Tensor Restricted Isometry Property for Multilinear Sparse System of Genomic Interactions

Alexandra Fry   Carmeliza Navasca
alifry@uab.edu   cnavasca@uab.edu

Department of Mathematics
University of Alabama at Birmingham
Birmingham, Alabama 35294-1170

January 3, 2015

1 Introduction

Research scientists in the biomedical sciences perform large scale experiments referred to as high throughput research. High throughput methods enable research involving large data (e.g. the human genome) to investigate the effects of individual genes and to eventually answer fundamental questions. These large scale experiments allow scientists to investigate the effects of individual genes, i.e. identifying genes involved in a specific function. By studying each single-deletion strain, biologists are able to deduce the relevance of a particular gene to a specific function or functions. However, studies with single gene suppression at a time have shown a weak correlation between the genes and the functions due to redundancy in the genome [5, 4]; i.e. a similar gene or backup copies can perform the function of the deleted one [5].

Moving beyond pairwise interactions and to organisms with larger genomes is challenging. We model pairwise and high-order gene interactions through multilinear systems. Since sparsity is an important characteristic of the models structure [5], which means relatively few of all conceivable interactions of genes are significant, we build from the existing theory from compressed sensing [2, 3]. Compressed sensing efficiently reconstructs signals by finding solutions to underdetermined linear systems. It takes advantage of the signal's compressibility, allowing the entire signal to be determined from relatively few measurements. Our contribution is to extend the application of compressed sensing in gene interaction analysis within the tensor paradigm.

Multiple interactions with up to D genes are modeled as multilinear equations which can be structured as a high-order tensor equations [1]. Informally, tensors are multi-dimensional arrays; the order of a tensor is the number modal directions. A vector is first-order tensor, a matrix is a second-order tensor. Multilinear systems are generalizations of linear systems [1]. A matrix defines a linear transformation such that $L(x) = \mathbf{A}x$ as would a tensor define a multilinear transformation.

2 Tensor Equation for Multiple Gene Interactions

Genomic interactions are modeled as a sparse system [5]. We develop a tensor representation of the underlying sparse multilinear system for the genomic data. One knockdown inhibits one gene in the pathway; two knockdowns (pairwise) inhibit two genes. Let a be the genome vector with a nonzero entry a_i be the knockdown of the ith gene. If M experiments are performed, there will be M distinct one gene knockdown. Thus, a linear system representation can be set-up to detect the relevance of these knockdowns; i.e. $y_k^{(1)} = \sum_i^N a_{ki}x_i$ where a is a row vector with one nonzero value on the ith entry, $y^{(1)} \in \mathbb{R}^M$ with entries signifying the amount of viral replication with one gene knockdown and x is the unknown sparse vector with nonzero entries reflect the relevance of the gene. The one knockdown set-up may not be sufficient since other genes may act like a proxy in the pathway for the blocked gene. For the pairwise interaction of inhibited genes, the multilinear case is defined as $y_k^{(2)} = \sum_{i<j} a_{ki}a_{kj}x_{ij}$ where a_{ki} is the ith knockdown gene in the kth observation and $y^{(2)} \in \mathbb{R}^M$ with entries signifying viral replication with two-gene knockdowns. For D gene interactions is the following equation: $y_k^{(D)} = \sum_{i<j<\cdots<D}^N a_{ki}a_{kj}\ldots a_{kD}x_{ij\ldots D}$. Thus, the equation for all interactions up to D is

$$y_k^{(1,2,\ldots,D)} = \sum_i a_{ki}x_i + \sum_{i<j} a_{ki}a_{kj}x_{ij} \tag{1}$$
$$+ \ldots + \sum_{i_1<i_2<\cdots<i_D} a_{ki_1}a_{ki_2}\ldots a_{ki_D}x_{i_1i_2\ldots i_D}$$

where x_i is a vector entry in $x \in \mathbb{R}^N$, x_{ij} is a matrix entry in $\mathbf{X} \in \mathbb{R}^{N\times N}$, and $x_{i_1i_2\ldots i_D}$ is a tensor entry in $\mathcal{X} \in \mathbb{R}^{\underbrace{N\times\cdots\times N}_{\mathrm{D}}}$.

We denote the scalars in $\mathbb{R}$ with lower-case letters $(a, b, \ldots)$ and the vectors with bold lower-case letters $(\mathbf{a}, \mathbf{b}, \ldots)$. The matrices are written as bold upper-case letters $(\mathbf{A}, \mathbf{B}, \ldots)$ and the symbol for tensors are calligraphic letters $(\mathcal{A}, \mathcal{B}, \ldots)$. The subscripts represent the following scalars: $(\mathcal{A})_{ijk} = a_{ijk}$, $(\mathbf{A})_{ij} = a_{ij}$, $(\mathbf{a})_i = a_i$, unless noted otherwise. The superscripts indicate the length of the vector or the size of the matrices. For example, $\mathbf{b}^K$ is a vector with length K and $\mathbf{B}^{N\times K}$ is a $N \times K$ matrix.

2.1 Tensor Equation via Embedding

We reformulate (1) by *embedding* the one-knockdown system into the pairwise interaction system, and then the pairwise equation into the higher order interactions. Recall that the one-interaction is

$$y_k^{(1)} = \sum_i^N a_{ki} x_i = \mathbf{A}x \tag{2}$$

whereas the two-interaction model is $y_k^{(2)} = \sum_{i<j} a_{ki} a_{kj} x_{ij}$. The pairwise model can be formulated as

$$\mathcal{A} * \mathbf{X} = y^{(2)} \tag{3}$$

where $(\mathcal{A} * \mathbf{X})_k = \sum_{ij}^N \mathcal{A}_{kij} \mathbf{X}_{ij} = y_k^{(2)}$ with $\mathcal{A}_{kij} = a_{ki} a_{kj}$. The dimensions are: $\mathcal{A}$ is a third-order tensor of size $M \times N \times N$, $\mathbf{X}$ is a $N \times N$ matrix and $y^{(2)}$ is a vector of size M.

Incorporating (2) into (3), we obtain

$$y_k^{(1,2)} = \sum_{i,j}^{N+1,N} \mathcal{A}_{kij} \mathbf{X}_{ij} = (\mathcal{A} * \mathbf{X})_k \tag{4}$$

(see Figures 1-2) where $\mathcal{A}_{kij} = a_{ki} a_{kj}$. Henceforth, $\mathcal{A}$ is a tensor of size $M \times N+1 \times N$, $\mathbf{X}$ is a matrix of size $N+1 \times N$ and y_k is a vector of size M. The generalization of the multilinear system up to order D is $y_k^{(1,\ldots,D)} = \sum_{i_1, j_2, \ldots, i_D} \mathcal{A}_{i_1 j_1 \ldots i_D k} \mathbf{X}_{i_1 i_2 \ldots i_D}$ where $\mathcal{A}$ and $\mathbf{X}$ are $(D+1)$th and Dth order tensors.

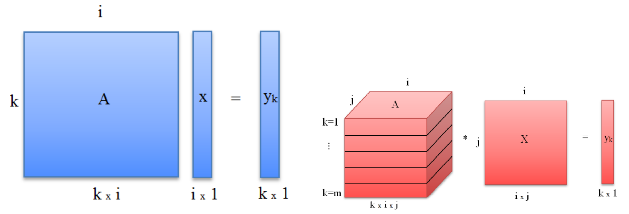

Figure 1: (Multi)linear system for one gene knockdown (left) and two-gene knockdowns (right).

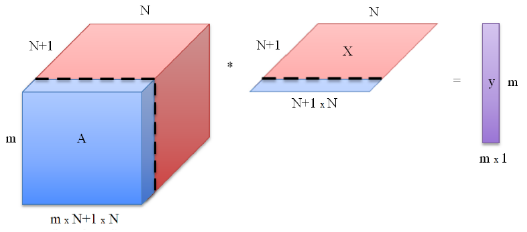

Figure 2: Tensor model for up to two-gene interactions

2.2 Tensor Restricted Isometry Property

Let $\mathbf{X} \in \mathbb{R}^{n\times m}$. The Frobenius norm is defined as $\|\mathbf{X}\|_F = (\Sigma_{i,j}^{n,m} x_{ij}^2)^{\frac{1}{2}}$. The l_1 norm of a vector $v \in \mathbb{R}^n$ is $\|v\|_1 = \Sigma_{i=1}^n |v_i|$, while the l_2 norm is $\|v\|_2 = (\Sigma_{i=1}^n {v_i}^2)^{\frac{1}{2}}$. The vectorization map is defined as $vec(\mathbf{W} \in \mathbb{R}^{N\times M}) \rightarrow \mathbf{w} \in \mathbb{R}^{N\cdot M}$ where $\mathbf{W}_{ij} \rightarrow \mathbf{w}_k$ with $k = (i-1) * N + j$ for $i = 1, \ldots, N$ and $j = 1, \ldots, M$.

We will provide the Tensor Restricted Isometry Property (TRIP) for the necessary conditions for a sparse and unique solution to (3). We define the nullspace of $\mathcal{A}$ as $\mathcal{N}(\mathcal{A}) = \{\mathbf{X} \in \mathbb{R}^{N+1\times N} | \mathcal{A} * \mathbf{X} = 0\}$. We denote $\Sigma_s \subseteq \mathbb{R}^n$ s.t. $\Sigma_s = \{x \in \mathbb{R}^n | s << n$ at most nonzero elements$\}$. We say $x \in \Sigma_s$ is s-sparse.

Lemma 2.1 *Assume $\Sigma_{2s} \cap N(\mathcal{A}) = \{0\}$. Then the multilinear system $\mathcal{A} * \mathbf{X} = y$ (3) has a unique s-sparse solution.*

Proof: Assume $\mathbf{X} \in \Sigma_s$ for some s. Suppose there are two different s-sparse solutions $\mathbf{X}_1$, $\mathbf{X}_2 \in \Sigma_s$, i.e. $\mathcal{A} * \mathbf{X}_1 = y$, $\mathcal{A} * \mathbf{X}_2 = y$. Note $\mathbf{X}_1 \neq \mathbf{X}_2$ implies $\mathbf{X}_1 - \mathbf{X}_2 \neq 0$. Then, $\mathcal{A} * (\mathbf{X}_1 - \mathbf{X}_2) = \mathcal{A} * \mathbf{X}_1 - \mathcal{A} * \mathbf{X}_2 = y - y = 0$ implies $\mathbf{X}_1 - \mathbf{X}_2 \in \mathcal{N}(\mathcal{A})$. Also, $\mathbf{X}_1 - \mathbf{X}_2 \in \Sigma_{2s}$. So, $\mathbf{X}_1 - \mathbf{X}_2 \in \mathcal{N}(\mathcal{A}) \cap \Sigma_{2s}$. But $\mathcal{N}(\mathcal{A}) \cap \Sigma_{2s} = \{0\}$. Therefore, $\mathbf{X}_1 - \mathbf{X}_2 = 0$, i.e. $\mathbf{X}_1 = \mathbf{X}_2$ which is a contradiction. □

Remark 2.1 *From the given assumption that $\Sigma_{2s} \cap N(\mathcal{A}) = \{0\}$, all nonzero elements in the null space of tensor $\mathcal{A}$ have at least $2s+1$ nonzero components. If $\mathcal{A} * \mathbf{X} = y$ has more than one s-sparse solution, then $\mathcal{N}(\mathcal{A})$ must contain a nonzero $2s$-sparse matrix.*

Lemma 2.2 *Suppose there exists positive constants c_1 and c_2 such that $c_1 \leq \|\mathcal{A} * \mathbf{U}\|_2^2 \leq c_2$ for every $\mathbf{X} \in \Sigma_{2s}$, where we define $\mathbf{U} = \frac{\mathbf{X}}{\|\mathbf{X}\|_F}$. Then $\Sigma_{2s} \cap N(\mathcal{A}) = 0$.*

Note that $\|\mathbf{U}\|_F^2 = 1$. If there exists a positive c_1 such that $c_1 \leq \|\mathcal{A} * \mathbf{U}\|_2^2$ for all $\mathbf{U} \in \Sigma_{2s}$. It follows that no 2s-sparse matrix with norm one is in the nullspace of $\mathcal{A}$. Moreover,

$$\max_{\mathbf{U}} \|\mathcal{A} * \mathbf{U}\|_2^2 = \max_{vec(\mathbf{U})} \|\mathbf{A}_{mat} \cdot vec(\mathbf{U})\|_2^2$$

where $\mathbf{A}_{mat} \in \mathbb{R}^{M\times(N+1)\cdot N}$ is the matrix unfolding of $\mathcal{A}$. A maximum of $\|\mathbf{A}_{mat}\cdot vec(\mathbf{U})\|_2^2$ is attained since $vec(\mathbf{U}) \to \|\mathbf{A}_{mat}\cdot vec(\mathbf{U})\|_2^2$ is continuous over a compact subset of 2s-sparse vectors in $\mathbb{R}^{(N+1)\cdot N}$.

Theorem 2.1 *If $\mathcal{A} \in \mathbb{R}^{M\times N+1\times N}$ and $\mathbf{X} \in \mathbb{R}^{N+1\times N}$ satisfies the Tensor Restricted Isometry Property (TRIP):*

$$(1-\delta_s)\|\mathbf{X}\|_F^2 \le \|\mathcal{A} * \mathbf{X}\|_2^2 \le (1+\delta_s)\|\mathbf{X}\|_F^2 \tag{5}$$

for $\delta_s \in (0,1)$ and some $s \ge 1$, then any s-sparse solution to $\mathcal{A} * \mathbf{X} = y$ is unique.

Proof: It follows from Lemma 2.2 that

$$c_1 \le \|\mathcal{A} * \mathbf{U}\|_2^2 \le c_2.$$

If we take $c_1 = 1-\delta_s$ and $c_2 = 1+\delta_s$ for $\delta_s \in (0,1)$, then

$$(1-\delta_s) \le \|\mathcal{A} * \mathbf{X}\|_2^2 * \frac{1}{\|\mathbf{X}\|_F^2} \le (1+\delta_s).$$

Now from Lemma 2.2 and Lemma 2.1, we obtain the uniqueness of the s-sparse solution of $\mathcal{A} * \mathbf{U} = y$. □

It is well-known from compressed sensing [3, 2] that ℓ_1 minimization recovers the sparse signals. For our case, the optimal solution $\mathbf{X}^*$ is the unique solution to the following optimization problem,

$$\min \|vec(\mathbf{X})\|_{\ell_1} \text{ subject to } \mathbf{A}_{mat}\cdot vec(\mathbf{X}) = y. \tag{6}$$

Define $f(t) = \|vec(\mathbf{X}^* + t\mathbf{N})\|_{\ell_1}$ where $\mathbf{N} \in \mathcal{N}(\mathcal{A})$. Suppose $\mathcal{A} * \mathbf{X}^{**} = y$ where $\mathbf{X}^{**} \neq \mathbf{X}^*$. Take $\mathbf{N} = \mathbf{X}^{**} - \mathbf{X}^* \neq 0$ and $\mathbf{N} \in \mathcal{N}(\mathcal{A})$. Then $vec(\mathbf{X}^{**}) = f(1) > f(0) = vec(\mathbf{X}^*)$.

Theorem 2.2 *If $f(t)$ has a unique global minimum at $t = 0$ for nonzero $\mathbf{N} \in \mathcal{N}(\mathcal{A})$, then $\mathbf{X}^*$ is the unique solution of (6).*

Proof: WLOG, assume $vec(\mathbf{X}^*) = (x_1^*, 0, \dots, 0)$ and fix any nonzero $vec(\mathbf{N}) = (n_1, \cdots, n_{(N+1)\cdot N})$. The function

$$\begin{aligned} f(t) &= \|vec(\mathbf{X}^* + t\mathbf{N})\|_{\ell_1} && (7)\\ &= \sum_{i=1}^{(N+1)\cdot N} |x_i^* + tn_i| && (8)\\ &= |x_1^* + tn_1| + |t| \sum_{i=2}^{(N+1)\cdot N} |n_i| && (9) \end{aligned}$$

has critical numbers at $t = 0$ and $t = -\frac{x_1^*}{n_1}$. Now, our goal is to show that $f(0) < f(-\frac{x_1^*}{n_1})$ which is equivalent to

$$|n_1| < \sum_{i=2}^{(N+1)\cdot N} |n_i| \tag{10}$$

We use the re-ordering trick [3, 2] of the indices of $vec(\mathbf{N})$. Let $T_1 = \{\sigma_1, \sigma_2\}$ where σ_1 and σ_2 are the indices of the two largest components where $\sigma_1, \sigma_2 \in \{n_2, n_3, \dots, n_{(N+1)\cdot N}\}$. Continue the process: $T_2 = \{\sigma_3, \sigma_4\}, T_3 = \{\sigma_5, \sigma_6\}, \dots T_s = \{\sigma_{(N+1)\cdot N-1}, \sigma(N+1)\cdot N\}$ if the vector length is even. Let $T = \{1\} \cup T_1$ and $T^c = \cup_{i=2}^s T_i$. We denote $n_{T_i} \in \mathbb{R}^{(N+1)\cdot N}$ be a vector where all the components are zero except at the indices of T_i. Observe that n_T is a 3-sparse vector. Recall that $\mathbf{N} \in \mathcal{N}(\mathcal{A})$ so that

$$\mathbf{0} = \mathbf{A}_{mat} * vec(\mathbf{N}) = \mathbf{A}_{mat} * n_T + \mathbf{A}_{mat} * n_{T^c}$$

which implies

$$\mathbf{A}_{mat} * n_T = -\mathbf{A}_{mat} * n_{T^c}. \tag{11}$$

Then,

$$|n_1| \le |n_T|_2^2 \le \frac{1}{\sqrt{1-\delta}}\|\mathbf{A}_{mat} * n_T\|_2$$

due to the LHS of TRIP. It follows from (11) that $\frac{1}{\sqrt{1-\delta}}\|\mathbf{A}_{mat} * n_T\|_2 = \frac{1}{\sqrt{1-\delta}}\|\mathbf{A}_{mat} * n_{T^c}\|_2$. Now by using the triangle inequality and the RHS of TRIP, we have

$$|n_1| \le \frac{\sqrt{1+\delta}}{\sqrt{1-\delta}} \sum_{i=2}^{s} \|n_{T_i}\|_2. \tag{12}$$

Using the fact

$$\|n_{T_i}\|_2 \le \frac{1}{\sqrt{2}}\|n_{T_i}\|_{\ell_1},$$

from [2], we obtain

$$|n_1| \le \frac{\sqrt{1+\delta}}{\sqrt{2(1-\delta)}} \sum_{i=2}^{(N+1)\cdot N} |n_i|.$$

If $\frac{\sqrt{1+\delta}}{\sqrt{2(1-\delta)}} < 1$, then we satisfy (10).

□

3 Conclusions

We develop a tensor based multilinear system framework which describes genomic interactions. We have shown that TRIP gives the conditions for the sparse multilinear system to have a unique solution. Moreover, TRIP was key in proving the recovery of sparse signals through ℓ_1 minimization.

In our future work, we will provide some numerical methods based on ℓ_1 minimization to approximate the solution to the multilinear system.

References

[1] M. Brazell, N. Li, C. Navasca and C. Tamon. "Solving Multilinear Systems via Tensor Inversion," *SIAM Matrix Analysis,* 34 (2), pp. 542-570, 2013.

[2] K. Bryan and T. Leise. "Making Do with Less: An Introduction to Compressed Sensing," *SIAM Review*, vol. 55, No. 3, pp. 547-566, 2013.

[3] E. Candès and T. Tao. "Decoding by linear programing," *IEEE Trans. Inform. Theory,* 51, pp. 4203-4215, 2005.

[4] L. Hao, A. Sakurai, T. Watanabe, E. Sorenson, and C.A. Nidom et al., "Drosophila RNAi screen identifies host genes important for influenza virus replication," *Nature*, vol. 454, pp. 890-893, 2008.

[5] B. Nazer and R. Nowak. "Sparse Interactions: Identifying High-Dimensional Multilinear Systems via Compressed Sensing," *Proceedings of the 48th Annual Allerton Conference on Communication, Control and Computation*, Monticello, IL, 2010.