Error Correcting Algorithms for Sparsely Correlated Regressors

Anonymous Authors[1]

Abstract

Autonomy and adaptation of machines requires that they be able to measure their own errors. We consider the advantages and limitations of such an approach when a machine has to measure the error in a regression task. How can a machine measure the error of regression sub-components when it does not have the ground truth for the correct predictions? A compressed sensing approach applied to the error signal of the regressors can recover their precision error without any ground truth. It allows for some regressors to be *strongly correlated* as long as not too many are so related. Its solutions, however, are not unique - a property of ground truth inference solutions. Adding ℓ_1–minimization as a condition can recover the correct solution in settings where error correction is possible. We briefly discuss the similarity of the mathematics of ground truth inference for regressors to that for classifiers.

1. Introduction

An autonomous, adaptive system, such as a self-driving car, needs to be robust to self-failures and changing environmental conditions. To do so, it must distinguish between self-errors and environmental changes. This chicken-and-egg problem is the concern of *ground truth inference algorithms* - algorithms that measure a statistic of ground truth given the output of an ensemble of evaluators. They seek to answer the question - Am I malfunctioning or is the environment changing so much that my models are starting to break down?

Ground truth inference algorithms have had a spotty history in the machine learning community. The original idea came from (Dawid et al., 1979) and used the EM algorithm to solve a maximum-likelihood equation. This enjoyed a brief renaissance in the 2000s due to advent of services like

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Amazon Turk. Our main critique of all these approaches is that they are parametric - they assume the existence of a family of probability distributions for how the estimators are committing their errors. This has not worked well in theory or practice (Zheng et al., 2017).

Here we will discuss the advantages and limitations of a non-parametric approach that uses compressed sensing to solve the ground truth inference problem for noisy regressors (Corrada-Emmanuel & Schultz, 2008). *Ground truth* is defined in this context as the correct values for the predictions of the regressors. The existence of such ground truth is taken as a postulate of the approach. More formally,

Definition 1 (Ground truth postulate for regressors). *All regressed values in a dataset can be written as,*

$$\hat{y}_{i,r} = y_{i,true} + \delta_{i,r}, \qquad (1)$$

where $y_{i,true}$ does not depend on the regressor used.

In many practical situations this is a very good approximation to reality. But it can be violated. For example, the regressors may have developed their estimates at different times while a $y(t)_{i,\text{true}}$ varied under them.

We can now state the ground truth inference problem for regressors as,

Definition 2 (Ground truth inference problem for regressors). *Given the output of R aligned regressors on a dataset of size D,*

$$((\hat{y}_{1,1}, \hat{y}_{2,1}, \ldots, \hat{y}_{D,1}), \ldots, (\hat{y}_{1,R}, \hat{y}_{2,R}, \ldots, \hat{y}_{D,R})),$$

estimate the error moments for the regressors,

$$\overline{\delta_{r_1}\,\delta_{r_2}} := \frac{1}{D}\sum_{i=1}^{D}\delta_{i,r_1}\,\delta_{i,r_2}, \qquad (2)$$

and

$$\overline{\delta_r} := \frac{1}{D}\sum_{i=1}^{D}\delta_{i,r}, \qquad (3)$$

without the true values, $\{y_{i,true}\}$.

The separation of moment terms that are usually combined to define a covariance[1] between estimators is deliberate

---

[1]To wit, covariance can be expressed as $\sigma_{r_i,r_j} = \overline{\delta_{r_1}\,\delta_{r_2}} - \overline{\delta_{r_1}}\,\overline{\delta_{r_2}}$.

and relates to the math for the recovery as the reader will understand shortly.

As stated, the ground truth inference problem for sparsely correlated regressors was solved in (Corrada-Emmanuel & Schultz, 2008) by using a compressed sensing approach to recover the $R(R+1)/2$ moments, $\overline{\delta_{r_1}\delta_{r_2}}$, for unbiased ($\overline{\delta_r} \approx 0$) regressors. Even the case of some of the regressors being strongly correlated is solvable. Sparsity of non-zero correlations is all that is required. Here we point out that the failure to find a *unique* solution for biased regressors still makes it possible to detect and correct biased regressors under the same sort of engineering logic that allows bit flip error correction in computers.

2. Independent, unbiased regressors

We can understand the advantages and limitations of doing ground truth inference for regressors by simplifying the problem to that of independent, un-biased regressors. The inference problem then becomes a straightforward linear algebra one that can be understood without the complexity required when some unknown number of them may be correlated.

Consider two regressors giving estimates,

$$\hat{y}_{i,r_1} = y_{i,\text{true}} + \delta_{i,r_1} \qquad (4)$$
$$\hat{y}_{i,r_2} = y_{i,\text{true}} + \delta_{i,r_2}. \qquad (5)$$

By the Ground Truth Postulate, these can be subtracted to obtain,

$$\hat{y}_{i,r_1} - \hat{y}_{i,r_2} = \delta_{i,r_1} - \delta_{i,r_2} \qquad (6)$$

Note that the left-hand side involves *observable* values that do not require any knowledge of $y_{i,\text{true}}$. The right hand side contains the error quantities that we seek to estimate. Squaring both sides and averaging over all the datapoints in the dataset we obtain our primary equation,

$$\Delta^2_{r_1,r_2} := \frac{1}{D}\sum_{i=1}^{D}(\hat{y}_{i,r_1} - \hat{y}_{i,r_2})^2$$
$$= \overline{\delta^2_{r_1}} - 2\overline{\delta_{r_1}\delta_{r_2}} + \overline{\delta^2_{r_2}}. \qquad (7)$$

Since we are assuming that the regressors are independent in their errors ($\overline{\delta_{r_1}\delta_{r_2}} \approx 0$), we can simplify 7 to,

$$\overline{\delta^2_{r_1}} + \overline{\delta^2_{r_2}} = \Delta^2_{r_1,r_2}. \qquad (8)$$

This is obviously unsolvable with a single pair of regressors. But for three it is. It leads to the following linear algebra equation,

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \overline{\delta^2_1} \\ \overline{\delta^2_2} \\ \overline{\delta^2_3} \end{bmatrix} = \begin{bmatrix} \Delta^2_{1,2} \\ \Delta^2_{1,3} \\ \Delta^2_{2,3} \end{bmatrix} \qquad (9)$$

Figure 1. Least squares recovery of the error moments for three independent regressors.

An application of this simple equation to a synthetic experiment with three noisy regressors is shown in Figure 1. Just like any least squares approach, and underlying topology for the relation between the different data points is irrelevant. Hence, we can treat, for purposes of experimentation, each pixel value of a photo as a ground truth value to be regressed by the synthetic noisy regressors - in this case with uniform error. To highlight the multidimensional nature of equation 6, we randomized each of the color channels but made one channel more noisy for each of the pictures. This simulates two regressors being mostly correct, but a third one perhaps malfunctioning. Since even synthetic experiments with independent regressors will result in spurious non-zero cross-correlations, we solved the equation via least squares[2].

3. Biased, independent regressors

So why are these impressive results not better known and a standard subject in Statistics 101 courses? There may be various reasons for this. The first one is that statistics concerns itself mostly with the imputation of the parameters of a model for the signal being studied, not the error of the regressors with themselves. We are not trying to impute properties of the true signal, but of the error signal between the regressors. A regressor may put out a signal $\hat{y}_{i,r}$, but its error signal $\delta_{i,r}$ could be completely different. Additionally, Statistics has historically swayed from moment methods (such as the approach taken here) to maximum likelihood methods and back. Moment methods are much more practi-

---

[2]The full compressive sensing solution being wholly unnecessary in this application where we know the regressors are practically independent.

cal now with the advent of big data and cheap computing power.

The other more important reason is that the above math fails for the case of *biased* regressors. We can intuitively understand that because eq. 6 is invariant to a global bias, Δ, for the regressors. We are not solving for the *full average* error of the regressors but their *average precision* error,

$$
\begin{bmatrix} \overline{\epsilon_1} \\ \vdots \\ \overline{\epsilon_r} \end{bmatrix} = \begin{bmatrix} \overline{\delta_1} \\ \vdots \\ \overline{\delta_r} \end{bmatrix} + \begin{bmatrix} \Delta \\ \vdots \\ \Delta \end{bmatrix} \tag{10}
$$

We can only determine the error of the regressors modulus some unknown global bias. This, by itself, would not be an unsurmountable problem since global shifts are easy to fix. From an engineering perspective, accuracy is cheap while precision is expensive[3]. The more problematic issue is that it would not be able to determine correctly who is biased if they are biased *relative* to each other.

Let us demonstrate that by using eq 6 to estimate the average bias, $\overline{\delta_r}$, for the regressors. Averaging over both sides, we obtain for three independent regressors, the following equation[4],

$$
\begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \overline{\delta_1} \\ \overline{\delta_2} \\ \overline{\delta_3} \end{bmatrix} = \begin{bmatrix} \Delta_{1,2} \\ \Delta_{1,3} \\ \Delta_{2,3} \end{bmatrix} \tag{11}
$$

The rank of this matrix is two. This means that the matrix has a one-dimensional null space. In this particular case, the subspace is spanned by a constant bias shift as noted previously. Nonetheless, let us consider the specific case of three regressors where two of them have an equal constant bias,

$$
\begin{bmatrix} \overline{\delta_1} \\ \overline{\delta_2} \\ \overline{\delta_3} \end{bmatrix} = \begin{bmatrix} 0 \\ \Delta \\ \Delta \end{bmatrix} . \tag{12}
$$

This would result in the Δ_{r_1,r_2} vector,

$$
\begin{bmatrix} \Delta_{1,2} \\ \Delta_{1,3} \\ \Delta_{2,3} \end{bmatrix} = \begin{bmatrix} -\Delta \\ -\Delta \\ 0 \end{bmatrix} . \tag{13}
$$

The general solution to Eq. 10 would then be,

$$
\begin{bmatrix} \overline{\delta_1} \\ \overline{\delta_2} \\ \overline{\delta_3} \end{bmatrix} = \begin{bmatrix} -\Delta \\ 0 \\ 0 \end{bmatrix} + c \begin{bmatrix} \Delta \\ \Delta \\ \Delta \end{bmatrix} . \tag{14}
$$

---

[3]Examples are (a) the zeroing screw in a precision weight scale, (b) the number of samples needed to measure a classifier's accuracy when it is of unknown accuracy versus when we know it is either, say, 1% or 99% accurate. The former situation is more accurate on average but less precise. The latter one, precise but inaccurate.

[4]The observable statistic, Δ_{r_1,r_2}, is equal to $1/D \sum (\hat{y}_{i,r_1} - \hat{y}_{i,r_2})$.

This seems to be a failure for any ground truth inference for noisy regressors. Lurking underneath this math is the core idea of compressed sensing: pick the value of c for the solutions to eq. 14 that minimizes the ℓ_1 norm of the recovered vector. When such a point of view is taken, non-unique solutions to ground truth inference problems can be re-interpreted as error detecting and correcting algorithms. We explain.

4. Error detection and correction

Suppose, instead, that only one of the three regressors was biased,

$$
\begin{bmatrix} \overline{\delta_1} \\ \overline{\delta_2} \\ \overline{\delta_3} \end{bmatrix} = \begin{bmatrix} \Delta \\ 0 \\ 0 \end{bmatrix} . \tag{15}
$$

This would give the general solution,

$$
\begin{bmatrix} \overline{\delta_1} \\ \overline{\delta_2} \\ \overline{\delta_3} \end{bmatrix} = \begin{bmatrix} \Delta \\ 0 \\ 0 \end{bmatrix} + c \begin{bmatrix} \Delta \\ \Delta \\ \Delta \end{bmatrix} , \tag{16}
$$

with c an arbitrary, constant scalar. If we assume that errors are sparse, then an ℓ_1-minimization approach would lead us to select the solution,

$$
\begin{bmatrix} \overline{\delta_1} \\ \overline{\delta_2} \\ \overline{\delta_3} \end{bmatrix} = \begin{bmatrix} \Delta \\ 0 \\ 0 \end{bmatrix} . \tag{17}
$$

The algorithm would be able to detect and correct the bias of a single regressor. If we wanted more reassurance that we were picking the correct solution then we could use 5 regressors. When the last two have constant bias, the general solution is,

$$
\begin{bmatrix} \overline{\delta_1} \\ \overline{\delta_2} \\ \overline{\delta_3} \\ \overline{\delta_4} \\ \overline{\delta_5} \end{bmatrix} = \begin{bmatrix} -\Delta \\ -\Delta \\ -\Delta \\ 0 \\ 0 \end{bmatrix} + c \begin{bmatrix} \Delta \\ \Delta \\ \Delta \\ \Delta \\ \Delta \end{bmatrix} . \tag{18}
$$

With the corresponding ℓ_1-minimization solution of,

$$
\begin{bmatrix} \overline{\delta_1} \\ \overline{\delta_2} \\ \overline{\delta_3} \\ \overline{\delta_4} \\ \overline{\delta_5} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \Delta \\ \Delta \end{bmatrix} . \tag{19}
$$

This is the same engineering logic that makes practical the use of error correcting codes when transmitting a signal over a noisy channel. Our contribution is to point out that the same logic *also* applies to *estimation errors* by regressors trying to recover the true signal.
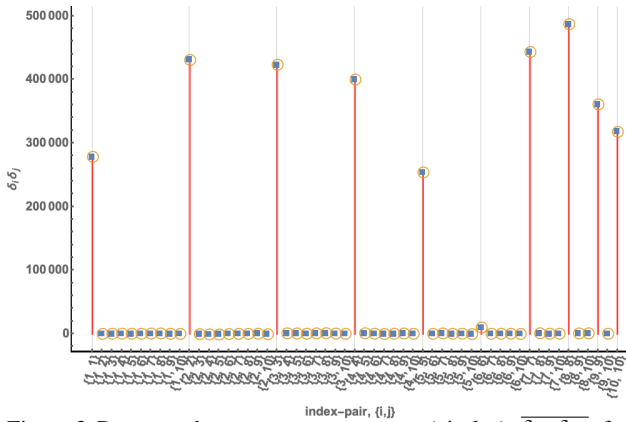
Figure 2. Recovered square error moments (circles), $\overline{\delta_{r_1}\delta_{r_2}}$, for the true error moments (squares) of 10 synthetic regressors on the pixels of a 1024x1024 image. Recovering algorithm does not know which vector components correspond to the strong diagonal signal, the (i,i) error moments.

5. Conclusions

A compressed sensing algorithm for recovering the average error moments of an ensemble of noisy regressors exists. Like other ground truth inference algorithms, it leads to non-unique solutions. However, in many well-engineered systems, errors are sparse and mostly uncorrelated when the machine is operating normally. Algorithms such as this one can then detect the beginning of malfunctioning sensors and algorithms.

We can concretize the possible applications of this technique by considering a machine such as a self-driving car. Optical cameras and range finders are necessary sub-components. How can the car detect a malfunctioning sensor? There are many ways this already can be done (no power from the sensor, etc.). This technique adds another layer of protection by potentially detecting anomalies earlier. In addition, it allows the creation of *supervision* arrangements such as having one expensive, precise sensor coupled with many cheap, imprecise ones. As the recovered error moment matrix in Figure 2 shows, many noisy sensors can be used to benchmark a more precise one (the (sixth regressor {6,6} moment in this particular case). As (Corrada-Emmanuel & Schultz, 2008) demonstrate, it can also be used on the final output of algorithms. In the case of a self-driving car, a depth map is needed of the surrounding environment - the output of algorithms processing the sensor input data. Here again, one can envision supervisory arrangements where quick, imprecise estimators can be used to monitor a more expensive, precise one.

There are advantages and limitations to the approach proposed here. Because there is no maximum likelihood equation to solve, the method is widely applicable. The price for this flexibility is that no generalization can be made. There is no theory or model to explain the observed errors

- they are just estimated robustly for each specific dataset. Additionally, the math is easily understood. The advantages or limitations of a proposed application to an autonomous, adaptive system can be ascertained readily. The theoretical guarantees of compressed sensing algorithms are a testament to this (Foucart & Rauhaut, 2013). Finally, the compressed sensing approach to regressors can handle *strongly*, but *sparsely*, correlated estimators.

We finish by pointing out that non-parametric methods also exist for classification tasks. This is demonstrated for independent, binary classifiers (with working code) in (Corrada-Emmanuel, 2018). The only difference is that the linear algebra of the regressor problem becomes polynomial algebra. Nonetheless, there we find similar ambiguities due to non-unique solutions to the ground truth inference problem of determining average classifier accuracy without the correct labels. For example, the polynomial for unknown prevalence (the environmental variable) of one of the labels is quadratic, leading to two solutions. Correspondingly, the accuracies of the classifiers (the internal variables) are either x or $1 - x$. So a single classifier could be, say, 90% or 10% accurate. The ambiguity is removed by having enough classifiers - the preferred solution is where one of them is going below 50%, not the rest doing so.

References

Corrada-Emmanuel, A. Ground truth inference of binary classifier accuracies - the independent classifiers case. https://github.com/andrescorrada/ground-truth-problems-in-business/blob/master/classification/IndependentBinaryClassifiers.pdf, 2018.

Corrada-Emmanuel, A. and Schultz, H. Geometric precision errors in low-level computer vision tasks. In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, pp. 168–175, Helsinki, Finland, 2008.

Dawid, P., Skene, A. M., Dawidt, A. P., and Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, pp. 20–28, 1979.

Foucart, S. and Rauhaut, H. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, New York, 2013.

Zheng, Y., Li, G., Li, Y., Shan, C., and Cheng, R. Truth inference in crowdsourcing: Is the problem solved? In *Proceedings of the VLDB Endowment*, volume 10, no. 5, 2017.