Time Series Experiment Design Under One-Shot Sampling: The Importance of Condition Diversity

Xiaohan Kang and Bruce Hajek
Department of Electrical and Computer Engineering
and Coordinated Science Laboratory
University of Illinois at Urbana–Champaign
Urbana, IL 61801, USA
Email: {xiaohank,b-hajek}@illinois.edu

Faqiang Wu and Yoshie Hanzawa
Department of Biology
California State University, Northridge
Northridge, CA 91330, USA
Email: {faqiang.wu,yoshie.hanzawa}@csun.edu

***Abstract*—Many biological data sets are prepared using one-shot sampling, in which each individual organism provides only one sample. Time series therefore do not follow trajectories of individuals over time. However, samples collected at different times from individuals grown/raised under the same conditions share the same perturbations of the biological processes, and hence behave as surrogates for multiple samples from a single individual at different times. This implies the importance of growing/raising individuals under multiple conditions if one-shot sampling is used. This paper models the condition effect explicitly by correlated perturbations in the variations driving the expression dynamics, quantifies the performance of the generalized likelihood-ratio test for network structure, and illustrates the difficulty in network reconstruction under one-shot sampling when the condition effect is absent.**

I. Introduction

Time series data is important for studying biological systems because of their dynamic nature. Ideally it is desirable to use time series with *multi-shot sampling*, where each individual[1] is sampled multiple times to produce the time series data, as in Fig 1. Then the natural biological variations in different individuals can be leveraged for statistical inference, and thus inference can be made even if the samples are prepared under the same experimental condition.

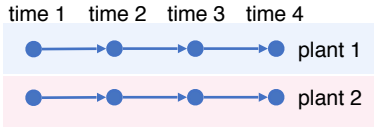


Fig. 1: Multi-shot sampling. Each plant is observed four times.

However, in many experiments multi-shot sampling is not possible. Due to stress response of the organisms and/or the large amount of cell tissue required for accurate measurements, the dynamics of the relevant biological process in an individual of the organism cannot be observed at multiple times without interference. For example, in an RNA-seq experiment an individual plant is often only sampled once in its entire life, leaving the dynamics unobserved at other times. See Section II-A for a review of literature on this subject. We call the resulting time series data, as illustrated in Fig 2, a time series with *one-shot sampling*. Because the time series

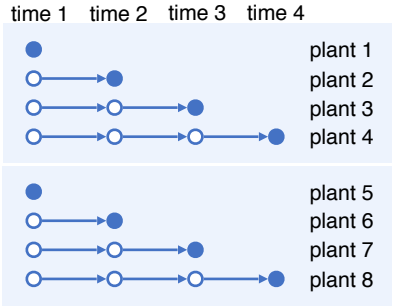


Fig. 2: One-shot sampling. Each plant is observed once.

with one-shot sampling do not follow the trajectories of the same individuals, they do not capture all the correlations in the biological processes. For example, the trajectory of observations on plants 1–2–3–4 and that on 1–6–7–4 in Fig. 2 are statistically identical.

To address this problem, instead of getting multi-shot time series of single individuals, one can grow multiple individuals under each condition with a variety of conditions, and get one-shot time series of the single conditions. The one-shot samples from the same condition then become a surrogate for multi-shot samples for a single individual, as illustrated in Fig 3. In essence, if we view the *condition effect* on the samples

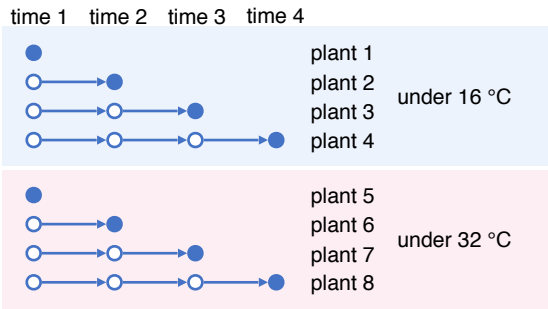


Fig. 3: One-shot sampling with two different conditions.

[1]We sometimes call the individual organism a "plant" in this paper, while the model and analysis apply to animals and microorganisms as well.

as being random, then there should be a positive correlation among samples grown under the same condition. It is similar to the effect of biological variation of a single individual sampled at different times (if such sampling were possible).

For each condition, the one-shot samples at different times are also complemented by *biological replicates*, which are samples from independent individuals taken at the *same* time used to reduce measurement errors and/or biological variations. See Section II-B for a review on how replicates are used for biological inference. With a budget over the number of samples, a balance must be kept between the number of conditions, the number of sampling times and the number of replicates.

To illustrate and quantify the effect of one-shot sampling in biological inference, we introduce a simple dynamic gene expression model with a condition effect. We consider a hypothesis testing setting and model the dynamics of the gene expression levels at different sampling times by a dynamic Bayesian network (DBN), where the randomness comes from biological and environmental variations. The condition-dependent variation of gene j is the same for all plants in that condition and makes up a fraction γ_j of all the variation, the remaining of which is biological variation and is independent for the individuals. The fraction γ_j is called the *condition correlation coefficient* of gene j for the varying conditions and it indicates the strength of the condition effect on the expression of gene j. For example, for varying temperature conditions a differentially expressed gene may have a γ of 0.4 while a non-differentially expressed gene may have a γ of 0.1. In contrast to GeneNetWeaver [1], where the condition effect is modeled by the same random perturbation to the network coefficients, in our model the condition effect is characterized by correlation in the variation terms of the dynamics. Note in both models samples from different individuals under the same condition are statistically independent given the randomness associated with the condition.

The contributions of this paper are threefold.

1) A composite hypothesis testing on the gene regulatory network is formulated and a gene expression dynamic model that explicitly capture the per-gene condition effect and the gene regulatory interactions is proposed.
2) The performance on single-gene self-regulation recovery under one-shot and multi-shot sampling is analyzed, indicating that multiple-condition experiments can somewhat mitigate the shortcomings of one-shot sampling.
3) The difficulty of network reconstruction under one-shot sampling with no condition effect is illustrated. This difficulty connects network analysis and differential expression analysis, two common tasks in large-scale genomics studies, in the sense that the part of network involving non-differentially expressed genes may be harder to reconstruct.

The paper is organized as follows. Section II discusses the related work on one-shot sampling and biological replicates. Section III describes a hypothesis testing setting with one-shot or multi-shot time series observation model and condition correlation coefficient γ. Section IV studies self-regulation sign recovery for a single gene. Section V shows the information limitation of network reconstruction for one-shot sampling with no condition effect. Section VI concludes this paper.

II. Related work

A. One-shot sampling in the literature

This subsection reviews the sampling procedures reported in several papers measuring gene expression levels in biological organisms with samples collected at different times to form time series data. In all cases, the sampling is one-shot, in the sense that a single plant or cell is only sampled at one time.

1) Microorganisms: In the transcriptional network inference challenge from DREAM5 [2], three compendia of biological data sets were provided based on microorganisms (*E. coli*, *S. aureus*, and *S. cerevisiae*), and some of the data corresponded to different sampling times in a time series. Being based on microorganisms, the expression level measurements involved multiple individuals per sample.

2) Plants: In [3], the plants are exposed to nitrate, which serves as a synchronizing event, and samples are taken from three to twenty minutes after the synchronizing event. The statement "… each replicate is independent of all microarrays preceding and following in time" suggests the experiments are based on one-shot sampling. In contrast, the state-space model with correlation between transcription factors in an earlier time and the regulated genes in a later time fits multi-shot sampling. [4] studied the gene expression difference between leaves at different developmental stages in rice. The 12th, 11th and 10th leaf blades were collected every 3 days for 15 days starting from the day of the emergence of the 12th leaves. While a single plant could provide multiple samples, namely three different leaves at a given sampling time, no plant was sampled at two different times. Thus, from the standpoint of producing time series data, the sampling in this paper was one-shot sampling. [5] devised the phenol-sodium dodecyl sulfate (SDS) method for isolating total RNA from *Arabidopsis*. It reports the relative level of mRNA of several genes for five time points ranging up to six hours after exposure to a synchronizing event, namely being sprayed by a hormone *trans*-zeatin. The samples were taken from the leaves of plants. It is not clear from the paper whether the samples were collected from different leaves of the same plant, or from leaves of different plants.

3) Animals: [6] likely used one-shot sampling for their $-24, 60, 120, 168$ hour time series data from mouse skeletal muscle C2C12 cells without specifying whether the samples are all taken from different individuals. [7] produced time series data by extracting cells from a human, seeding the cells on plates, and producing samples in triplicate, at a series of six times, for each of five conditions. Multiple cells are used for each sample with different sets of cells being used for different samples, so this is an example of one-shot sampling of *in vitro* experiment in the sense that each plate of cells is one individual. The use of (five) multiple conditions can serve as a surrogate for a single individual set of cells to

gain the effect of multi-shot sampling. Similarly, the data sets produced by [8] involving the plating of HeLa S3 cells can be classified as one-shot samples because different samples are made from different sets of individual cells. Interestingly, the samples are prepared under one set of conditions, so the use of different conditions is not adopted as a surrogate for multi-shot sampling. However, a particular line of cells was selected (HeLa S3) for which cells can be highly synchronized. Also, the paper does not attempt to determine causal interactions.

B. On biological replicates

This section summarizes how data from biological replicates is used in the analysis of experimental data for the purpose of differential expression analysis of genes, or inference of gene regulatory networks.

In many biological experiments, independent biological replicates are used to reduce the variation in the measurements and to consequently increase the power of the statistical tests. [9] studies the effect of biological replication together with dye switching in microarray experiments and recommends biological replication when precision in the measurements is desired. [10] studies the tradeoff between biological replication and sequencing depth under a sequencing budget limit in RNA-seq differential expression (DE) analysis. They propose a metric for cost effectiveness that suggests a sequencing depth of 10 million reads per library of human breast cells and 2–6 biological replicates for optimal RNA-seq DE design. [11] studies the number of necessary biological replicates in RNA-seq differential expression experiments on *S. cerevisiae* quantitatively with various statistical tools and concludes with the usage of a minimum of six biological replicates. A recent survey ([12]) suggests a minimum of three replicates for RNA-seq experiments whenever sample availability allows.

The choice of replication strategy depends on how the statistical algorithm uses the replicate data. In many differential analysis software packages replicates are treated as independent samples with identical experimental conditions. For example, in edgeR [13] and sleuth [14] the logarithm of the abundance of gene i in sample m is assumed to be $x_m^* \beta_i$, where x_m is the column vector of design characteristics with respect to p variates for sample m and β_i the column vector of the associated effects of the p variates to gene i. Replicate samples can then be used to expand the design matrix x with identical columns. Note that, as a result, replicates are not necessary for edgeR and sleuth because samples with different design characteristics can all contribute to the estimation of β. It is then not clear whether replicates bring more benefit than the sheer additional amount of data compared to samples under different conditions.

For regulatory network reconstruction there is even less consensus on how replicates should be used. One straightforward way is to reduce the replicates into a single set of data by averaging either directly or after a random resampling of the original replicated data. In this case the mean of the replicates are used as a better estimate of the population than each single replicate, while higher moments of the empirical distribution of the replicates are practically ignored. Another way adopted in [3] is to account for all four potential transitions between two replicates in two adjacent sampling times in their machine learning algorithm due to the one-shot nature of the replicates.

III. Stochastic model of time-series samples

This section formulates the hypothesis testing problem of learning the structure of the gene regulatory network (GRN) from gene expression data with one-shot or multi-shot sampling. The GRN is characterized by an unknown adjacency matrix. Given the GRN, a dynamic Bayesian network model is used for the gene expression evolution with time. A parameter γ_j is used for each gene j to explicitly capture the correlation between the variations in the gene expression levels of the individuals under the same experimental condition.

A. Notation

For any positive integer n, let $[n] = \{1, 2, \ldots, n\}$. We use $(f(x))_{x \in \mathcal{I}}$ to denote the family of elements in the set $\{f(x) \colon x \in \mathcal{I}\}$ indexed by the index set $\mathcal{I}$. The indicator function on a statement or a set P is denoted by $\mathbb{1}_P$. The n-by-n identity matrix is denoted by I_n. The transpose of matrix A is denoted by A^*.

B. Model for gene regulatory network topology

Let n be the number of genes and let $A \in \mathbb{R}^{n \times n}$ be the unknown adjacency matrix of the GRN. The sign of the entry a_{ij} of A for $i \neq j$ indicates the type of regulation of j by i, and the absolute value the strength of the regulation. A zero entry $a_{ij} = 0$ with $i \neq j$ indicates no regulation of j by i. The diagonal of A characterizes gene self-regulation. Let $\mathcal{G} = \{G_1, G_2, \ldots, G_{|\mathcal{G}|}\}$ be a finite set of network structures and let D be a mapping from $\mathbb{R}^{n \times n}$ to $\mathcal{G}$. Then $\mathbb{R}^{n \times n}$ is partitioned by the associated network structures. Fix a loss function $l \colon \mathcal{G}^2 \to \mathbb{R}$. Let $Y \in \mathcal{Y}$ be the random observation and let $\delta \colon \mathcal{Y} \to \mathcal{G}$ be an estimator for the structure. The performance of an estimator is evaluated by the expected loss $\mathbb{E}\, l(D(A), \delta(Y))$. This is a hypothesis testing problem with composite hypotheses $\left\{D^{-1}(G) \colon G \in \mathcal{G}\right\}$. This paper considers network reconstruction up to regulation type with $D(A) = (\operatorname{sgn}(A_{ij}))_{(i,j) \in ([n] \times [n])}$ and $l(S, S') = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}_{\{S_{ij} \neq S'_{ij}\}}$, where $\operatorname{sgn}(s) = \mathbb{1}_{\{s>0\}} - \mathbb{1}_{\{s<0\}}$. In other words, the expected loss is the fraction of ordered pairs of genes with wrong signs. The observation model of Y is described in Sections III-C and III-D.

C. Model for gene expression dynamics

This subsection models the gene expression dynamics of individuals by dynamic Bayesian networks with a parameter γ_j for gene j, which is the condition correlation coefficient of variation in the expression level dynamics of gene j.

Let K, T and C be the number of individuals, sampling times, and conditions, respectively. Let $X_j^k(t) \in \mathbb{R}$ be the expression level of gene $j \in [n]$ in individual $k \in [K]$ at time $t \in [T]$, and let $c_k \in [C]$ be the label that indicates the condition for individual k. The gene expression levels evolve

according to the *Gaussian linear model*[2] (GLM) with initial condition $X_j^k(0) = 0$ for any $j \in [n]$, $k \in [K]$ and recursion[3]

$$X_j^k(t+1) = \sum_{i=1}^{n} X_i^k(t) A_{ij} + \sqrt{\gamma_j} W_{\text{cond},j}^{c_k}(t+1) + \sqrt{1-\gamma_j} W_{\text{biol},j}^{k}(t+1), \tag{1}$$

for $j \in [n]$, $k \in [K]$, and $t \in \{0, 1, \ldots, T-1\}$, where $\left(W_{\text{cond},j}^{c}(t)\right)_{(c,j,t)\in[C]\times[n]\times[T]}$, $\left(W_{\text{biol},k}^{j}(t)\right)_{(j,k,t)\in[n]\times[K]\times[T]}$ are collections of independent standard Gaussian random variables that are used to drive the dynamics, and $\gamma_j \in [0,1]$ the fraction of condition-dependent variation for gene j. Here the last two terms in (1) denotes the condition-dependent variation and biological variation respectively. To write (1) in matrix form, we let $X(t) = \left(X_j^k(t)\right)_{(k,j)\in[K]\times[n]}$ and $W(t) = \left(W_j^k(t)\right)_{(k,j)\in[K]\times[n]}$ be K-by-n matrices, where $W_j^k(t) = \sqrt{\gamma_j} W_{\text{cond},j}^{c_k}(t) + \sqrt{1-\gamma_j} W_{\text{biol},j}^{k}(t)$. Then

$$X(t+1) = X(t)A + W(t+1) \tag{2}$$

and hence

$$X(t) = \sum_{\tau=1}^{t} W(\tau) A^{t-\tau}. \tag{3}$$

D. Model for sampling method

This paper considers a *full factorial design* with CRT samples obtained under C conditions, R replicates and T sampling times, denoted by $Y = (Y^{c,r,t})_{(c,r,t)\in[C]\times[R]\times[T]}$. Here the triple index for each sample indicates the condition, replicate and time. Consider two sampling methods that generate Y: multi-shot sampling and one-shot sampling.

1) Multi-shot sampling: Assume an individual can be sampled multiple times. This sampling model requires $K = CR$ and $c_k = \lceil \frac{k}{R} \rceil \in [C]$ for all $k \in [K]$. For all c, r, t let

$$Y^{c,r,t} = X^{(c-1)R+r}(t) + \sigma_Z Z^{c,r,t}, \tag{4}$$

with $\left(Z_j^{c,r,t}\right)_{(j,c,r,t)\in[n]\times[C]\times[R]\times[T]}$ being a collection of independent standard Gaussian random variables modeling the observation noise, and σ_Z is the observation noise level. We see that for fixed c and r the observations of different times are from the same individual indexed by $k = (c-1)R + r$.

2) One-shot sampling: Assume an individual can be sampled only once. This model requires $K = CRT$ and $c_k = \lceil \frac{k}{RT} \rceil \in [C]$ for all $k \in [K]$. Let

$$Y^{c,r,t} = X^{(c-1)RT+(r-1)T+t}(t) + \sigma_Z Z^{c,r,t}. \tag{5}$$

Again σ_Z is the observation noise level and Z's standard Gaussian random variables. Note that for fixed c and r the observations of different times are from *different* individuals.

[2]A logarithm transform is often applied to the estimated concentrations of proteins such that lowly expressed genes have negative expression levels.

[3]Here it is assumed that the data is scaled so $\mathrm{Var}(X_j^k(t+1) \mid X^k(t)) = 1$.

IV. Performance evaluation of network structure estimators

This section studies the performance of network structure estimators with multi-shot and one-shot sampling data. First, general properties of the two sampling methods are obtained. Then the generalized likelihood-ratio test (GLRT) is studied. Finally simulation results on single-gene network reconstruction with GLRT are shown.

A. General properties

By (3), (4) and (5), the samples Y are jointly Gaussian with zero mean. The covariance of the random tensor Y is derived under the two sampling methods in the following.

Under multi-shot sampling, the samples under different conditions are independent and hence uncorrelated. Consider $Y^{c,r,t}$ and $Y^{c,r',t'}$, which are two samples under the same condition and collected at times t and t'. The covariance matrix between $Y^{c,r,t}$ and $Y^{c,r',t'}$ is the sum of the covariance matrices of their common variations at times τ for $1 \le \tau \le t \wedge t'$ multiplied by $(A^*)^{t-\tau}$ on the left and $A^{t'-\tau}$ on the right, plus covariance for the observation noise. The covariance matrices of the variations are identity if the two samples are from the same individual (i.e., $r = r'$), and the diagonal matrix $\Gamma = \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_n)$ otherwise. This yields:

$$\mathbb{E}((Y^{c,r,t})^* Y^{c',r',t'}) = \begin{cases} \sum_{\tau=1}^{t\wedge t'} (A^*)^{t-\tau} A^{t'-\tau} + \sigma_Z^2 \mathbb{1}_{\{t=t'\}} I_n & \text{if } (c,r) = (c',r'), \\ \sum_{\tau=1}^{t\wedge t'} (A^*)^{t-\tau} \Gamma A^{t'-\tau} & \text{if } c = c' \text{ and } r \neq r', \\ 0 & \text{if } c \neq c'. \end{cases}$$

Under one-shot sampling the only difference compared with multi-shot sampling is that two samples indexed by (c,r,t) and (c,r,t') are from different individuals if $t \neq t'$. So

$$\mathbb{E}((Y^{c,r,t})^* Y^{c',r',t'}) = \begin{cases} \sum_{\tau=1}^{t\wedge t'} (A^*)^{t-\tau} A^{t-\tau} + \sigma_Z^2 I_n & \text{if } (c,r,t) = (c',r',t'), \\ \sum_{\tau=1}^{t\wedge t'} (A^*)^{t-\tau} \Gamma A^{t'-\tau} & \text{if } c = c' \text{ and } (r,t) \neq (r',t'), \\ 0 & \text{if } c \neq c'. \end{cases} \tag{6}$$

For any fixed network structure estimator:

1) If $\gamma_j = 1$ for all j and C, R and T are fixed, the joint distribution of the data is the same for both types of sampling. (So the performance of the estimator would be the same for multi-shot and one-shot sampling.)
2) If $\gamma_j = 1$ for all gene j, $\sigma_Z^2 = 0$ (no observation noise) and C, T are fixed, the joint distribution of the data is the same for both types of sampling (as noted in item 1) and any replicates beyond the first are identical to the first. (So the performance of the estimator can be obtained even if all replicates beyond the first are discarded.)
3) Under multi-shot sampling, when C, R, T are fixed with $R = 1$, the joint distribution of the data does not depend on γ for all $\gamma \in [0,1]$. (So the performance of the estimator would be the same for all $\gamma \in [0,1]$.)

B. GLRT

The GLRT (see, e.g., page 38, Chapter II.E in [15]) is given by $\delta(y) = D(\hat{A}_{\mathrm{ML}}(y))$, where $\hat{A}_{\mathrm{ML}}(y)$ is the maximum-likelihood estimate for A based on the covariance of Y in Section IV-A given the observation $Y = y$.

Proposition 1: Suppose all genes have the same condition effect; i.e., $\gamma_j = \gamma$ for all $j \in [n]$. Then GLRT (with the knowledge of γ) has the following properties.

1) Under multi-shot sampling with $\sigma_Z^2 = 0$ (no observation noise), the GLRT performance for sign estimation will be the same for all (R, γ) excluding $(R \geq 2, \gamma = 1)$.
2) Under one-shot sampling and $\gamma = 0$, the log likelihood of the data as a function of A (i.e. the log likelihood function) is invariant with respect to replacing A by $-A$. (So, for the single-gene $n = 1$ case, the log likelihood function is an even function of A, so the GLRT will do no better than random guessing.)

For 2 it suffices to notice in (6) the covariance is invariant with respect to changing A to $-A$. A proof of 1 is given below.

Proof of 1): We first prove it for the case of a single gene with constant T and a constant number of individuals, CR. To do that we need to look at the likelihood function closely.

Because the trajectories for different conditions are independent (for given parameters (A, γ)), we shall first consider the case with a single condition; i.e., $C = 1$. There are hence R trajectories of length T. Then the covariance matrix of the length-R driving vector used at time t for the trajectories is

$$\mathrm{Cov}(W(t)) = (1-\gamma)I_R + \gamma J_R =: \Sigma.$$

When $\gamma > 0$, Σ is not the identity matrix multiplied by some constant; i.e., the noise vector $W(t)$ is colored across replicates. It can be checked when $\gamma < 1$ the matrix Σ is positive definite. Then there exists an orthogonal matrix U and a diagonal matrix Λ with positive diagonal elements such that $\Sigma = U\Lambda U^*$. Let $\Sigma^{-1/2} = U\Lambda^{-1/2}U^*$ and let

$$\widetilde{X}(t) = \Sigma^{-1/2}X(t),$$
$$\widetilde{W}(t) = \Sigma^{-1/2}W(t)$$

for all $t \in [T]$. Then the trajectories for the R replicates in a single condition become:

$$\widetilde{X}(t+1) = \widetilde{X}(t)A + \widetilde{W}(t+1).$$

It can be checked that after the linear transformation by $\Sigma^{-1/2}$, which does not depend on A, the new driving vectors are white (i.e., $\mathrm{Cov}(\widetilde{W}(t)) = I_R$). It follows that the distribution of $\widetilde{X}|(A,\gamma)$ is the same as the distribution of $X|(A,0)$ (i.e. $\gamma = 0$). Therefore, for $x = (x^r(t))_{(r,t)\in[R]\times[T]}$, if we let $L_X(x|A,\gamma)$ denote the likelihood of $X = x$ for parameters A, γ, then

$$L_X(x|A,\gamma) = L_{\widetilde{X}}(\widetilde{x}|A,\gamma)d(R,T,\gamma) = L_X(\widetilde{x}|A,0)d(R,T,\gamma),$$

where $d(R,T,\gamma) = (\det\Sigma)^{-T/2}$ is a function of R, T and γ, and $\widetilde{x}(t) = \Sigma^{-1/2}x(t)$.

Now consider the likelihood function for all CRT samples with general C. It is the product of C likelihood functions for the samples prepared under the C different conditions. It is thus equal to $d(R,T,\gamma)^C$ times the likelihood of the transformed expression levels $\widetilde{x}$, which is the likelihood function for $\gamma = 0$ and a total of CRT samples. The form of the product depends on C and R only through CR, because under the transformation, all CR trajectories are independent. Hence, for fixed A, γ, C, R, T the distribution of the maximum likelihood estimate of A, when the samples are generated using a given $\gamma > 0$ (so the R individuals under each condition are correlated) and the likelihood function also uses γ, is the same as the distribution of the maximum likelihood estimate of A when $\gamma = 0$ (in which case the CR individual trajectories are independent and identically distributed). Formally,

$$\begin{aligned}
\mathbb{E}_\gamma\, l(D(A), \delta(Y)) &= \mathbb{E}_\gamma\, l(D(A), D(\arg\max_A L_X(X|A,\gamma))) \\
&= \mathbb{E}_\gamma\, l(D(A), D(\arg\max_A L_X(\widetilde{X}|A,0))) \\
&= \mathbb{E}_0\, l(D(A), D(\arg\max_A L_X(X|A,0))) \\
&= \mathbb{E}_0\, l(D(A), \delta(Y)),
\end{aligned}$$

where $\mathbb{E}_\gamma$ denotes that the condition effect of the random elements X and Y is γ. The above fails if $\gamma = 1$ and $R \geq 2$ because then Σ is singular. It also fails if γ is unknown to the GLRT.

For the general model with multiple genes, if γ is the same for each gene, 1) holds as before – for the proof, apply left multiplication by $\Sigma^{-\frac{1}{2}}$ for each gene, time, and condition to all R samples in the condition. View (2) as an update equation for an $R \times n$ matrix for each group of R samples in one condition. One column of length R per gene, and one row per sample. ■

C. Simulations

The Gaussian linear model is used to simulate one-shot sampling data with a single gene. In order to compare one-shot and multi-shot sampling, we view the main expense to be proportional to the number of samples. We thus fix a total budget of $CRT = 180$ samples and consider full factorial design with C and R varying with $CR = 30$, and $T = 6$. Results are plotted in Fig. 4. The four plots on the left are for one-shot sampling and the four on the right are for multi-shot sampling. For each plot the observed probability of sign (of A) error is shown for $\gamma \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and for R ranging over the divisors of 30 from smallest to largest. Plots 4a–4d show the performance for the GLRT algorithm assuming no observation noise ($\sigma_Z^2 = 0$) and known γ. Plots 4e–4h show the performance for the GLRT algorithm assuming observation noise with known variance $\sigma_Z^2 = 1$, and with both γ and A unknown to the algorithm.

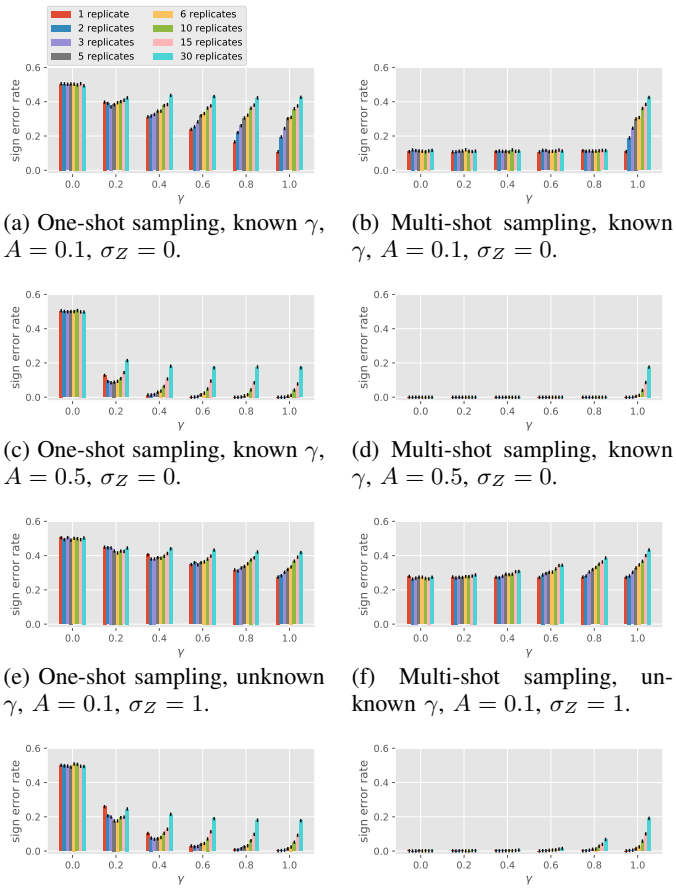
The numerical simulations reflect the following observations implied by the analytical model.

1) Under one-shot sampling, when $\gamma = 0$, the GLRT is equivalent to random guessing.

2) The GLRT performs the same under one-shot and multi-shot sampling when $\gamma = 1$.
3) Under no observation noise, the performance for multi-shot sampling is the same for all $\gamma < 1$.

Some empirical observations are in order.

1) Multi-shot sampling outperforms one-shot sampling (unless $\gamma = 1$, where they have the same error probability).
2) For one-shot sampling, the performance improves as γ increases. Regarding the number of replicates R per condition, if $\gamma = 0.2$ (small condition effect), a medium number of replicates (2 to 5) outperforms the single replicate strategy. For larger γ, one replicate per condition is the best.
3) For multi-shot sampling, performance worsens as γ increases. One replicate per condition ($R = 1$) is best.
4) Comparing 4a–4d vs. 4e–4h, we observe that the performance degrades with the addition of observation noise, though for moderate noise ($\sigma_Z = 1.0$) the effect of observation noise on the sign error is not large. Also, the effect of the algorithm not knowing γ is not large.



(a) One-shot sampling, known γ, $A = 0.1$, $\sigma_Z = 0$.

(b) Multi-shot sampling, known γ, $A = 0.1$, $\sigma_Z = 0$.

(c) One-shot sampling, known γ, $A = 0.5$, $\sigma_Z = 0$.

(d) Multi-shot sampling, known γ, $A = 0.5$, $\sigma_Z = 0$.

(e) One-shot sampling, unknown γ, $A = 0.1$, $\sigma_Z = 1$.

(f) Multi-shot sampling, unknown γ, $A = 0.1$, $\sigma_Z = 1$.

(g) One-shot sampling, unknown γ, $A = 0.5$, $\sigma_Z = 1$.

(h) Multi-shot sampling, unknown γ, $A = 0.5$, $\sigma_Z = 1$.

Fig. 4: Error rates of the GLRT for the sign of a single gene with confidence intervals based on 10 000 simulations.

V. Information limitation for reconstruction under one shot sampling without condition effect

In previous sections we see that when $\gamma_j = 0$ for all j, GLRT is the same as random guessing for the single gene case. This leads us to the following question: is the network reconstruction with no condition effect ($\gamma_j = 0$ for all j) information theoretically possible? In this section we examine this question under general estimator-independent settings. Note in this case the trajectories of all individuals are independent given A regardless of $(c_k)_{k\in[K]}$.

As we have seen in Proposition 1 part 2, when $\gamma_j = 0$ for all j, the distribution of the observed data Y is invariant under adjacency matrix A or $-A$, implying any estimator will have a sign error probability no better than random guessing for the average or worst case over A and $-A$. Here, instead of sign error probability, we consider the estimation for A itself.

The extreme case with infinite number of samples available for network reconstruction is considered to give a lower bound on the accuracy for the finite data case. Note that with infinite number of samples a sufficient statistic for the estimation of the parameter A is the marginal distributions of $X^1(t)$; no information on the correlation of $\left(X^1(t)\right)_{t\in[T]}$ across time t can be obtained. A similar observation is made in [16] for sampling stochastic differential equations.

We first consider the transient case with $X(0) = 0$ as stated in Section III-C. With infinite data the covariance matrix $\Sigma_t \triangleq \mathrm{Cov}(X(t)) = \sum_{\tau=1}^{t}(A^*)^{t-\tau}A^{t-\tau}$ can be recovered for $t \in [T]$. Now we want to solve A from $(\Sigma_t)_{t\in[T]}$. As a special case, if $A^*A = \rho I_n$ (i.e., $\rho^{-1/2}A$ is orthogonal) then we will have $\Sigma_t = \sum_{\tau=0}^{t-1}\rho^\tau I_n$. As a result, given $(\Sigma_t)_{t\in[T]}$ in the above form, A can only be determined up to orthogonality, with $\frac{n(n-1)}{2}$ degrees of freedom remaining. In general case it is not clear if A can be recovered from $(\Sigma_t)_{t\in[T]}$.

Now consider the case where X^k is in steady state; i.e., $X(0)$ is random such that $\mathrm{Cov}(X(t))$ is invariant with t. With infinite amount of data we can get the covariance matrix Σ, which satisfies $\Sigma = A^*\Sigma A + I$. Since covariance matrices are symmetric, we have $\frac{n(n+1)}{2}$ equations for n^2 variables in A. Thus A is in general not determined by the equation uniquely. In fact, note that Σ is positive definite. Then by eigendecomposition $\Sigma = Q\Lambda Q^*$, where Q is an orthogonal matrix and Λ the diagonal matrix of the eigenvalues of Σ. Then $\Lambda = (Q^*AQ)^*\Lambda(Q^*AQ) + I$. Let $B = QAQ^*$. Then $\Lambda = B^*\Lambda B$. By the Gram–Schmidt process, B can be determined with $\frac{n(n-1)}{2}$ degrees of freedom. So the network cannot be recovered from the stationary covariance matrix.

VI. Conclusion

One-shot sampling can miss a lot of potentially useful correlation information. Often gene expression data collected from plants is prepared under one-shot sampling. One factor that can partially mitigate the shortcomings of one-shot sampling is to prepare samples under a variety of conditions or perturbations. Samples grown under the same condition can then be thought of as a surrogate for an individual plant.

To clarify issues and take a step towards quantifying them, we proposed a model with a parameter $\gamma \in [0, 1]$ which is the correlation coefficient induced by preparation of samples under the same condition. The model includes the joint expression level dynamics of multiple genes, although in this paper we focused on the case of a single gene and its self-regulation.

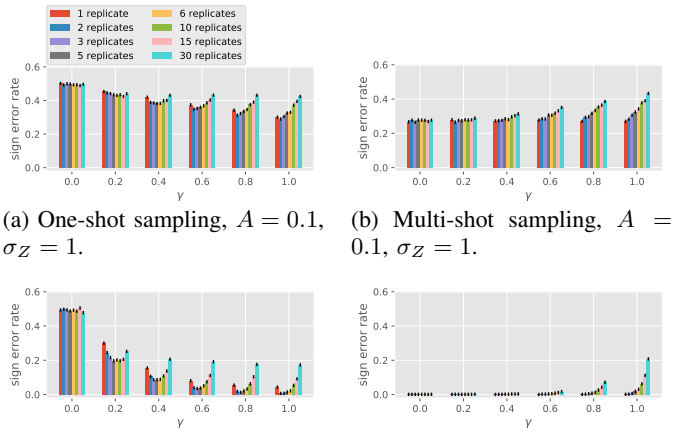
Our results suggest that the preparation of samples under different conditions can help mitigate the shortcomings of one-shot sampling. However, this comes with several caveats. (i) Multi-shot sampling outperforms one-shot sampling in the numerical simulations presented for a single gene network. (ii) Not all genes are responsive to changes in conditions. Thus, the condition effect is gene dependent (which is why we introduced gene dependent γ_j in our multiple gene model). While we found that not knowing γ for a single gene (but estimating it) did not significantly degrade the performance, it remains to know if the same holds for the multi-gene model. In addition, parts of the network involving genes not differentially expressed with conditions may be more difficult to infer. (iii) There is little agreement across the biology literature about how to model the impact of condition on the gene regulatory network. In some cases, it is not even clear that we are observing the same network structure as conditions vary.

Appendix A
Joint A, γ and Observation Noise Estimation

The simulation results in Fig. 4 are for the GLRT estimator that is assumed to know the total variation level of the driving random variables (i.e. the $W_j^k(t)$ random variables), which for brevity we normalized to one. To account for not knowing the variance we could multiply the W's by a positive parameter σ, such that σ that is not known to the algorithm. In addition, it was assumed that the GLRT knows the variance of the observation noise, σ_Z^2. This section presents additional simulations, for a GLRT estimator that needs to estimate all four parameters, A, γ, σ, and σ_Z, again for recovery of the sign of regulation for a single gene. The sign error rates with the true $A = 0.1, 0.5$ and one-shot and multi-shot sampling are shown in Fig. 5. It can be seen in comparison with Fig. 4 that the performance of GLRT is not significantly worse without the knowledge of σ and σ_Z.

The scatter plots of the estimate $\hat{A}$ against the actual A and that of the estimate $\hat{\gamma}$ against the actual γ with different joint prior on A and γ and $C = 10$, $R = 3$, $T = 6$ are shown in Figs. 6–10. Each plot is based on 1000 simulations. For all these figures, the plots on the left are for one-shot sampling and the plots on the right are for multi-shot sampling. Also, for all these figures the GLRT algorithm did not have access to σ and σ_Z. Fig. 6 shows scatterplots that indicate how well A and γ can be estimated when the true value of (A, γ) is uniformly distributed over $[-1, 1] \times [0, 1]$. Figs 7–10 show how well A and γ can be estimated when only one of these parameters is uniformly distributed and the other is fixed at 0.1 or 0.5.

We can see that for all cases shown, the estimates of A and γ are more accurate for multi-shot sampling vs. one-shot sampling. Comparing Figs. 7 and 8, we see that increasing A slightly helps the estimation of γ. Comparing Figs. 9 and 10, we see that increasing γ greatly improves the estimation of A for one-shot sampling, while it slightly hurts the estimation of A for multi-shot sampling. These observations are consistent with the performance of GLRT for sign recovery seen in Figs. 4 and 5. Namely, for one-shot sampling the performance significantly improves as γ increases, while for multi-shot sampling the performance slightly degrades as γ increases (but it is much worse for $\gamma \approx 1$).



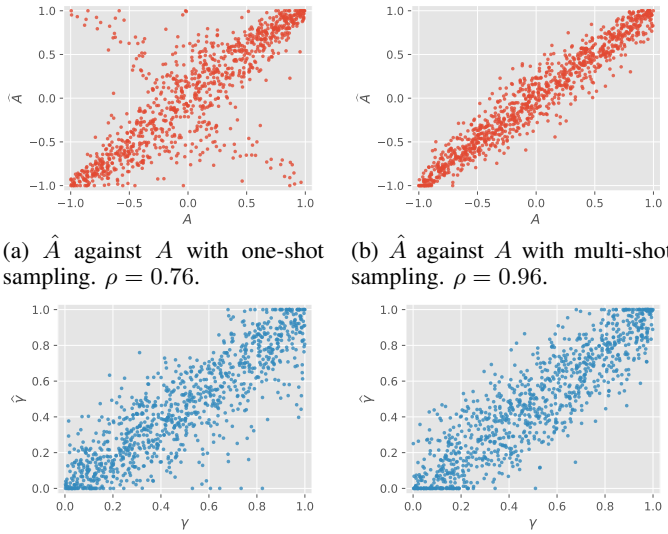
(a) One-shot sampling, $A = 0.1$, $\sigma_Z = 1$.

(b) Multi-shot sampling, $A = 0.1$, $\sigma_Z = 1$.

(c) One-shot sampling, $A = 0.5$, $\sigma_Z = 1$.

(d) Multi-shot sampling, $A = 0.5$, $\sigma_Z = 1$.

Fig. 5: Error rates of the GLRT for the sign of a single gene with confidence intervals based on 10 000 simulations. All four parameters (A, γ, σ and σ_Z) are unknown to the algorithm.



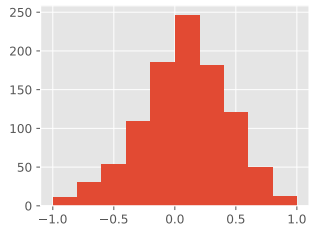
(a) $\hat{A}$ against A with one-shot sampling. $\rho = 0.76$.

(b) $\hat{A}$ against A with multi-shot sampling. $\rho = 0.96$.

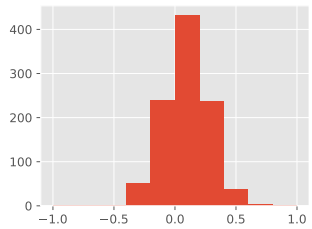
(c) $\hat{\gamma}$ against γ with one-shot sampling. $\rho = 0.87$.

(d) $\hat{\gamma}$ against γ with multi-shot sampling. $\rho = 0.89$.

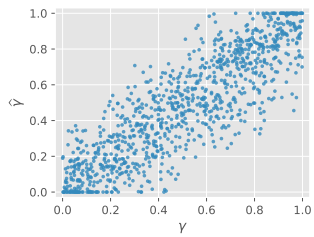
Fig. 6: Joint estimation scatter plots with $\sigma = 1$, $\sigma_Z = 1.0$ and $(A, \gamma) \sim \mathsf{Unif}([-1, 1] \times [0, 1])$.



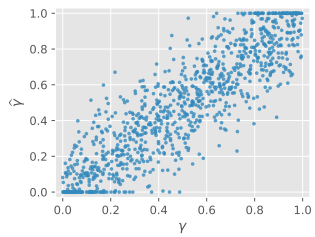
(a) Histogram of $\hat{A}$ for fixed $A = 0.1$ with one-shot sampling.



(b) Histogram of $\hat{A}$ for fixed $A = 0.1$ with multi-shot sampling.

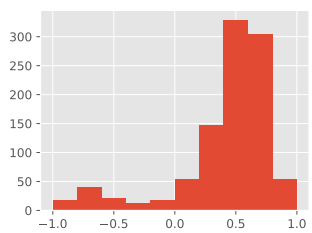


(c) $\hat{\gamma}$ against γ with one-shot sampling and fixed $A = 0.1$. $\rho = 0.87$.

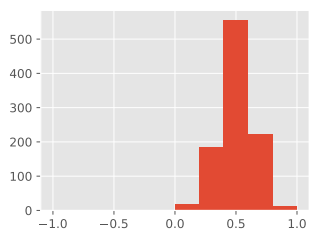


(d) $\hat{\gamma}$ against γ with multi-shot sampling and fixed $A = 0.1$. $\rho = 0.88$.

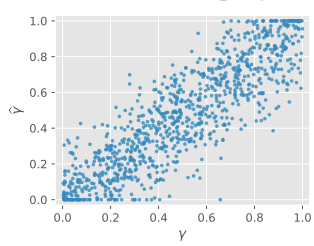
Fig. 7: Joint estimation plots with $\sigma = 1$, $\sigma_Z = 1$, fixed $A = 0.1$ and $\gamma \sim \mathsf{Unif}([0, 1])$.



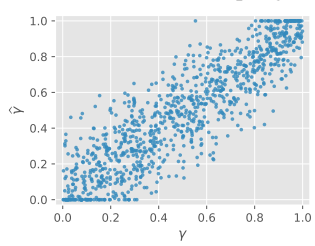
(a) Histogram of $\hat{A}$ for fixed $A = 0.5$ with one-shot sampling.



(b) Histogram of $\hat{A}$ for fixed $A = 0.5$ with multi-shot sampling.

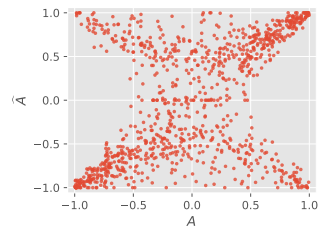


(c) $\hat{\gamma}$ against γ with one-shot sampling and fixed $A = 0.5$. $\rho = 0.88$.

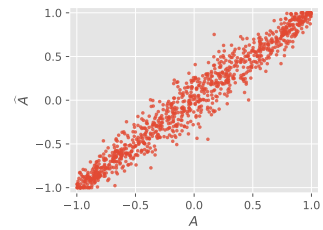


(d) $\hat{\gamma}$ against γ with multi-shot sampling and fixed $A = 0.5$. $\rho = 0.89$.

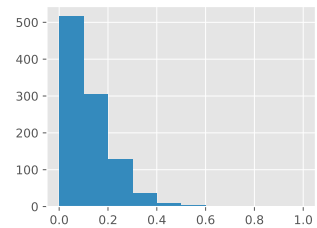
Fig. 8: Joint estimation plots with $\sigma = 1$, $\sigma_Z = 1$, fixed $A = 0.5$ and $\gamma \sim \mathsf{Unif}([0, 1])$.



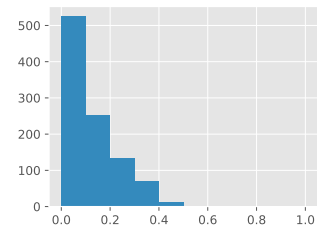
(a) $\hat{A}$ against A with one-shot sampling and fixed $\gamma = 0.1$. $\rho = 0.41$.



(b) $\hat{A}$ against A with multi-shot sampling and fixed $\gamma = 0.1$. $\rho = 0.98$.

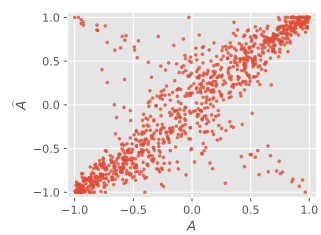


(c) Histogram of $\hat{\gamma}$ for fixed $\gamma = 0.1$ with one-shot sampling.

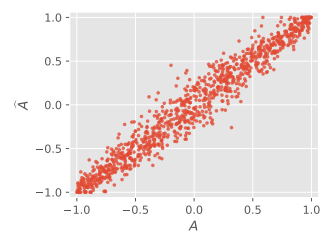


(d) Histogram of $\hat{\gamma}$ for fixed $\gamma = 0.1$ with multi-shot sampling.

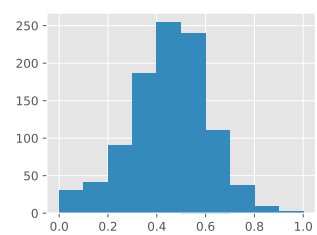
Fig. 9: Joint estimation scatter plots with $\sigma = 1$, $\sigma_Z = 1$, fixed $\gamma = 0.1$ and $A \sim \mathsf{Unif}([-1, 1])$.



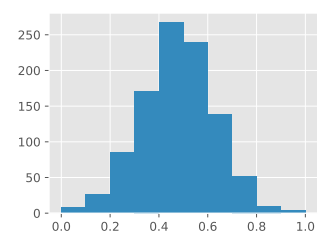
(a) $\hat{A}$ against A with one-shot sampling and fixed $\gamma = 0.5$. $\rho = 0.82$.



(b) $\hat{A}$ against A with multi-shot sampling and fixed $\gamma = 0.5$. $\rho = 0.97$



(c) Histogram of $\hat{\gamma}$ for fixed $\gamma = 0.5$ with one-shot sampling.



(d) Histogram of $\hat{\gamma}$ for fixed $\gamma = 0.5$ with multi-shot sampling.

Fig. 10: Joint estimation plots with $\sigma = 1$, $\sigma_Z = 1$, fixed $\gamma = 0.5$ and $A \sim \mathsf{Unif}([-1, 1])$.

REFERENCES

[1] T. Schaffter, D. Marbach, and D. Floreano, "GeneNetWeaver: *in silico* benchmark generation and performance profiling of network inference methods," *Bioinform.*, vol. 27, no. 16, pp. 2263–2270, 2011.

[2] D. Marbach, J. Costello, R. Küffner, N. Vega, R. Prill, D. Camacho, K. Allison, M. Kellis, J. Collins, and G. Stolovitzky, "Wisdom of crowds for robust gene network inference," *Nature methods*, vol. 9, no. 8, pp. 796–804, Aug. 2012.

[3] G. Krouk, P. Mirowski, Y. LeCun, D. E. Shasha, and G. M. Coruzzi, "Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate," *Genome Biology*, vol. 11, no. 12, p. R123, 2010.

[4] C. Yamaoka, Y. Suzuki, and A. Makino, "Differential expression of genes of the Calvin–Benson cycle and its related genes during leaf development in rice," *Plant and Cell Physiology*, vol. 57, no. 1, pp. 115–124, 2016.

[5] M. Taniguchi, T. Kiba, H. Sakakibara, C. Ueguchi, T. Mizuno, and T. Sugiyama, "Expression of *Arabidopsis* response regulator homologs is induced by cytokinins and nitrate," *FEBS Letters*, vol. 429, no. 3, pp. 259–262, Jun. 1998.

[6] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, May 2010.

[7] V. Y. Kiselev, V. Juvin, M. Malek, N. Luscombe, P. Hawkins, N. L. Novère, and L. Stephens, "Perturbations of PIP3 signalling trigger a global remodelling of mRNA landscape and reveal a transcriptional feedback loop," *Nucleic Acids Research*, vol. 43, no. 20, pp. 9663–9679, Nov. 2015.

[8] M. Whitfield, G. Sherlock, A. Saldanha, J. Murray, C. Ball, K. Alexander, J. Matese, C. Perou, M. Hurt, P. Brown *et al.*, "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," *Molecular Biology of the Cell*, vol. 13, no. 6, pp. 1977–2000, Mar. 2002.

[9] M. Liang, A. G. Briggs, E. Rute, A. S. Greene, and A. W. Cowley, "Quantitative assessment of the importance of dye switching and biological replication in cDNA microarray studies," *Physiological Genomics*, vol. 14, no. 3, pp. 199–207, 2003.

[10] Y. Liu, J. Zhou, and K. P. White, "RNA-seq differential expression studies: more sequence or more replication?" *Bioinformatics*, vol. 30, no. 3, pp. 301–304, 2014.

[11] N. J. Schurch, P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, M. Blaxter, and G. J. Barton, "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?" *RNA*, 2016.

[12] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, "A survey of best practices for RNA-seq data analysis," *Genome Biology*, vol. 17, no. 1, p. 13, Jan 2016.

[13] D. J. McCarthy, Y. Chen, and G. K. Smyth, "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," *Nucleic Acids Research*, vol. 40, no. 10, pp. 4288–4297, 2012.

[14] H. Pimentel, N. L. Bray, S. Puente, P. Melsted, and L. Pachter, "Differential analysis of RNA-seq incorporating quantification uncertainty," *Nature Methods*, vol. 14, pp. 687–690, Jul 2017.

[15] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Springer-Verlag New York, 1994.

[16] J. Bento, M. Ibrahimi, and A. Montanari, "Learning networks of stochastic differential equations," in *Advances in Neural Information Processing Systems (NIPS)*, 2010, pp. 172–180. [Online]. Available: http://papers.nips.cc/paper/4055-learning-networks-of-stochastic-differential-equations.pdf