problems. Figure 3 plots the fraction of successes for each of the three algorithms, as a function of error density ρ. The extended ℓ^1-minimization is denoted "$L^1 - [A\ \mathtt{I}]$" (red curve), while the alternative approach of [4] is denoted "$L^1 - \perp$ comp" (blue curve). Whereas both competitors break down around 40% corruption, the extended ℓ^1-minimization continues to succeed with high probability even beyond 60% corruption.

[5] This comparison requires $n \ll m$ although our method is not limited to this case.



(a) $\|\boldsymbol{x}_0\|_0 = 1$



(b) $\|\boldsymbol{x}_0\|_0 = m^{1/2}$

Fig. 4. **Error correction in weak proportional growth.** (a), (b): Simulated examples with $\delta = 0.25$, $\nu = 0.05$. Fraction of successful recoveries as a function of error density ρ, for varying m. In (a), $\|\boldsymbol{x}_0\|_0 = 1$, while in (b), $\|\boldsymbol{x}_0\|_0 = m^{1/2}$. In both cases, as m increases, the fraction of errors that can be corrected approaches 1.

b) Error correction capacity. While the previous experiment demonstrates the advantages of the extended ℓ^1-minimization (2) for the CAB model, Theorem 1 suggests that more is true: As the dimension increases, the fraction of errors that the extended ℓ^1-minimization can correct should approach one. We generate problem instances with $\delta = 0.25$, $\nu = 0.05$, for varying $m = 100, 200, 400, 800, 1600$. We again plot the fraction of correct recoveries as a function of ρ in Figure 4 (a) and (b). In Figure 4(a), we fix $k_1 = 1$, while in (b), k_1 grows as $k_1 = m^{1/2}$. In both cases, as m increases, the fraction of errors that can be corrected also increases.

c) Phase Transition in Total Proportional Growth. Theorem 1 does not provide any explicit information about the behavior of ℓ^1-minimization when the signal support k_1 grows proportionally to m: $k_1/m \to \rho_1 \in (0, 1)$. Based on intuition from more homogeneous polytopes (especially [9]), we might expect that when k_1 also exhibits proportional growth, an asymptotically sharp phase transition between guaranteed recovery and guaranteed failure will occur at some critical error fraction $\rho^* \in (0, 1)$. We investigate this empirically

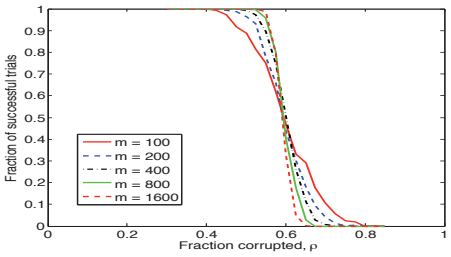


Fig. 5. **Phase transition in total proportional growth.** When $\|\boldsymbol{x}_0\|_0$ grows linearly, we observe an asymptotically sharp phase transition. Here, $\nu = 0.05, \delta = 0.25, \|\boldsymbol{x}_0\|_0 = 0.05m$.

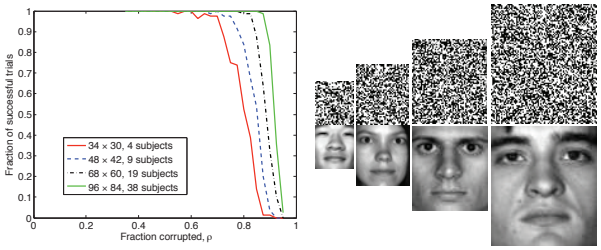


Fig. 6. **Error correction with real face images.** Left: fraction of correct recoveries for varying levels of occlusion. Right: examples of correct recovery for each resolution. The fraction of corruption is chosen so that the probability of correct recovery is 50%.

here by again setting $\delta = 0.25$, $\nu = 0.05$, but this time allowing $k_1 = 0.05m$. Figure 5 plots the fraction of correct recovery for varying error fractions ρ, as m grows: $m = 100, 200, 400, 800, 1600$. In this proportional growth setting, we see an increasingly sharp phase transition, near $\rho = 0.6$.

d) Error correction with real face images. Finally, we return to the motivating example of face recognition under varying illumination and random corruption. We use the Extended Yale B face database [10], which tests illumination sensitivity of face recognition algorithms. We form the matrix A from images in Subsets 1 and 2, which contain mild-to-moderate illumination variations. Each column of the matrix A is a $w \times h$ face image, stacked as a vector in $\mathbb{R}^m$ ($m = w \times h$). Here, the weak proportional growth setting corresponds to the case when the total number of image pixels grows proportionally to the number n of face images. Since the number of images per subject is fixed, this is the same as the total image resolution growing proportionally to the number of subjects. We vary the image resolutions through the range 34×30, 48×42, 68×60, 96×84. The matrix A is formed from images of $4, 9, 19, 38$ subjects, respectively, corresponding to $\delta \approx 0.09$. Here, $\nu \approx 0.3$. In face recognition, the sublinear growth of $\|\boldsymbol{x}_0\|_0$ comes from the fact that the observation should ideally be a linear combination of only images of the same subject. Various estimates of the required number of images, k_1, appear in the literature, ranging from 5 to 9. Here, we fix $k_1 = 7$, and generate the (clean) test image synthetically as a linear combination of k_1 training images from a single subject. For each resolution considered, and for each error fraction, we generate 75 trials. Figure 6 plots the fraction of successes as a function of the fraction of corruption. As predicted by Theorem 1, the fraction of errors that can be corrected again approaches 1 as the data size increases.

4. DISCUSSIONS AND FUTURE WORK

This work analyzes one scenario, motivated by a practical imaging application, in which the performance of ℓ^1-minimization greatly exceeds what might be expected based on existing theory. We believe that similar analysis of other practical applications may likewise reveal phenomena of broad practical and theoretical interest. Even for this simple model, there is still much to be done. In particular, while the ℓ^1-minimizer is known to be stable under noise, it would be interesting to provide an explicit stability bound, as a function of ν. We would also like to investigate the relevance of this result to compressive image acquisition problems, by analyzing how much error tolerance remains after randomly projecting $\boldsymbol{y}$ onto a low dimensional subspace.

5. REFERENCES

[1] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *submitted to SIAM Review*, 2007.

[2] J. Wright and Y. Ma, "Dense error correction via ℓ^1-minimization," Tech. Rep. UILU-ENG-08-2210, DC 237, `http://perception.csl.uiuc.edu/~jnwright/Wright08-IT.pdf`, University of Illinois at Urbana-Champaign, 2008.

[3] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *To appear in IEEE Trans. PAMI*, 2008.

[4] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. IT*, vol. 51, no. 12, 2005.

[5] D. Donoho, "For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[6] D. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ^1-norm near solution approximates the sparest solution," *preprint*, 2004.

[7] D. Donoho, "Neighborly polytopes and sparse solution of underdetermined linear equations," *preprint*, 2005.

[8] D. Needell and R. Vershynin, "Signal recovery from inaccurate and incomplete measurements via regularized orthogonalized matching pursuit," preprint `http://www.math.ucdavis.edu/~dneedell/`, 2007.

[9] D. Donoho and J. Tanner, "Counting faces of randomly projected polytopes when the projection radically lowers dimension," preprint, `http://www.math.utah.edu/~tanner/`, 2007.

[10] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. PAMI*, vol. 27, no. 5, pp. 684–698, 2005.