Recovering approximate single cell distribution from aggregate measurements

Pratik Worah [1]

Abstract

We design an algorithm that approximately recovers single cell data, for example, scRNA-seq data for a small subset of genes, using aggregate measurements, for example, unlabeled in-situ hybridization data and RNA-seq data, from a sample. The technical crux of our algorithm involves compressed sensing based sparse recovery when the measurement matrix is unknown, only its statistical distribution is known.

1. Introduction

The importance of single cell data is increasing as we try to understand fundamental processes in biology and apply them in medicine. However, the costs of gathering single cell data are prohibitive, so a method to even approximately recover single cell data from a small number of inexpensive aggregate level measurements may prove useful. In this paper, we explore one such idea using compressed sensing type algorithms. We show, using simulations, that it is possible to approximately recover the underlying single cell distribution of RNA expression using ℓ_1 minimization, even when the problem does not fit the standard compressed sensing paradigm.

To summarize our results: From a computational biology perspective, our contribution is encapsulated in Algorithm 1 which uses compressed sensing without knowing the exact measurement matrix, just its statistical properties, to approximately reconstruct the original microscopic distribution using only a few aggregate measurements of the sample. For lack of data and resources, we don't run experiments, but we run simulations (in Section 3) using publicly available scRNA-seq data ((Bastidas-Ponce et al., 2019; Bergen et al., 2020))[1] as the ground truth, to see how Algorithm 1 and its derivatives may eventually perform in the real world.

[1]Google Research, USA. Correspondence to: Pratik Worah <pworah@google.com>.

[1]Their data is available at https://scvelo.readthedocs.io/en/stable/scvelo.datasets.pancreas/ and https://github.com/theislab/pancreatic-endocrinogenesis/

From a theoretical perspective, a full version of this paper slightly generalizes a state of the art result in compressed sensing for Gaussian matrices by (Raskutti et al., 2010).

1.1. Background

A typical genomic assay can fall somewhere on the granularity spectrum of measuring single cell genomic data, for example, scRNA-seq or fluorescence in-situ hybridization (FISH) on the one hand; to mapping just the aggregate average value, for example, RNA-seq, on the other hand. Clearly, the former capture more information than the latter, as they measure the entire distribution of gene expression in a tissue, as opposed to just the average or total gene expression. FISH, in particular, has already found some use in healthcare (see for example (Hu et al., 2014)). At the same time methods like FISH have practical drawbacks. For example, it requires elaborate sample preparation in order to measure the fluorescence intensity (see (Haroon et al., 2013) for the details involved in a one-off in-solution FISH).

The assay implied by Algorithm 1 does rely on in-situ hybridization, but it does not require fluorescent markers or labels. Moreover, since it does not rely on measuring intensity, it should not require elaborate sample preparation like FISH. All it requires is that the hybridization take place first in the cellular compartments and then cellular compartments be made permeable (say via detergents) to perform RNA-seq to generate an observation – a co-ordinate for the input vector y to Algorithm 1.

The high-level idea is that a small amount of the statistical information of the single cell distribution is transferred to the aggregate measurement via the hybridization reaction in individual cellular compartments, see Assumption 2.2 for formal details. Therefore, if we make enough independent measurements then we may be able to approximately recover the entire distribution from aggregate measurements only, in certain circumstances.

2. Technical overview of our results and methods

Problem formulation: Suppose we have a set of noisy measurements $C := \{c_1, ..., c_N\}$ corresponding to N "cells". For example, the measurements could be the (spliced and

unspliced) RNA concentrations corresponding to Cpe (Carboxypeptidase E[2]) in a sample of $N = 250$ cells from the pancreatic scRNA-seq data by (Bastidas-Ponce et al., 2019). Suppose further that we bin (i.e., partition and round) the concentrations into n intervals. For example, for the Cpe sample, set $n = 100$, which gives: for values varying from $M_1 := 0$ to $M_2 := 35$ – the minimum and maximum values in our sample, 100 bins (intervals) of width 0.35 each. Let $x_{M_1}, x_{M_1+1*0.35}, ..., x_{M_2}$ denote the number of cells with measurement values $M_1, M_1 + 1 * 0.35, ..., M_2$ respectively. Thus the vector x represents the (non-normalized) distribution of the measurement values. Note that tuple of measurements C and the distribution x are unknown, and the goal of this paper is to approximate x using few measurements (i.e., small m).

While measurements, in this case RNA concentrations, can take on a continuum of values, they typically tend to cluster into a few values, in any snapshot, for informative genes. Figure 1 illustrates this fact. Out of 4000 genes (post-

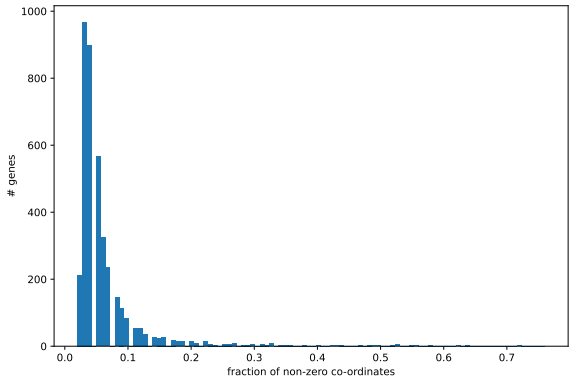

Figure 1. RNA expression is clustered into few values for most genes

filtering) in the murine pancreatic data-set of (Bastidas-Ponce et al., 2019), fewer than 5% of them had more than 10% of co-ordinates values set as non-zero when measured up to two significant figures of accuracy. Therefore, we may think of x as a sparsely supported distribution, likely characterizing various cell types.

Assumption 2.1. We assume that the unknown distribution to be recovered, i.e., x, is s-sparse, for some $s \simeq \log^{O(1)}(n)$.

The critical question to consider is the nature of the noise in the measurement set C. It's critical because the noise distribution captures a small amount of statistical information in each cellular compartment that will be passed on to the aggregate measurement. Here's where the biochemistry of hybridization plays a role: For a cell i, the measurement c_i above (say the concentration for Cpe) would be made by an in-situ hybridization oligonucleotide probe. Note that we don't need to observe the value of c_i (for any i). All that is needed is that the excess amounts of probe are present in the cell i, so that fraction of the probe proportional to c_i hybridizes with Cpe RNA, in cell i. Of course, not all probes will hybridize. Moreover, most probes will hybridize independently of one another, and therefore one can expect roughly Gaussian statistics. Therefore, we will work under the following assumption:

Assumption 2.2. We assume that the mean fraction of probes in any cell i that hybridize will be proportional to its cellular concentration c_i, the proportionality constant depending on the forward and backward hybridization reaction rates (of Cpe in our running example). Moreover, the variance will also be proportional to the concentration (of Cpe in our running example). Furthermore, we assume that the distribution of $c_1, ..., c_N$ is an independent multinomial (equivalently Gaussian, for our purposes) with mean vector $\{k \cdot c_1, ..., k \cdot c_N\}$ and variance $\{k' \cdot c_1, ..., k' \cdot c_N\}$, for proportionality constants k and k' that can be estimated empirically.

Finally, we repeat the noisy measurements above m times, i.e., we take m samples $C(1), ..., C(m)$ and we assume that we can observe the sums: $\{\sum_{i\in[N]} c_i(j) : j \in [m]\}$ for $C(1), .., C(m)$.[3] The result is a m-dimensional vector of observations y. The sums denote a simple aggregate measurement, which tell us what total fraction of RNA probes were hybridized. Note that the information about the distribution x is now present in the mean and the variance of each of the m observations in y.

In more mathematical terms, we have the following compressed sensing problem at hand:

$$(V + \Gamma)x = y, \tag{1}$$

where V is a $m \times n$ matrix corresponding to the means i.e.,

$$V = \begin{pmatrix} M_1 & M_1+1 & M_1+2 & \dots \\ \vdots & \vdots & \ddots & \\ M_1 & M_1+1 & & M_2 \end{pmatrix}$$

where we have assumed that M_1 and M_2 are the minimum and maximum values of the c_i; and Γ is a matrix of mean zero independent binomial random variables corresponding to the noise and captures the variances. Thus, by Assumption 2.2, Γ is chosen (by nature) from a Gaussian random ensemble of the form:

$$\begin{pmatrix} N(0,k'M_1) & N(0,k'(M_1+1)) & N(0,k'(M_1+2)) & \dots \\ \vdots & \vdots & \ddots & \\ N(0,k'M_1) & N(0,k'(M_1+1)) & & N(0,k'M_2) \end{pmatrix}$$

[2]Cpe is a gene involved in the synthesis of neuropeptides and peptide hormones.

[3]Notation: $[N] \equiv \{1, 2, .., N\}$.

where $N(0, \sigma^2)$ denotes a Normal random variable with mean 0 and variance σ^2, and for simplicity we have assumed a partition of $[M_1, M_2]$ into bins of width 1. Note vector y is known (observations) and x (distribution) is unknown. Moreover, while V is known, the values of the noise matrix Γ are unknown – this differentiates our problem from a standard compressed sensing problem, where the matrix $V + \Gamma$ is known.

2.1. Compressed sensing with unknown measurement matrix

Suppose we replaced Γ by an independent sample with the same statistics. So let $\tilde{\Gamma}$ be a sampled noise matrix and Γ the actual noise implied by the experiment. Both have the same distribution by construction, but they are independent of each other. Similarly, let x be the actual probability distribution and $\tilde{x}$ be a probability distribution that is the solution of Equation 2. Suppose $\tilde{x} = x + \Delta$ for some Δ. Then we have from our problem set-up:

$$(V + \tilde{\Gamma})\tilde{x} = (V + \Gamma)x \quad (2)$$

$$(V + \tilde{\Gamma})(x + \Delta) = (V + \Gamma)x \quad (3)$$

$$(V + \tilde{\Gamma})\Delta = \Gamma x - \tilde{\Gamma}x. \quad (4)$$

Since x is s-sparse and each row of Γ and $\tilde{\Gamma}$ consists of independent Gaussians, the Hoeffding bound (see for example (Dembo & Zeitouni, 1998)) implies the following observation on the noise and hence can be used to bound the reconstruction error (which is given in a full version of this paper).

Lemma 2.3. *The ℓ_2 norm of $(V + \tilde{\Gamma})\Delta$, i.e., the magnitude of η, is $O(ms)$, where the constants in the $O(\cdot)$ notation depends on the range of our measurements.*

2.1.1. Algorithm

Lemma 2.3 implies that we may solve the following compressed sensing problem to recover x:

$$\tilde{\Gamma}x = y = \Gamma x + \eta, \quad (5)$$

where the ℓ_2 norm of the noise term ϵ is $O(ms)$, and we require that $x \geq 0$. Here we have used a (known) matrix $\tilde{\Gamma}$ with a similar distribution as the (unknown) matrix Γ. Note that the bound by (Raskutti et al., 2010) shows that $m = O(\log n)$ samples suffice to reconstruct the sparse vector in the problem formulated thus far. If we have several instances of the problem where $\tilde{\Gamma}$ and Γ are drawn repeatedly and independently then it is natural to expect that averaging over the solution of several such instances may lead to a reasonable approximation of the (unknown) x.

This discussion prompts the averaging method in Algorithm 1. In the algorithm, we assume that M_1 and M_2 denote the minimum and maximum values of the observations are known.

Algorithm 1 Approximate reconstruction of distribution

1: Input: Measurement vector $y \in \mathbb{R}^m$, consisting of m (noisy) aggregate observations, a guess of the sparsity s of the unknown distribution to be found, k – the averaging parameter, and η – a parameter that upper bounds the ℓ_2 norm of the noise.
2: Output: An approximation of the underlying n-dimensional probability distribution that generated y.
▷ Begin algorithm:
3: Center the observations y around 0 (by subtracting the mean $\bar{y}$ from each co-ordinate). Set $y' = y - (\bar{y}, .., \bar{y})$.
4: Sample a $m \times n$ random Gaussian ensemble Γ. The variance matrix of each row being a diagonal matrix, with the diagonal: $(vM_1, v(M_1 + 1), ..., vM_2)$. Here v is a constant that is set so that the total variance of each row equals the variance of the co-ordinates of y.
5: The constraint in the compressed sensing problem is: $\|\tilde{\Gamma}z - y'\|_2 \leq \eta$. Solve the convex program to recover z.
6: Sort all coordinates of z in inverse order, and set all but first s to 0, to obtain z_s.
7: Repeat steps 3, 4 and 5 $k \simeq \log n$ times; compute $\frac{z_s(1)+...+z_s(k)}{k}$, zero out all but its s largest values, normalize its mass to 1 so the resulting $\tilde{p}$ is a probability distribution supported on s points and return it.

In fact, we can further optimize our reconstruction algorithm as follows: We introduce new types of hybridization probes so that a probe can bind upto two Cpe RNA segments simultaneously in the cell compartments, and then eventually count the total number of such "doubly" hybridized probes. This generates a measurement corresponding to $\sum_{i\in[N]} c_i^2$ as opposed to $\sum_{i\in[N]} c_i$ – a quadratic (non-linear) measurement. Since we assume a Gaussian noise model where the noise (variance) is proportional to the concentration of the product (see Assumption 2.2), the corresponding variances will also scale as $\sum_{i\in[N]} c_i^2$.

There's nothing unique about quadratics, other different non-linear function based probes can be designed and used. Suppose that $\mathcal{F}$ denotes the set of non-linear functions that can be successfully implemented using hybridization probes – *permissible non-linear functions*. Then following (2) above, Algorithm 1 can be modified so that the rows of measurement matrices Γ are picked according to a set of Normal distributions with variances specified by functions in $\mathcal{F}$. Let $\mathcal{M}$ be the set of such *permissible covariance matrices*. A longer discussion about permissible measurement probes is given in the appendix.

3. Simulations on pancreatic scRNA-seq data

In this section, we provide details of our simulation and some examples of recovered distributions and error estimates as well.

The plots in Figures 2 and 3 use a uniform random sample of $N = 250$ cells from the $\simeq 4000$ cells in the original data (Bastidas-Ponce et al., 2019). We assume $n = 100$, i.e., range of expression values for any single gene (say Cpe) are rounded into 100 bins, for computational tractability reasons we keep n small. The original distribution of a gene like Cpe for our sample of 250 cells would then be the histogram of: number of cells/N (y-axis) vs (rounded) gene expression value (x-axis). Recall that, from Figure 1, most genes will have sparse histograms i.e., they typically take only a few different values in our snapshot. Therefore, we assume sparsity parameter $s = 20$ in Algorithm 1, for our simulations. Figures 2 and 3 show the actual and reconstructed histogram (using Algorithm 1, and few samples, i.e., m).

Genes like Cpe and Nnat play a role in hormone secretion so they are 0 valued for non-secretory cells, and the latter as a group are far more numerous than any individual subgroup of secretory cells. So they would completely overwhelm all other bars/values in the histogram. Therefore, we remove the bucket corresponding to 0 value from our input. In any actual practice, we would need to separate out secretory and non-secretory cells before we ran an algorithm like ours.

We provide further details about our experimental set-up, especially the non-linear functions, corresponding to custom probes, used in reconstructing the data in the appendix.

Remark 3.1. In Figure 4, we plot the reconstruction error as a function of sample size (m). It is worth noting that 20 samples were enough for Cpe reconstruction to converge. This suggests that the Wasserstein distance between recovered and original distributions can behave much better than the ℓ_1 distance, used in standard compressed sensing, which grows linearly in m (cf. Lemma 2.3).

4. Discussion

This work proposes an algorithm to approximately reconstruct single cell distribution from a small number of aggregate measurements. On the algorithmic side, we propose an algorithm. A full version of this paper, submitted elsewhere, extends current compressed sensing results to show formal bounds for its performance and correctness. We use simulations using publicly available scRNA-seq data from (Bastidas-Ponce et al., 2019) as the ground truth, to compare the original and approximately reconstructed distributions, where the latter is computed using our algorithm. Due to limited resources we could not verify the simulations in actual experiments using custom hybridization probes.

References

Bastidas-Ponce, A., Sophie Tritschler, L. D., Scheibner, K., Tarquis-Medina, M., Salinno, C., Schirge, S., Burtscher, I., Böttcher, A., Theis, F. J., Lickert, H., and Bakht, M. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, 146(12):dev173849, 2019.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38: 1408–1414, 2020.

Dembo, A. and Zeitouni, O. *Large Deviations Techniques and Applications*. Springer Verlag, New York, 1998. ISBN 0-387-98406-2.

Haroon, M. F., Skennerton, C. T., Steen, J. A., Lachner, N., Hugenholtz, P., and Tyson, G. W. Chapter One - In-Solution Fluorescence In Situ Hybridization and Fluorescence-Activated Cell Sorting for Single Cell and Population Genome Recovery. In DeLong, E. F. (ed.), *Microbial Metagenomics, Metatranscriptomics, and Metaproteomics*, volume 531 of *Methods in Enzymology*, pp. 3–19. Academic Press, 2013. doi: https://doi.org/10.1016/B978-0-12-407863-5.00001-0. URL https://www.sciencedirect.com/science/article/pii/B9780124078635000010.

Hu, L., Ru, K., Zhang, L., Huang, Y., Zhu, X., Liu, H., Zetterberg, A., Cheng, T., and Miao, W. Fluorescence in situ hybridization (FISH): an increasingly demanded tool for biomarker research and personalized medicine. *Biomarker Research*, 2(3), 2014.

Raskutti, G., Wainwright, M. J., and Yu, B. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(78):2241–2259, 2010. URL http://jmlr.org/papers/v11/raskutti10a.html.

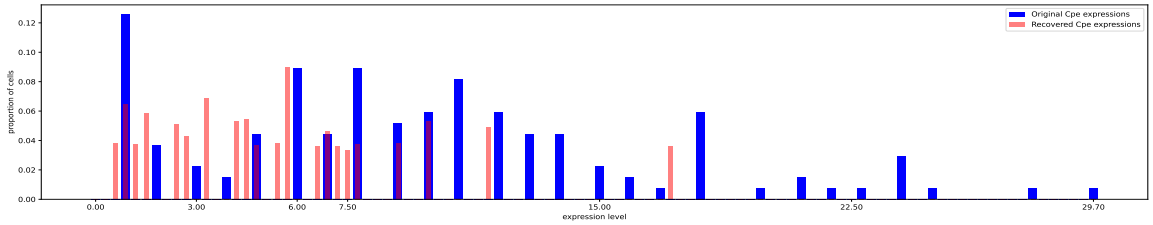

Figure 2. Recovered and original distributions using ℓ_1 minimization algorithm with a better optimized experiment design for Cpe ($N = 250, n = 100, m = 10, s = 20$).

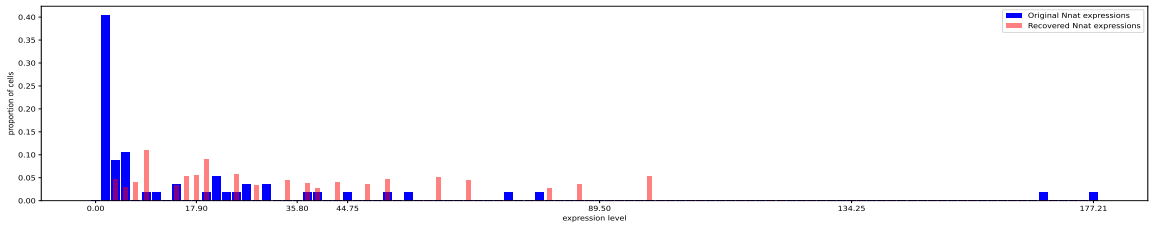

Figure 3. Recovered and original distributions using ℓ_1 minimization algorithm with a better optimized experiment design for Nnat ($N = 250, n = 100, m = 10, s = 20$).

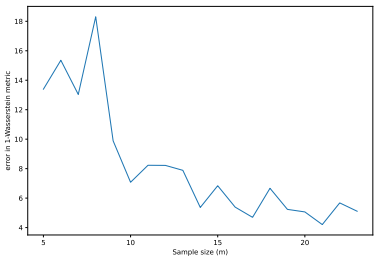

Figure 4. Error (measured by 1-Wasserstein distance between reconstructed and original distributions) as a function of sample size (m). Based on the trends, the error for Cpe decreases with m.

A. Permissible probes

Suppose we can design probes that hybridize with two RNA strands as opposed to one, then the concentration of "doubly" hybridized product formed is $\sum_{i\in[N]} c_i^2$. Therefore, the mean and variance are proportional to $\sum_{i\in[N]} c_i^2$. So we can essentially measure quadratic functions of cellular concentrations using appropriate hybridization probes. We say $f(x) = x^2$ is a *permissible function* and ensembles of the form:[4]

$$\begin{pmatrix} N(0,k'M_1^2) & N(0,k'(M_1+1)^2) & N(0,k'(M_1+2)^2) & \dots \\ \vdots & \vdots & \ddots & \\ N(0,k'M_1^2) & N(0,k'(M_1+1)^2) & & N(0,k'M_2^2) \end{pmatrix}$$

lead to a *permissible measurement matrix* corresponding to the *permissible covariance matrix*: $\mathrm{Diag}(k'M_1^2, ..., k'M_2^2)$.[5]

Similarly we can measure third moments also. It is likely that we can't compute higher moments than three, as the chances of any meaningful amount of hybridization would be too small. On the other hand, we can also design probes that bind to our target RNA T or a background RNA B, but not both. If we know that the levels of B are constant in most cells then we can effectively measure using functions of the form $\frac{a}{b+c_T}$, where c_T is the concentration of our target RNA in the cell compartment, and a and b are constants depending on the background RNA used. They correspond to covariance matrices of the form $\mathrm{Diag}(\frac{a}{b+M_1}, \frac{a}{b+M_1+1}, ..., \frac{a}{b+M_2})$.

Therefore, it is likely that a rich set of "low degree" probes can be constructed. In our simulation, we used the following class of measurement functions:

$$F := \{x, x^2, \frac{a}{b+x}, \frac{a}{(b+x)(c+x)}, \frac{a'}{b'+x}, x^3\}.$$

Let F_i denote the i^{th} element of F. Of course, nothing prevents us from repeating the same measurement twice, for example, F_1, F_1, F_1, F_1, F_1. But it will have diminishing returns as the smallest positive singular value will be smaller as well. In particular, for our experiments with $m = 10$, we used the sequence of measurements: $F_1, F_2, ..., F_6, F_1, F_2, .., F4$.

[4] We have assumed bins of width 1 for simplicity.

[5] Notation: Diag denotes the diagonal matrix with the specified diagonal.