Recovering a sparse linear dynamical system

Pratik Worah [1]

Abstract

We design a ℓ_1 optimization based sparse recovery algorithm to recover a linear system of differential equations from snapshots of the data at few different points in time, when the number of non-zero coefficients is known to be sparse, and also prove an upper-bound on the number of samples required for unique recovery. Typically, dynamical systems governing gene expressions are naturally sparse – a gene can be directly controlled by only a handful of genes. Therefore, as an application, we recover the coefficients of a part of the underlying dynamical system relevant to the expression of the gene Carboxypeptidase E (Cpe) from scRNA-seq data in murine pancreatic tissue.

1. Introduction

The problem of ascertaining which gene directly influences which other gene, and to what degree, is a fundamental problem in genomics. From a mathematical perspective, it is equivalent to computing the dynamical system underlying gene expression mechanisms. In this note, we define and study one concrete version of the problem, where we work under the assumption that one gene can be directly influenced by only a small number of other genes.

Mathematically, the problem we study in this note can be succinctly stated as follows. Given a high-dimensional noisy linear dynamical system represented as a multi-variate Ornstein-Uhlenbeck (OU) process that has a sparse[1] speed matrix[2], i.e., any row of the speed matrix has few non-zero entries, we want to recover the coefficients of the speed matrix from a small number of samples of the position and velocity vectors of the dynamical system.

A natural way to solve the above problem is via compressed sensing algorithms, for example, basis pursuit (equivalently ℓ_1 minimization). However, for such algorithms to work the null space property or one of its stronger cousins has to hold true. Typically, that requires a condition on the covariance matrix when the distribution of each entry in the measurement matrix is a Gaussian (see for example (Raskutti et al., 2010) for the most general of such conditions). The condition in (Raskutti et al., 2010) assumes that the rows of the sensing/measurement matrix are independent, which is not the case for the matrices derived from noisy linear *dynamical systems*. That said, the linear nature of our dynamical system still allows us to extend their theorems, which rely on Gordon's inequality, to the case of mildly dependent rows. Note that we do not prove any theorems in this brief note, and our main theoretical contribution is an overview in Section 3 with remarks (Remarks 3.1 and 3.2) that summarize our main ideas. They will be proven in a full version of the paper. As an application, we use single cell RNA-seq data from (Bastidas-Ponce et al., 2019) and RNA velocity package scvelo[3] from (Bergen et al., 2020) to recover the speed matrix using the well-known basis pursuit algorithm (see Figure 1). Section 4 contains some details about our experimental setup.

2. Compressed sensing: Basics

A basic problem in the field of compressed sensing is to reconstruct an unknown sparse[4] signal $\phi \in \mathbb{R}^n$ using data from a (small) sample of measurements. These measurements are represented by a matrix $X \in \mathbb{R}^{m \times n}$, where m is the number of samples. Typically, $m \simeq \log^{O(1)} n$. Moreover, the m sample observations y are related to ϕ as: $X\phi = y$ or in the case of noisy measurements: $\|X\phi - y\|_2 \leq \eta$. The matrix X is assumed to be known.

The central problem in compressed sensing is to reconstruct (or recover) the s-sparse unknown signal ϕ (for $s \ll n$) from y, while keeping $m \ll n$ (see (Foucart & Rauhut, 2013)). The crux of the solution lies in solving a convex

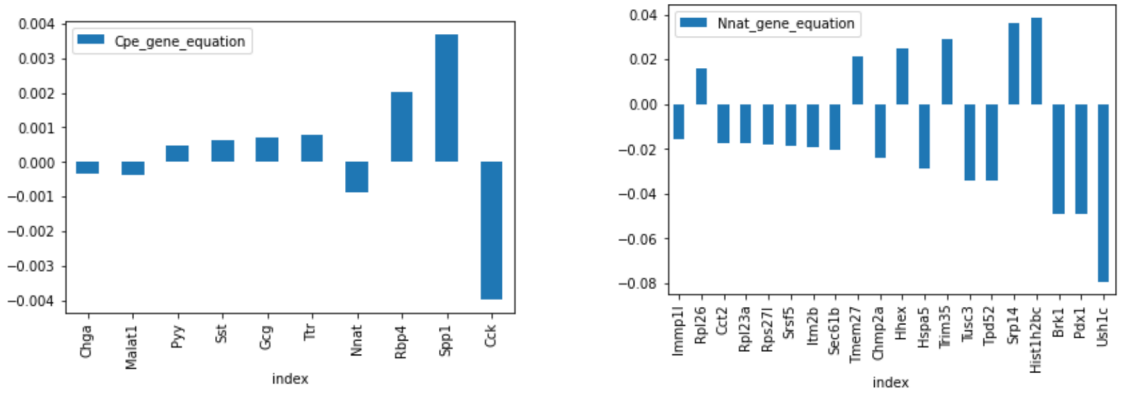
[1]Google Research, USA. Correspondence to: Pratik Worah <pworah@google.com>.

[1]Here sparse means poly-logarithmic in the dimension of the system.

[2]See Equation 3 for a definition.

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

[4]A n-dimensional real vector is said to be s-sparse if it is non-zero on at most s co-ordinates.



(a) Recovered Φ specifying regulatory genes for Cpe

(b) Recovered Φ specifying regulatory genes for Nnat

Figure 1. Regulatory genes for Cpe and Nnat. The x-axis denotes the gene name, the y-axis denotes the (negative) influence it has on the expression of the gene: (a) Cpe and (b) Nnat. Note y-axis signs are inverted as there is a negative sign in front of the speed matrix Φ in Equation 3.

program (CP) of the form:

$$\min_{x} \quad \|\phi\|_1 \qquad (1)$$
$$s.t. \quad \|X\phi - y\|_2 \leq \eta.$$

Definition 2.1. A matrix X has the (s, γ)-null space property (NSP) if

$$\forall v \in \mathrm{Ker}(X) \setminus \{0\},\ \forall |T| \leq s, \|v_T\|_1 \leq \gamma \|v_{T^c}\|_1.$$

While the number of fundamental results in compressed sensing are too numerous to list here, we point the reader who is new to the area to the book (Foucart & Rauhut, 2013). Below we state a canonical theorem from compressed sensing literature.

Theorem 2.2. *(Chen, 2012), (Cahill et al., 2016) Let $\mathcal{S}$ be the set of s-sparse vectors in $\mathbb{R}^n$. If X has (s, γ)-NSP, then any minimizer $\hat{x}$ of CP in Equation 1 satisfies*

$$\|\hat{\phi} - \phi\|_1 \leq \frac{4\sqrt{2}\sqrt{N}}{(1-\gamma)\lambda(X)}\eta + \frac{4(1+\gamma)}{\sqrt{2}(1-\gamma)}\sigma_s(\phi), \qquad (2)$$

where x is the original signal to be recovered, λ is the smallest positive singular value of X and $\sigma_s(x) := \min_{z \in \mathcal{S}} \|x - z\|_1$.

Definition 2.3. (Raskutti et al., 2010) Suppose the rows of a $m \times n$ matrix X are chosen according to independent multivariate normal distributions with $n \times n$ covariance matrix Υ. Then X is said to follow a (s, α, γ)-restricted eigenvalue (RE) condition of order γ, if $\|\Upsilon^{1/2} v\|_2 \geq \alpha \|v\|_2$, for all v that satisfy a (s, γ)-NSP with respect to X.

Note that a RE condition of order (s, α, γ), for any constant α, implies a (s, γ)-NSP. Hence the former is stronger than the latter.

3. Overview of our results

Let $x(t) \in \mathbb{R}^n$ denote the position vector of our dynamical system in question. Let m denote the number of our snapshots of $x(t)$, taken at times $\{t_1, t_2, ..., t_m\}$, and let $\Phi \in \mathbb{R}^{n \times n}$ denote the speed matrix for the underlying OU process, i.e.,

$$dx(t) = \Phi(\mu - x(t))dt + \sigma dW_t, \qquad (3)$$

where $\mu \in \mathbb{R}^n$ is the long term mean expression level of each gene and $\sigma \in \mathbb{R}^{n \times n}$ denotes the fluctuation in measurements, $W_t \in \mathbb{R}^n$ denotes Weiner noise. For theoretical purposes, we will assume $\mu \equiv \vec{0}$, as the process can be centered, if necessary. Note that for our running example for gene expression levels which can not assume negative values, $x(t)$ can be chosen to be the log of the expression levels, in case the diffusivity σ is significant compared to the mean.

The measurement matrix X is obtained by concatenating the snapshot vectors $[x(t_1), x(t_2), ..., x(t_m)]$. We assume the time derivatives $\dot{x}(t)$ can be measured as well, these are our observations (y in Equation 1). In the gene expression example, they are computed from steady-state RNA velocity data (Bergen et al., 2020), (Bastidas-Ponce et al., 2019).

Note that $x(t)$, and hence the measurement matrix, is Gaussian, but the rows are not independent, unlike the usual assumption in compressed sensing papers (see for example (Raskutti et al., 2010), (Foucart & Rauhut, 2013)). When rows are independent Gaussians, the paper (Raskutti et al., 2010) shows that the required number of snapshots for the recovery of any row of Φ is small (around $O(\log n)$), under the RE condition (see Definition 2.3). This leads to the basic compressed sensing question: how many samples are enough for our case, with a Gaussian measurement matrix

coming from an OU process, as above?

The correlations between any two rows of our measurement matrix decay rapidly with time as long as the real part of the eigenvalues of the recovered matrices is negative and not too small. We state our main result as a remark, without proof, in this note. Since we do not provide full proofs, our results are stated as remarks, instead of theorems.

Remark 3.1. Let $t_1, ..., t_m$ be the snapshot times of the OU process in Equation 3, such that $|t_i - t_j| \geq C$, for a large enough constant C, for all i and j in $[m]$. For large t, positive definite $\sigma\sigma^T$, and measurement matrix $X \in \mathbb{R}^{m \times n}$ (as defined above) with $N(0, \Upsilon)$ rows, there exist constants K, c, c' such that:

$$\frac{\|Xv\|_2}{\sqrt{m}} \geq K\|\Upsilon^{1/2}v\|_2 - \rho(X)\sqrt{\frac{\log n}{m}}\|v\|_1, \qquad (4)$$

with probability at least $1 - c' \exp(-cm)$, where $\rho(X) := \max_{i \in [n]} X_{i,i}$ and K depends upon C.

As in (Raskutti et al., 2010) this implies a condition stronger than the null space property and ensures recovery with high confidence. The proof relies on generalizing the first part of the proof in (Raskutti et al., 2010) that uses Gordon's inequality and may be interesting in itself. Some details are sketched later in this section.

However, if we assume C is very large in Remark 3.1 then the rows of the measurement matrix X are independent, and (Raskutti et al., 2010) implies successful recovery using basis pursuit, as long as Υ satisfies the RE condition. However, in our case Υ depends upon Φ! Thus even in this "independent" setting, we still need to derive sufficient conditions on Φ which can be inspected *after recovery* to have high confidence that the covariance matrix Υ satisfies the RE condition of order (s, α, γ), for some α and γ. Only after that verification, based on the values of α and γ, one can say that the number of snapshots used, i.e., m, was large enough for successful recovery, with high probability. This motivates the following remark.

Remark 3.2. Assuming $\forall i, j \in [m] : |t_i - t_j| \to \infty$, then Υ satisfies RE condition for some positive constants (s, α, γ) if:

1. $\sigma\sigma^T$ is positive definite with minimum eigenvalue lower bounded by a positive constant, or
2. $\{\Phi\sigma\sigma^T, \Phi^2\sigma\sigma^T, ..., \Phi^k\sigma\sigma^T\}$ spans $\mathbb{R}^n$ for some finite k.

Note that the coefficient α although positive, may go to 0 as $t_i \to \infty$.

The proof of Remark 3.2, although not involved by itself, relies on Hörmander's condition (see (Hörmander, 1967)) and requires some background in Lie algebra and differential equations. We skip the proof in this note.

It is worth noting that the number of samples required for successful recovery grows as $O(\frac{\log n}{\alpha^2})$, see for example (Raskutti et al., 2010). Therefore, characterizing the constant α is an important problem when $\sigma\sigma^T$ is not positive definite, and corresponds to the stochastic process version of the minimum positive singular value in Chen's bound for the deterministic setting (Theorem 2.2).

Next, we briefly sketch the modification involved in generalizing the proof in (Raskutti et al., 2010) to incorporate auto-correlations from our OU process setting. The proof of the main theorem in (Raskutti et al., 2010), which inspired Remark 3.1, proceeds in three parts: (1) An upper-bound on $\mathbb{E}[M(r, \Phi)]$, where $M(r, \Phi) := 1 - \inf_{v \in V(r)} \frac{\|\Phi v\|_2}{\sqrt{m}}$,[5] (2) a sharp concentration result for $M(r, \Phi)$ around its expectation, and (3) a peeling argument to show the final high probability statement holds. For our purposes, we only need to modify (1), the crux of which relies on Gordon's inequality.

The point of the using Gordon's inequality is to upper bound the Euclidean norm of a linear combination of Gaussians using the supremum of a linear combination of a different set of Gaussians. Gordon's inequality allows us to upper bound the expectation of the former with that of the latter, as long as the total variance of the latter is greater than that of the former. The crux of the entire method is verifying this relationship between the total variance of the two sets of Gaussians. See the book (Foucart & Rauhut, 2013) for more details about Gordon's inequality.

In our case, the Gaussians are correlated, so calculating the variances requires us to keep track of $O(mn)$ covariance matrices as well. This may seem daunting, but it is easily accomplished by observing that the cross-covariance matrix $\Gamma(t_i, t_j)$ between the n-dimensional Gaussians $x(t_i)$ and $x(t_j)$ with $t_i - t_j = C$ are given by:

$$\Gamma(t_i, t_j) = e^{-C\Phi} \int_0^{t_j} e^{-t\Phi}\sigma\sigma^T e^{-t\Phi^T} dt. \qquad (5)$$

For large t_j, the integral in Equation 5, denoted by Γ', where $\Gamma = e^{-C\Phi}\Gamma'$, can be written as:

$$\Phi\Gamma' + \Gamma'\Phi^T = \sigma\sigma^T/2. \qquad (6)$$

See the results in (Vatiwutipong & Phewchean, 2019) for a proof of the above. Therefore, the mn covariance matrices are just a matrix exponential times a common covariance matrix (given by Γ'). This makes the total variance calculation for Gordon's inequality tractable. We defer the details to the full version of the paper.

[5] Here $V(r) := \{v \in \mathbb{R}^n : \|\Upsilon^{1/2}v\|_2 = 1, \|v\|_1 \leq r\}$.

Finally, we remark that a potential generalization, that combines the Remarks 3.1 and 3.2, to the case when C is finite and $\sigma\sigma^T$ is not necessarily positive definite would be interesting. The proof would need to combine the Hörmander condition based ideas in the independent case and take into account the auto-correlations and build over the proof in (Raskutti et al., 2010).

4. Further details: scRNA-seq data

In this section, we provide some details about how we obtain the graphs in Figure 1, based on the scRNA-seq data in (Bastidas-Ponce et al., 2019) (also see (Bergen et al., 2020)) for about 4000 cells murine pancreatic tissue. The vectors $x(t)$ are merely the total RNA expression levels for a cell as sampled from the data. Each cell is assumed to provide a (noisy) snapshot of the underlying dynamical system. For the velocity, we use the definition of RNA velocity as measured by the difference in spliced and unspliced RNA levels in (Bergen et al., 2020) to approximate $\dot{x}(t)$. Finally, we use the basis pursuit algorithm (see (Foucart & Rauhut, 2013)) to infer the speed matrix Φ in Equation 1 for the genes Cpe and Nnat, by solving the second order cone program (using cvx). Figure 1 is the result of application of basis pursuit algorithm for the rows corresponding to the genes Cpe and Nnat.

Note that, since linear systems of differential equations can encode for higher order derivatives, the notion of RNA acceleration and other higher order terms are implicit in our dynamical system.

Also note that, for computational efficiency, we remove low count genes from the data using scvelo, the recovered coefficients in Figure 1 may leave out some intermediate genes. So for the purposes of this discussion "directly influence" should be understood as confined to the universe of unfiltered genes.

We can use Figure 1 to make the following initial observations:

1. In Figure 1, Cck, the gene directly responsible for secreting the peptide hormone Cholecystokinin seems to have a positive effect on Cpe, which is required for maturation of Cholecystokinin, as explained in this endocrinology encyclopedia entry. So this is expected. That Spp1 (Osteopontin) negatively effects Cpe in the pancreas could be an interesting corollary that needs experimental verification. Osteopontin has been known to positively correlate with the diabetic state (Cai et al., 2018).

2. This naturally leads to the question: how do we know that we have enough data samples vs the number of genes to have uniquely and accurately recovered the interactome (dynamical system)? The answer lies in the upper-bounds, sketched in the theoretical overview in Section 3.

5. Conclusion

In this note, we summarized an application of compressed sensing algorithms to recover genes which directly effect the expression of a given gene. More generally, the idea can be used to recover the coefficients of the equations of any linear dynamical system which satisfies the sparsity assumption. We sketched theoretical bounds on the sample size that can allow us to have high degree of confidence on the recovered dynamical system coefficients.

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.

Cahill, J., Chen, X., and Wang, R. The gap between null space property and the restricted isometry property. *arxiv:1506.03040v2*, 2016.

Cai, M., Bompada, P., Salehi, A., R.Acosta, J., B.Prasad, R., Ataca, D., Groopa, M. L. L., and Marinis, Y. D. Role of osteopontin and its regulation in pancreatic islet. *Biochemical and Biophysical Research Communications*, 495(1): 1425–1431, 2018.

Chen, X. Stability of compressed sensing for dictionaries and almost sure convergence rate for the kacmarz algorithm. *Diss. Vanderbilt University*, 2012.

Foucart, S. and Rauhut, H. A mathematical introduction to compressive sensing. *Birkhäuser New York, NY*, 2296-5009:625, 2013.

Hörmander, L. Hypoelliptic second order differential equations. *Acta Mathematica*, 119, 1967.

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.

Vatiwutipong, P. and Phewchean, N. Alternative way to derive the distribution of the multivariate Ornstein–Uhlenbeck process. *Advances in Difference Equations*, (276), 2019.