Shall we dense? Comparing design strategies for time series expression experiments

Emre Sefer and Ziv-Bar Joseph

Machine Learning Department, School of Computer Science, Carnegie Mellon University
{esefer,zivbj}@cs.cmu.edu

Abstract. Recent advances in sequencing technologies have enabled high throughout profiling of several types of molecular datasets including mRNAs, miRNAs, methylation, and more. Many studies profile one or more of these types of data in a time course. An important experimental design question in such experiments is the number of repeats that is required for accurate reconstruction of the signal being studied. While several studies examined this issue for *static* experiments which are often assumed to profile independent samples (for example different patients) much less work has focused on the importance of repeats for time series analysis. Due to budget and sample availability constraints, more repeats in such studies often imply less time points and vice versa.
Here we study this issue by comparing the performance of dense and repeat sampling of time series expression data. We first develop a theoretical framework that can analyze the expected error for these two strategies for a restricted yet expressive set of possible curves over a wide range of possible noise levels. We also analyze real expression data to compare these strategies. For both the theoretical analysis and experimental data we observe that under reasonable assumptions on noise, dense sampling usually outperforms the repeat strategy. Our results provide support to the large number of high throughput experiments that do not perform repeat measurements in each of the time points.

Supporting code and datasets: www.cs.cmu.edu/~esefer/genetheoretical

1. Introduction

High-throughput time-series experiments have been used to study several biological systems and processes. While early work primarily focused on profiling gene expression data using microarrays, recent advances in sequencing technologies enabled the profiling of many more types of biological data over time. These include RNA-Seq [17], microRNAs [13], ChIP-Seq [4], methylation [14] and other epigenetic events [9]. In such studies researchers often determine a start and end point (for example, using the time of infection as the start and 24h as the end in immune response studies [21]), and then select a (usually small) number of time points in between these start and end points in which the high-throughput experiments are performed.

Obviously, the more points that can be profiled between the start and end points, the more likely it is that the reconstructed trajectory for the data type being studied is accurate (gene or miRNA expression, histone modifications over time etc.). However, in practice the number of time points that are used in a study is usually very small [22]. The main limiting factor for most experiments is budget. While technology has greatly improved over the last two decades, high-throughput NGS studies still cost hundreds of dollars per specific experiment. This is a major issue for time series studies, especially those that need to profile multiple types of biological datasets (for example, studies that profile mRNA, miRNAs and methylation levels at each selected point). Another issue that can limit the number of experiments performed (and so the total number of time points that can be used) is biological sample availability. Thus, when designing such experiments researchers often need to balance the overall goals of reconstructing the most accurate temporal representation of the data types being studied and the need to limit the number of experiments as discussed above.

Given these constraints, an important question when designing high-throughput time-series studies is the need for *repeat* experiments. On the one hand, repeats are a hallmark of biological experiments [5] providing valuable information about noise and reliability of the measured values. On the other, repeats further reduce the number of time points that can be profiled which may lead us to miss key events between sampled points. If budget and/or sample availability is an issue, then even one repeat for each time point cuts the number of total points that can be profiled by half which can have a large impact on our ability to accurately reconstruct the trajectories of the biological data being profiled. Further, while repeats can be very useful in dealing with measurement noise, if we assume that the data being studied can indeed be represented by a (smooth) continuous curve, which is often the case [1], then the autocorrelation between successive points can also provide information about the noise in the data (we do not expect large variations between these points). In such cases, more time points, even at the expense of fewer or no repeats, may prove to be a better strategy for reconstructing the dynamics of the type of data being studied.

Indeed, when looking at the large number of time-series datasets deposited in GEO (roughly 25% of all datasets in GEO are time-series, primarily gene expression though other types are starting to appear as well), we observe that in many cases repeats have not been used in these studied [22]. However, to the best of our knowledge, no analysis to date was performed to determine the trade-offs between a dense sampling strategy (profiling more time points using one experiment per point) and repeat sampling (profiling fewer points, with more than once for each point). To study this issue, we use both theoretical analysis and analysis of real data. In our theoretical analysis, we consider a large number of piecewise linear curves and noise levels and compute the expected errors (in terms of the accuracy of the reconstructed curve) when using the two sampling methods. While expression and other profiles are usually not piecewise linear, these curves represent important types of biological responses (for example, gradual or single activation, cyclic behavior, increase and then return to baseline, etc.). Next, we analyze time-series gene expression data to determine the performance of these strategies on real biological data.

Overall, our results support the commonly used (though so far not justified) practice of reducing or eliminating repeat experiments in time-series high-throughput studies. For both the theoretical analysis when using reasonable noise levels and the biological data we analyzed, we see that dense sampling outperforms repeat sampling indicating that for such data auto-correlation can indeed be a useful feature when trying to reduce the impact of noise on the reconstructed curves.

1.1 Related Work

We are not aware of a detailed study that examined the trade-offs between the dense and repeat strategies for profiling time-series high-throughput molecular data. However, the issue of repeats and their impact on static datasets (where no relationship is assumed between consecutive experiments that are not repeats) has been extensively studied. For example, [8] analyzed the variation in a large number of repeat experiments of the *same samples* collected in different dates and determined that overall correlations between these experiments were high. Other have used repeat experiments for follow up analysis including to identify differentially expressed (DE) genes [18], and to improve the performance of clustering methods [16]. Interestingly, while most methods for identifying DE genes in static experiments rely on repeats, several methods for the identification of such DE genes in time-series studies rely on the overall trajectory of the genes [7, 1] which, as we discuss below, may not be best captured using repeats if the budget is limited.

More generally, the issue of repeated experiments in time series studies has been the focus of several statistical papers. For example, for epidemiological studies, tradeoffs were established between frequent measurements of a small number of patients and more infrequent measurements of a larger number of patients [12]. However, the major difference between high throughput biological datasets and most prior work that studied these tradeoffs is the fact that in the biological experiments *all* genes must be sampled at the same time at each experiment. In other words, rather than trying to infer a single curve/profile for each experiment, we are actually inferring tens of thousands of curves simultaneously. Thus, methods for the analysis of such data should consider a much larger set of possible outcomes and examine the impact of the two possible strategies (dense and repeat) in the context of such large number of potential curves.

2. Methods

The main goal of our theoretical analysis discussed below is to develop a framework for computing the expected difference in the resulting error (defined as the difference between the true underlying curves and the estimated curves) between the two possible strategies we are considering. Our goal is to develop methods that can compute such differences for a large set of possible curves. While we constrain our analysis to piecewise linear profiles, as mentioned above these can often represent the outcomes that researchers care about. Indeed, clustering methods based on such piecewise linear representation for time series data have been pretty popular [6] indicating that they can represent an important subset of the possible trajectories.

2.1 A Likelihood-based Framework

We compare two possible strategies for sampling in time series data: Dense and repeat. Dense sampling performs a single expression experiment at each time point whereas repeat performs 2 or more (depending on the setting) such experiments at each of its time points. Since we assume a fixed budget (which means that the number of experiments both method perform is the same), dense sampling is able to query more time points (using uniform sampling), but would have to pay a price in terms of accuracy at each point since no repeats are available.

To analyze the impact of repeats on the ability to accurately reconstruct the gene expression signal we use a probabilistic model. Since expression data is noisy, such model is better able to capture the uncertainty in the measurement repeats. For the theoretical analysis, we assume that the time series is studied between $[0, T_{\max}]$ and that there are k transition points in each expression profile (Figure 1). The value at each transition point can change (up or down) by 1, where 1 represents a unit change in our model (for example, log fold change of 2). Note that while this assumption restricts the set of potential expression profiles, the possible set of resulting curves is still rich enough to define an important subset of expression trajectories. The transition points themselves are *not restricted* in terms of their temporal occurrence and so do not need to coincide with the measured time points. More importantly, by varying k, we can model (using a piecewise linear model) several realistic trajectories. Finally, in many cases researchers are primarily interested in the transition points themselves (for example, the first time a gene becomes differentially expressed) and so such model captures an important aspect of the goals of time series gene expression analysis.

As common in time series expression studies [2], we assume that the value of the first time point for all genes is 0 (since expression data is mostly represented as a log ratio between a specific time point and time point 0). Following several papers [20], we assume that the distribution of noise for observed values is a Gaussian with mean 0 and standard deviation σ.

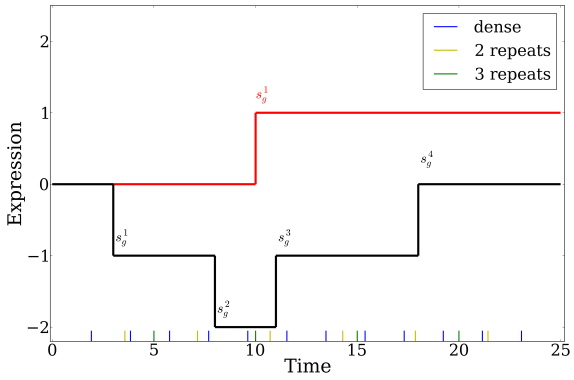


Fig. 1: A step function (red) and a more complex transition function (black). Transition times are denoted by s_g^i. Blue, yellow, and green lines at the bottom represent the sampled points by dense, 2 repeats, and 3 repeats strategies respectively.

More formally, we denote the observed data using *Dense method* for a gene g by D_g^d and for *Repeat method* by D_g^r. Let T^d and T^r be the set of measured time points for each method respectively, and n_r be the number of experiments used for each time point in *Repeat method*. From here on, we use *Dense* and *Repeat*$_{n_r}$ to refer dense and repeat methods with n_r repeats at each time point. For a given budget B, we assume $|T^d| = B$ and $|T^r| = \lceil \frac{B}{n_r} \rceil$. We assume an expression profile for a gene g is defined by transition times $S_g = \{s_g^1, \ldots, s_g^k\}$ and corresponding transition directions $C_g = \{c_g^1, \ldots, c_g^k\}$ where each $c_g^i \in \{-1, 1\}$. The goal of an experiment (using either of the sampling methods) is to detect, as accurately as possible, these transition times and directions. Let $S_r = \{s_r^1, \ldots, s_r^k\}$ and $C_r = \{c_r^1, \ldots, c_r^k\}$ denote the points and directions estimated by *Repeat* strategy. Similarly, let S_d and C_d denote the points and the directions estimated by *Dense*. We assume S_g, S_r, S_d to be sorted in increasing order, and define $f_{\text{mis}}(S_g, C_g, S_r, C_r)$ to be the difference between the area of the true gene profile curve defined by S_g and C_g, and area of the estimated curve defined by S_r, C_r. We compare both strategies by f_{mis}.

2.2 General Likelihood Function

Given the experiment values and k (the required number of transitions), we next need to select the set of transition points and directions for each method S_d, C_d and S_r, C_r. For this, we use the maximum likelihood (ML) criterion. Let A^r be the set of all k-point subsets of T^r that are candidates for S_g, and $Q = \{1, -1\}^k$ be the set of all possible transition directions for these points that are candidates for C_g. Each k-point subset $T^{'} = \{T_i^{'},\, i \in 1, \ldots, k\} \in A^r$ and a transition function $C = \{c_i,\, i \in 1, \ldots, k\} \in Q$ partitions $[0, T_{\max}]$ into $k+1$ intervals $I^{'} = \{I_i^{'} = [T_i^{'}, T_{i+1}^{'}),\, i \in 0, \ldots, k\}$ with corresponding values $\{v_i,\, i \in 0, \ldots, k\}$ where $T_0^{'} = 0$, $T_{k+1}^{'} = T_{\max}$, and $v_{i+1} = v_i + c_{i+1}$. Let $\mathcal{L}(T^{'}, C|D_g^r)$ denote the probability of the observed values for *Repeat* conditioned on transition times $T^{'}$ and directions C. Assuming independent Gaussian measurement noise, this likelihood can be formulated by:

$$\begin{aligned}\mathcal{L}(T^{'}, C|D_g^r) = p(D_g^r|T^{'}, C) &= \prod_{i=0}^{k} \prod_{t_a \le t_j^r < t_b,\, I_i^{'}=[t_a,t_b)} \prod_{z=1}^{n_r} p(d_{j:z}^r|v_i, \sigma) \\ &= \prod_{i=0}^{k} \prod_{t_a \le t_j^r < t_b,\, I_i^{'}=[t_a,t_b)} \prod_{z=1}^{n_r} \frac{1}{\sigma\sqrt{2\pi}}\, \exp^{-\frac{(d_{j:z}^r - v_i)^2}{2\sigma^2}} \end{aligned} \quad (1)$$

where $d_{j:z}^r \in D_g^r$ is the z'th repeat of j'th measured value, $t_j^r \in T^r$ is the set of time points for *Repeat*, and $p(d_{j:z}^r|v_i, \sigma)$ is Gaussian probability of observing $d_{j:z}^r$ given mean v_i and standard deviation σ. To find the ML estimate for S_r and C_r, we set $S_r, C_r = \text{argmax}_{T^{'} \in A^r, \overline{C} \in Q}\; p(D_g^r|T^{'}, \overline{C})$. A similar analysis can be carried out to determine ML estimate for S_d and C_d where we condition on observed values for each point in T^d and $n_r = 1$.

2.3 Analyzing a restricted set of profiles

While our goal is to evaluate the general likelihood function presented above, because of the combinatorial nature of the computation (over all selections of points and directions), it is impossible to compare the methods for completely unrestricted cases. We thus continue by discussing restriction on the general framework that on the one hand allow us to compute a closed form solution to the expected differences between the two methods in a reasonable (polynomial) time while at the same time capture a relevant and biologically important subset of the potential expression profiles.

We start by considering step functions. Such functions allow only a single transition (for example, a gene that is only up or down regulated at some point during the experiment and stays in that level until the end). While step functions are clearly highly restricted, there are many cases where genes with a step function like behavior are of interest, for example when looking for DE genes in a response experiment. For such genes, the key question is to determine the timing of the step event (time of activation). In Section 2.4, we allow functions with a larger (though still known) possible transitions, whereas in Section 2.5, we consider the most general case where both the location and direction of transitions are unknown.

For a step function, we only need to determine a single time point which leads to $s_r = s_r^1$, $s_d = s_d^1$, $s_g = s_g^1$, $c_r = c_r^1$, $c_d = c_d^1$, $c_g = c_g^1$. The likelihood function (1) becomes:

$$\mathcal{L}(s_r, c_r|D_g^r) = \prod_{t_j < s_r} \prod_{z=1}^{n_r} p(d_{i:z}^r|0) \prod_{t_j \ge s_r} \prod_{z=1}^{n_r} p(d_{i:z}^r|c_r) \quad (2)$$

where c_r is the direction change (here an activation so $c_r = 1$). For a step function that transitions from 0 to 1 at time s_g, expected error is:

$$E(f_{\text{mis}}) = \sum_{t_i^r \in T^r} p(s_r = t_i^r|c_r = 1, s_g, c_g, \sigma^2) \underbrace{\Big((T_{\max} - t_i^r)|c_g - c_r| + (t_i^r - s_g)|c_r|\Big)}_{f_{\text{mis}}(s_g, c_g, t_i^r, c_r)} \quad (3)$$

where $p(s_r = t_i^r | c_r, s_g, c_g, \sigma^2)$ is the probability of selecting the i'th time point that transitions into the value $c_r = 1$ conditioned on the actual step time, actual transition direction, and the noise in the measured data.

In order to select t_i as the step point, we need the likelihood defined by it and $c_r = 1$ to be higher than any other point and c_r. From here on, we drop the superscript r when referring to the sampled time points and values, and use the shorthand notation $\mathcal{L}(t_i, 1)$ to denote $\mathcal{L}(t_i^r, 1|D^r)$. Since transition direction is known:

$$p(s_r = t_i^r | c_r = 1, s_g, c_g, \sigma^2) = p(\mathcal{L}(t_i, 1) > \mathcal{L}(t_j, 1), \forall j \neq i) \tag{4}$$

where $\mathcal{L}(t_i, 1)$ is defined in Eq. 2. Computing $p(\mathcal{L}(t_i, 1) > \mathcal{L}(t_j, 1), \forall j \neq i)$ involves nested integrals over pairwise probabilities. Let $S_i = \{t_1, \ldots, t_{i-1}\}$ and $M_i = \{t_{i+1}, \ldots, t_T\}$ be the set of sorted time points that are smaller or larger than t_i respectively, and $p(\mathcal{L}(t_i, c_x) > \mathcal{L}(t_j, c_y))$ be the probability of likelihood defined by t_i and direction c_x being larger than the likelihood defined by t_j and c_y. For $t_j \in S_i$, both predicted curves have the same value up to t_j (0) as well as at and after t_i (1) since $c_x = c_y = 1$. Then, this pairwise probability $p(\mathcal{L}(t_i, 1) > \mathcal{L}(t_j, 1))$ can be expressed as in Eq. 5 in terms of log-likelihood comparison:

$$\begin{aligned} p(\mathcal{L}(t_i, 1) > \mathcal{L}(t_j, 1)) &= p\Big(\log(\mathcal{L}(t_i, 1)) - \log(\mathcal{L}(t_j, 1)) > 0 \Big) = p\Big(-\frac{1}{2\sigma^2} \Big(\sum_{m=j}^{i-1} \sum_{z=1}^{n_r} (d_{m:z})^2 - \sum_{m=j}^{i-1} \sum_{z=1}^{n_r} (d_{m:z} - 1)^2 \Big) > 0 \Big) \\ &= p\Big(\sum_{m=j}^{i-1} \sum_{z=1}^{n_r} d_{m:z} \leq n_r \frac{h}{2} \Big) = \Phi(n_r \frac{h}{2}, m_j^i, \sigma_j^i) \end{aligned} \tag{5}$$

where h is the number of measurements between t_j and t_{i-1} including both time points and $\Phi(n_r \frac{h}{2}, m_j^i, \sigma_j^i)$ is cdf of Gaussian with mean m_j^i and a standard deviation σ_j^i. Since we know s_g and c_g (the computation is conditioned on them) and we are dealing with Gaussian, the sum of the observations is also a Gaussian with mean $m_j^i = n_r \sum_{m=j,\, t_m \geq s_g}^{i-1} 1$ and standard deviation $\sigma_j^i = \sqrt{n_r \sum_{m=j}^{i-1} \sigma^2}$.

Repeating the pairwise comparison in Eq. 5 for all points in S_i and M_i returns set of distributions that need to be satisfied. For a step function, distributions returned by S_i and M_i are independent of each other, so the nested integral for Eq. 4 can be separated into two integrals each of which can be efficiently estimated by Gaussian quadrature or by MCMC [10] (See Appendix for details).

2.4 Analyzing profiles with multiple transitions

Following the analysis of step functions, where we focused on identifying a single change point, we now consider the more general (though still not the most general) case where we no the number of transition points and the direction (for example, $0, 1, 2, 1$) but do not know the specific time points in which they occur. In this case, we estimate $p(S_r = T^i | C_r, S_g, C_g, \sigma^2)$ for $T^i \in A^d$ in Eq. 3 which is defined as the probability of the likelihood defined by T^i to be higher than the likelihood defined by any other k-subset.

In order to estimate this probability, we follow the approach used for the step function, and define the pairwise probability $p(\mathcal{L}(T^i, C^r) > \mathcal{L}(T^j, C^r))$ as in:

$$p(\mathcal{L}(T^i, C^r) > \mathcal{L}(T^j, C^r)) = p\Big(\sum_{m=0}^{k} \sum_{\substack{t_a \leq t < t_b, \\ I_m^i = [t_a, t_b)}} \sum_{z=1}^{n_r} (d_{t:z} - v_m^r)^2 - \sum_{n=0}^{k} \sum_{\substack{t_a \leq t < t_b, \\ I_n^j = [t_a, t_b)}} \sum_{z=1}^{n_r} (d_{t:z} - v_n^r)^2 > 0 \Big) \tag{6}$$

where I^i and I^j are intervals defined by T^i and T^j, and $v_{m+1}^r = v_m^r + c_{m+1}^r$ is the corresponding value of the $m + 2$'th interval defined by transition directions C^r. For every k-subset T^i, there are $\binom{T}{k} - 1$ comparisons intersection of which define the integral boundaries for estimating Eq. 4. In contrast to step functions in Section 2.3, we cannot separate the estimation of the nested integral in Eq. 4 into two parts since there is no total ordering and independence between variables. In this case, we estimate the integral by sampling over the domain.

2.5 General Transition Functions

Finally, we arrive at the most general case where both the location and direction of transitions are unknown. Note that the number of transition k is an input for this computation, but since the goal of the modeling here is to determine how well dense and repeat methods do, we can easily perform the computation on all relevant values of k to reach the conclusions we are interested in for a specific noise model (it is unlike that genes would have more than 5-6 transitions in most time series studies, in fact in most cases they have much fewer). For the case of k possible transitions, expected error becomes:

$$E(f_{\text{mis}}) = \sum_{T^i \in A^r} \sum_{C^x \in Q} p(S_r = T^i, C_r = C^x | S_g, C_g, \sigma^2) f_{\text{mis}}(S_g, C_g, T^i, C_x) \tag{7}$$

where $p(S_r = T^i, C_r = C^x | S_g, C_g, \sigma^2)$ is the probability of selecting k-point subset T^i and set of transition directions $C^x = \{c_1^x, \ldots, c_k^x\}$ conditioned on the actual step time, actual transition direction, and the noise in the measured data. In Eq. 7, expectation is taken over all possible k-point subsets of T^r and all possible transition directions C^x of length k since we also do not know the transition directions. When estimating $p(S_r = T^i, C_r = C^x | S_g, C_g, \sigma^2)$, we want the likelihood defined by T^i and C^x to be higher than the likelihood defined by any other k-point subset T^j and k-length transition directions C^y pair. We follow the approach used for the step function, and define the pairwise probability $p(\mathcal{L}(T^i, C^x) > \mathcal{L}(T^j, C^y))$ as in:

$$p(\mathcal{L}(T^i, C^x) > \mathcal{L}(T^j, C^y)) = p\Big(\sum_{m=0}^{k} \sum_{\substack{t_a \le t < t_b, \\ I_m^i = [t_a, t_b)}} \sum_{z=1}^{n_r} (d_{t:z} - v_m^x)^2 - \sum_{n=0}^{k} \sum_{\substack{t_a \le t < t_b, \\ I_n^j = [t_a, t_b)}} \sum_{z=1}^{n_r} (d_{t:z} - v_n^y)^2 > 0 \Big) \tag{8}$$

where I^i and I^j are intervals defined by T^i and T^j, and $v_{m+1}^x = v_m^x + c_{m+1}^x$, v_{m+1}^y are the corresponding values of the $m+2$'th interval defined by transition directions C^x and C^y respectively. For every k-subset T^i and C^x, there are $\binom{T}{k} 2^k - 1$ comparisons defining the integral boundaries in estimating Eq. 4. Similar to Section 2.4, full ordering is not guaranteed between the variables, so nested integral in Eq. 4 can again be estimated via sampling.

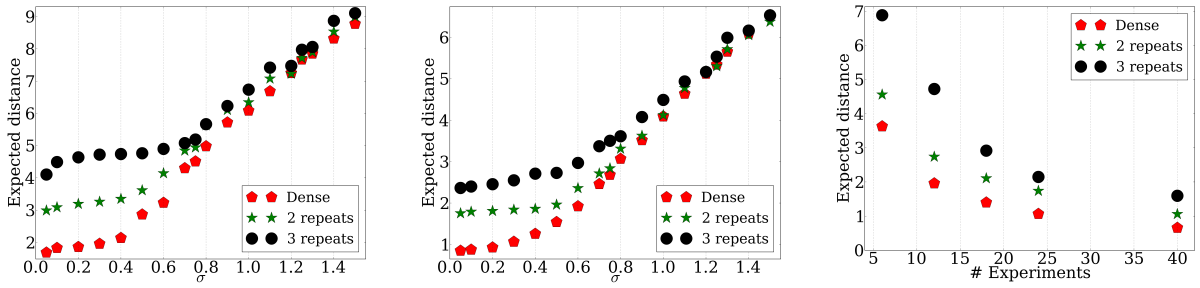
3. Results

To test the difference between using a dense sampling strategy with more time points profiled vs. a repeat strategy with fewer points (but the same number of experiments), we first used the theoretical framework discussed above to evaluate the expected performance of the two strategies and then compared them using real gene expression profiles. For the theoretical analysis, we assumed that gene expression was measured between 0 and $50h$ (similar to real experiments, for example [19]). In such setting, when we have a budget for x experiments, the dense method (*Dense*) performs x RNA-Seq (or microarray) experiments uniformly between 0 and $50h$, whereas $Repeat_2$ and $Repeat_3$ perform 2 and 3 experiments at $\frac{x}{2}$ and $\lceil \frac{x}{3} \rceil$ uniformly sampled points, respectively. As mentioned above, we assume that noise in each measurement for each gene is Gaussian (mean 0, and standard deviation σ ranging between 0.1 and 1.5).

3.1 Detecting transition time for step functions

We first evaluated the performance for step functions. These functions can represent genes that start as non active (0) and become active after a certain time point (1), where the goal is to determine the transition time (Methods). For each of the noise levels we consider, we randomly selected 100 transition points and evaluated the performance of the two strategies for the resuulting curves. As can be seen in Figures (2a)–(2b), for such profiles *Dense* performs better for noise levels lower than 0.9 for both 12 and 24 experiments. Note that because we assume that the difference between an active and non-active gene is 1, a standard deviation close

to 1 is unlikely and so values less than 0.9 are more likely in practice. Indeed, for most gene expression experiments σ is much lower than 1 when analyzing log scale values (for example, close to 0.3 for [3]). For such values, *Dense* is clearly much better than *Repeat*. We have also tested a larger number of experiments fixing the noise standard deviation at 0.3. As can be seen in Figure 2c, when the number of experiments increases beyond 24, the improvement seen for the dense strategy decreases However, even for a very large number of experiments (40 over 50 hours with a single transition) *Dense* still outperform *Repeat* when using $\sigma = 0.3$.



(a) 12 experiments (b) 24 experiments (c) $\sigma = 0.3$

Fig. 2: Comparison of the strategies for different number of experiments and noise levels. a) 12 experiments, b) 24 experiments, c) Different number of experiments for a fixed noise $\sigma = 0.3$.

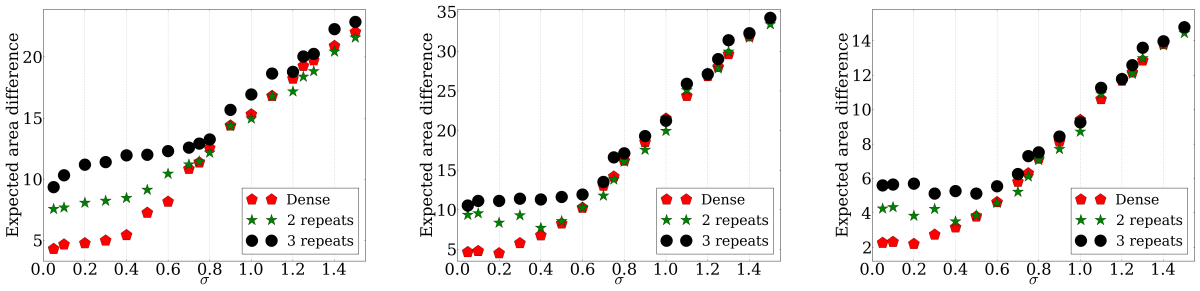
3.2 Analysis of more complicated transition functions

Following the analysis of the step function scenario we analyzed more complex transition profiles including monotonically increasing and non-monotonic transitions (Figures (3a)–(3b)) with 12 experiments. Specifically, we looked at a monotonically increasing function $0, 1, 2, 3$ representing a gene that is continuously up-regulated during the course of the study (common in response experiments, for example immune response [15]) and at a $0, 1, 0, 1$ representing a fluctuating gene (for example, for cases of cyclic activity such as cell cycle and circadian rhythms [11]). For these functions, we use the theoretical analysis above to compute the expected area difference between the true profile and the estimated profile for each of the methods (since the direction of the transitions are known, the differences are a function of inaccurate estimation of the transition time points). As can be seen in Figures (3a) and (3b), *Dense* outperforms the repeat methods when the noise is low to moderate. However, even for high noise values we see that the repeat based methods do not improve upon the dense method results indicating that even when the noise levels cannot be completely determined, using the dense strategy is at least going to lead to comparable results to the repeat methods, and in most cases would outperform them.

Figure 3c present results for the most general type of our theoretical framework. For this analysis, we fixed the number of transitions (in this case to 3) but do not assume that the directions are known. Thus, the analysis considers all possible 2^3 transition profiles as discussed in Methods. Again, even for this unrestricted version of the problem, we see that in noise levels up to 0.6 (which as mentioned above is much higher than often observed in practice) *Dense* outperforms repeat based sampling. Results are more mixed for higher noise levels, so there does not seem be a noise level in which repeat strongly dominates the dense sampling method.

3.3 Analysis of real biological data

The analysis above used our theoretical framework to compare the dense and repeat strategies for various profiles and noise levels. While such analysis is informative since it applies to any measurements resulting from the setting being considered, it is also important to analyze real biological expression data in order to compare the two strategies. For this, we used a gene expression dataset that profiled 22769 genes in *Anopheles Gambiae* for 48 hours. The study had two settings, both with 13 experiments over the duration being studied: 12 hours light/ 12



(a) Fluctuating $k = 3$ (b) Monotonically increasing (c) Fluctuating (12 experiments)

Fig. 3: Comparison of sampling strategies for different noise levels over 12 experiments in terms of expected area difference. a) Fluctuating profile with 3 transitions, b) Monotonically increasing profile. c) Comparison of dense and repeat strategies for recovering a profile for which the directions are unknown. The real data is generated from a fluctuating profile with 3 transitions, though these are not known in advance and so the likelihood function used to select the transition points and directions for both dense and repeat does not use this information. Results presented for different noise levels when performing 12 experiments.

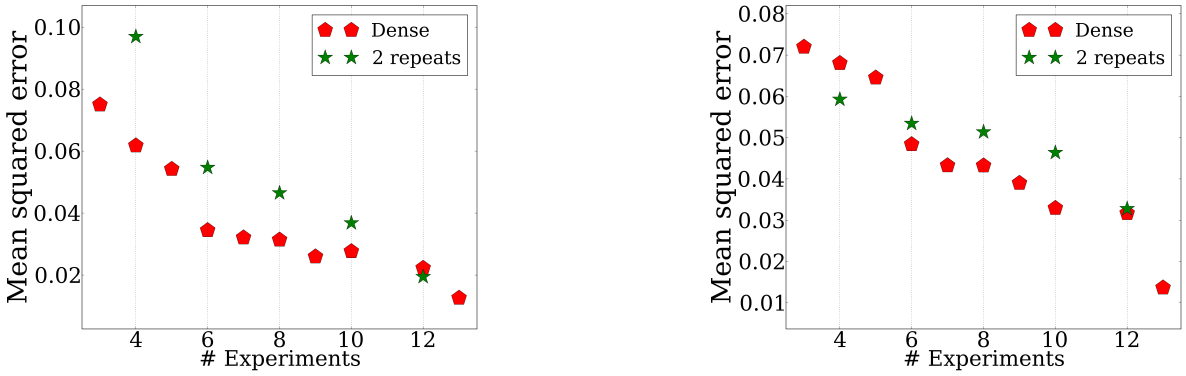
hours dark (LD) switching and constant dark (DD). Experiments were performed every 4 hours with 2 repeats for each time point used. As usual, we computed the values in each time point as log fold changes to the values at time point 0. For both strategies, we performed the following analysis: Given a specific number of experiments (upper bounded by 13, the total number of points sampled), we sample time points uniformly between 12 and $60h$ for each strategy. We use the value of the closest time point if a time point is not measured in the original dataset. For *Dense*, we randomly select one of the repeats at each of the time points that are used while both measurements are used for *Repeat* (though the total number of points used by repeat is half that of the ones used by dense). Next, we fit interpolating splines for each gene and estimate the mean squared error (MSE) by comparing to median values obtained when using *all* sampled points. Note that in all experiments at least half the experiments are not used (even when sampling 13 points for dense it only uses 1 experiment for each time point) and so the test data is not fully used in the reconstruction even when using the most number of points. We repeat this procedure 10000 times for *Dense* and *Repeat* and report the mean error.

We find that *Dense* outperforms *Repeat* in the both conditions studied and in some cases significantly so as in Figures 4a–4b (for example, when the budget only allows for 6-8 experiments in the LD setting). The performance difference between them decreases when the number of experiments increases. However, *Dense* is as good as, or better than, repeat for all settings.

Figures 5a–5c present the observed and reconstructed values using both *Dense* and *Repeat* for three of profiled genes using 8 experiments for the LD. This figure helps explain the differences between the performance of the two methods. For example, AGAP010735 which is known to activate the response to oxidative stress in *Anopheles Gambiae* declines rapidly and stays low for the reminder of the experiment. While *Dense* is able to capture this decline very early, it takes the *Repeat* method, which samples points less frequently, a much longer time to determine the actual magnitude of the decline leading to inaccurate reconstruction of the early response of this gene. Even more importantly, for AGAP000987 which was identified in this study as cycling with the condition (the key goal of this experiment), *Dense* indeed recovers the correct 2 cycles profile while *Repeat* completely misses the correct profile.

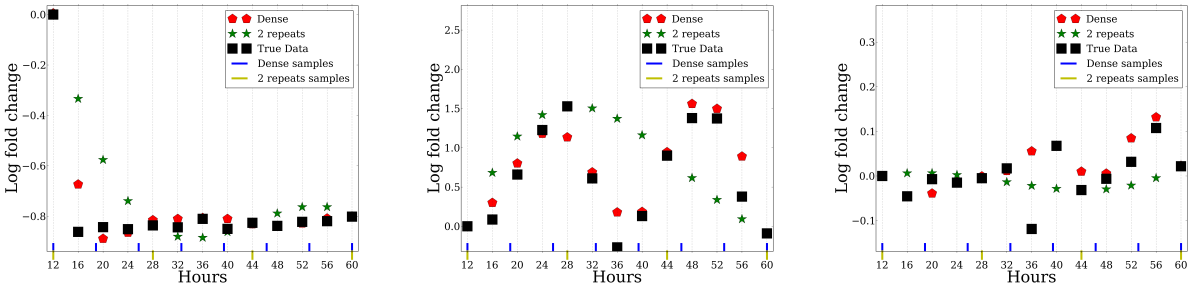
3.4 Comparisons using a subset of the genes

The above analysis looked at performance over all genes profiled in the experiment. However, in most case researchers tend to focus on a much smaller subset of genes (often the most varying)



(a) Light/Dark (LD) (b) Constant dark (DD)

Fig. 4: Comparison of strategies over all genes of *Anopheles Gambiae* by increasing number of experiments over a) LD data b) DD data.



(a) AGAP010735 (b) AGAP000987 (c) AGAP011186

Fig. 5: Comparison of the two strategies on individual genes by 8 experiments a) AGAP010735, b) AGAP000987, c) AGAP011186.

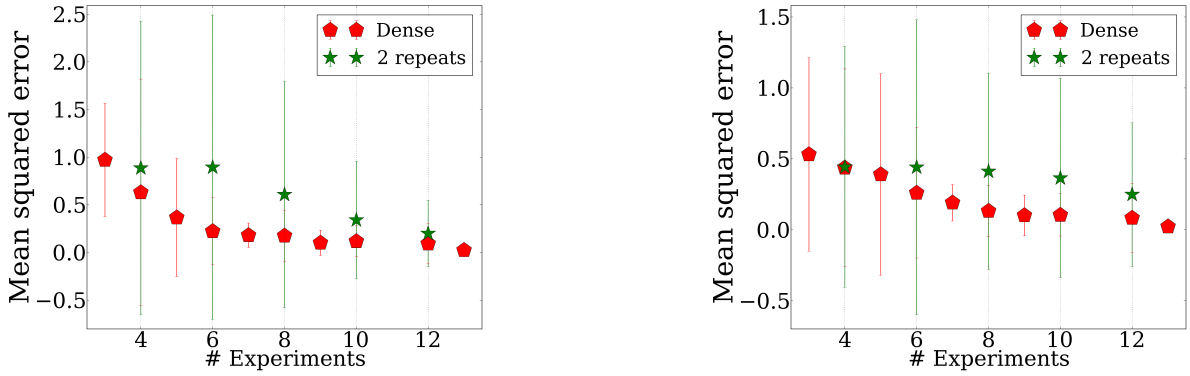
and any strategy for designing experiments should be able to recover an accurate representation for these genes. To study the difference between the dense and repeat strategies for these key genes, we used 536 rhythmic genes identified as rhythmic using a cosine wave-fitting algorithm for both LD and DD conditions by Rund et al. [11]. We use a spline fitting procedure as discussed in Section 3.3. Figures 6a–6b present the results obtained for this subset for LD and DD conditions respectively. As can be seen, we again observe that *Dense* performs better than $Repeat_2$ for this important subset.

Figures (7a)–(7b) present the gene specific performance differences for this smaller set of genes (as opposed to the average differences presented above). Genes are ordered based on the difference between the error obtained by *Dense* and *Repeat* strategies when using 8 experiments. Specifically, we observe that *Dense* performs better than *Repeat* for 468 (87%) and 523 (98%) out of the 536 rhythmic genes in the LD and DD datasets respectively. Even when increasing the number of experiments to 10 *Dense* still performs significantly better than $Repeat_2$ ($p < 0.01$, Wilcoxon rank-sum test).

4. Conclusion

While repeat experiment have been widely used in high-throughput analysis studies, they have been utilized to a much lesser extent when using the same technology to study time series data [22]. While it is hard to determine the exact causes for this practice, it is very likely that budget and sample quantity constraints have played a role. However, no systematic study examined the tradeoffs between more time points and more repeats for such studies.

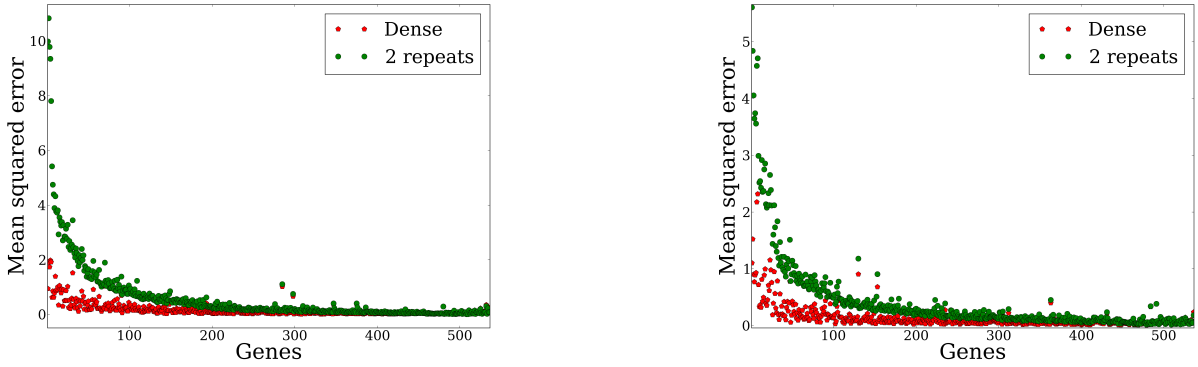
Here we have tried to address this issue using a combined theoretical and analysis framework. Our theoretical models consider the impact of various noise levels on the ability of each of these



(a) Light/Dark (LD)

(b) Dark/Dark (DD)

Fig. 6: Comparison of strategies over all genes exhibiting circadian and diel rhythms by increasing number of microarrays over a) LD data, b) DD data. Std. dev. of the error is estimated over the considered genes.



(a) Light/Dark (LD)

(b) Constant dark (DD)

Fig. 7: Comparison of strategies over all genes exhibiting circadian and diel rhythms where individual genes sorted by decreasing MSE difference between *Dense* and $Repeat_2$ when using 8 experiments over LD and DD data respectively.

strategies to correctly infer the underlying profile. As we show, by analyzing a restricted yet expressive set of piecewise linear curves, for reasonable noise levels a dense sampling strategy leads to better results than a repeat strategy that uses the same number of experiments. We obtain similar results when analyzing real biological gene expression data for both, the full set of genes being studied and a subset of the key genes identified in a specific study.

While we conclude that a dense sampling strategy is beneficial when the number of experiments is limited by external constraints, we do not claim that repeats do not provide additional and valuable information. Obviously, if such constraints do not exist, or if it is possible to increase the number of experiments performed, repeats are an important and useful strategy for identifying DE genes and for clustering and modeling their behavior. However, even those studies that do not utilize repeats can still extract some of these benefits (especially the ability to deal with noise) by relying on autocorrelation between successive, densely sampled, points.

The datasets and code used in this study are available from the supporting website. Our analysis is focused on gene expression data because it is still the largest type of data to be profiled by time series experiments. However, as mentioned in the Introduction, several other types of data are now being studied using similar experiments and future work is required to determine if our results hold for these types of data as well.

References

1. Bar-Joseph, Z., Gerber, G.K., Gifford, D.K., Jaakkola, T.S., Simon, I.: Continuous representations of time-series gene expression data. Journal of Computational Biology 10(3-4), 341–356 (2003)
2. Bar-Joseph, Z., Gitter, A., Simon, I.: Studying and modelling dynamic biological processes using time-series gene expression data. Nature Reviews Genetics 13(8), 552–564 (2012)
3. Blake, W.J., Kærn, M., Cantor, C.R., Collins, J.J.: Noise in eukaryotic gene expression. Nature 422(6932), 633–637 (2003)
4. Chang, K.N., Zhong, S., Weirauch, M.T., Hon, G., Pelizzola, M., Li, H., Huang, S.s.C., Schmitz, R.J., Urich, M.A., Kuo, D., et al.: Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in arabidopsis. Elife 2, e00675 (2013)
5. Cumming, G., Fidler, F., Vaux, D.L.: Error bars in experimental biology. The Journal of cell biology 177(1), 7–11 (2007)
6. Ernst, J., Bar-Joseph, Z.: Stem: a tool for the analysis of short time series gene expression data. BMC bioinformatics 7(1), 191 (2006)
7. Kim, J., Ogden, R.T., Kim, H.: A method to identify differential expression profiles of time-course gene data with fourier transformation. BMC bioinformatics 14(1), 310 (2013)
8. Mongan, M.A., Higgins, M., Pine, P.S., Afshari, C., Hamadeh, H.: Assessment of repeated microarray experiments using mixed tissue rna reference samples. BioTechniques 45(3), 283–292 (2008)
9. Paige, S.L., Thomas, S., Stoick-Cooper, C.L., Wang, H., Maves, L., Sandstrom, R., Pabon, L., Reinecke, H., Pratt, G., Keller, G., et al.: A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. Cell 151(1), 221–232 (2012)
10. Press, W.H.: Numerical recipes 3rd edition: The art of scientific computing. Cambridge university press (2007)
11. Rund, S.S., Hou, T.Y., Ward, S.M., Collins, F.H., Duffield, G.E.: Genome-wide profiling of diel and circadian gene expression in the malaria vector anopheles gambiae. Proceedings of the National Academy of Sciences 108(32), E421–E430 (2011)
12. Schmidt, W.P., Genser, B., Barreto, M.L., Clasen, T., Luby, S.P., Cairncross, S., Chalabi, Z.: Sampling strategies to measure the prevalence of common recurrent infections in longitudinal studies. Emerging themes in epidemiology 7(1), 5 (2010)
13. Schulz, M.H., Pandit, K.V., Cardenas, C.L.L., Ambalavanan, N., Kaminski, N., Bar-Joseph, Z.: Reconstructing dynamic microrna-regulated interaction networks. Proceedings of the National Academy of Sciences 110(39), 15686–15691 (2013)
14. Singer, Z.S., Yong, J., Tischler, J., Hackett, J.A., Altinok, A., Surani, M.A., Cai, L., Elowitz, M.B.: Dynamic heterogeneity and dna methylation in embryonic stem cells. Molecular cell 55(2), 319–331 (2014)
15. Teschendorff, A.E., Miremadi, A., Pinder, S.E., Ellis, I.O., Caldas, C.: An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol 8(8), R157 (2007)
16. Tjaden, B.: An approach for clustering gene expression data with error information. Bmc Bioinformatics 7(1), 17 (2006)
17. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L.: Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nature protocols 7(3), 562–578 (2012)
18. Tu, Y., Stolovitzky, G., Klein, U.: Quantitative noise analysis for gene expression microarray experiments. Proceedings of the National Academy of Sciences 99(22), 14031–14036 (2002)
19. Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., et al.: Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Molecular biology of the cell 13(6), 1977–2000 (2002)
20. Yip, K.Y., Alexander, R.P., Yan, K.K., Gerstein, M.: Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. PloS one 5(1), e8121–e8121 (2010)
21. Yosef, N., Regev, A.: Impulse control: temporal dynamics in gene transcription. Cell 144(6), 886–896 (2011)
22. Zinman, G.E., Naiman, S., Kanfi, Y., Cohen, H., Bar-Joseph, Z.: Expressionblast: mining large, unstructured expression databases. Nature methods 10(10), 925–926 (2013)

5. Appendix

Estimating Eq. 4 for a step function Repeating the pairwise comparison in Eq. 5 for all points in S_i and M_i return the folllowing $i-1$ and $T-i$ distributions to be satisfied respectively:

$$p\big(\sum_{a=1}^{j}\sum_{z=1}^{n_r} d_{a:z} \leq n_r\frac{j}{2}\big), \quad j \in 1,\ldots,i-1 \tag{9}$$

$$p\big(\sum_{a=i}^{j-1}\sum_{z=1}^{n_r} d_{a:z} \geq n_r\frac{j-i}{2}\big), \quad j \in i+1,\ldots,T \tag{10}$$

$d_{a:z}$ terms in Eq. 9 and Eq. 10 are independent of each other, so the probability of selecting t_i in Eq. 4 can be separated into two integrals as in:

$$p(s_r = t_i^r | c_r = 1, s_g, c_g, \sigma^2) = p\big(\mathcal{L}(t_i,1) > \mathcal{L}(t_j,1),\, \forall j \neq i\big) = p\big(\mathcal{L}(t_i,1) > \mathcal{L}(t_j,1),\, \forall j \in S_i\big)\; p\big(\mathcal{L}(t_i,1) > \mathcal{L}(t_j,1),\, \forall j \in M_i\big) \tag{11}$$

where $p\big(\mathcal{L}(t_i,1) > \mathcal{L}(t_j,1),\, \forall j \in S_i\big)$ is the probability of the likelihood defined by t_i being higher than the likelihood of all other points that are smaller than t_i. $d_{a:z}$ variables for each time point t_a in S_i have acyclic dependencies between them, $d_{a:z}$ variables depend only on the variables of time points between t_a and t_{i-1}. Due to the existence of this ordering between variables, $p\big(\mathcal{L}(t_i,1) > \mathcal{L}(t_j,1),\, \forall j \in S_i\big)$ can be expressed by the following nested integral:

$$p(\mathcal{L}(t_i,1) > \mathcal{L}(t_j,1),\, \forall j \in S_i) = \int_{-\infty}^{\frac{n_r}{2}} p(\hat{d}_{i-1}|m_{i-1}^i, \sigma_{i-1}^i) \int_{-\infty}^{n_r - \hat{d}_{i-1}} p(\hat{d}_{i-2}|m_{i-2}^{i-1}, \sigma_{i-2}^{i-1}) \cdots \int_{-\infty}^{n_r \frac{i-1}{2} - \sum_{t=2}^{i-1}\hat{d}_t} p(\hat{d}_1|m_1^2, \sigma_1^2)\, d_{\hat{d}_1} \cdots d_{\hat{d}_{i-2}} d_{\hat{d}_{i-1}} \tag{12}$$

where $\hat{d}_{i-1} = \sum_{z=1}^{n_r} d_{i-1:z}$ is a variable for summation of all repeats for the $i-1$'th time point. Each $\hat{d}_j$ is distributed gaussian with mean $m_j^{j+1} = n_r \sum_{m=j,\, t_m \geq s_g}^{j} 1$ and standard deviation $\sigma_j^{j+1} = \sigma\sqrt{n_r}$. The gaussians are independent of each other over interval $[-\infty, \frac{n_r}{2}]$, so this becomes:

$$p\big(\mathcal{L}(t_i,1) > \mathcal{L}(t_j,1),\, \forall j \in S_i\big) = A + \int_{-\infty}^{\frac{n_r}{2}} p(\hat{d}_{i-1}) \int_{\frac{n_r}{2}}^{n_r - \hat{d}_{i-1}} p(\hat{d}_{i-2}) \cdots \int_{\frac{n_r}{2}}^{n_r \frac{i-1}{2} - \sum_{t=2}^{i-1}\hat{d}_t} p(\hat{d}_1)\, d_{\hat{d}_1} \cdots d_{\hat{d}_{i-2}} d_{\hat{d}_{i-1}} \tag{13}$$

where $A = \prod_{j \in S_i} \Phi(\frac{n_r}{2}, m_j^{j+1}, \sigma_j^{j+1})$. Eq. 13 can be efficiently estimated by Gaussian quadrature or by MCMC [10]. We use similar derivation to estimate $p\big(\mathcal{L}(t_i,1) > \mathcal{L}(t_j,1),\, \forall j \in M_i\big)$. For large T^d, exact estimation of nested multidimensional integral in Eq. 11 can be complicated so we instead estimate its upper and lower bounds as below.

Estimating upper and lower bounds Exact estimation of nested multidimensional integral in Eq. 11 can be complicated for large T^d. In this case, we can rather estimate its lower and upper bounds quite efficiently. $\prod_{j \neq i} p(\mathcal{L}(t_i,1) > \mathcal{L}(t_j,1))$ gives a lower bound estimate since these pairwise terms are not originally independent. We can estimate an upper bound of $p\big(\mathcal{L}(t_i,1) > \mathcal{L}(t_j,1),\, \forall j \in S_i\big)$ as follows: Let $D(n)$ be upper bound of the integral defined only by the topmost n equations in (9). By approximating the multi-dimensional integral symmetrically, upper bound can be estimated recursively by:

$$D(n+1) = D(n)\Big(1 - \frac{1}{n+1}(1 - A_{i-n-1:i-n})\Big) \tag{14}$$

with base case $D(1) = A_{i-1:i} = \Phi(\frac{n_r}{2}, m_{i-1}^i, \sigma_{i-1}^i)$, and its upper bound is given by $D(i-1) = \frac{\prod_{j=1}^{i-1}(A_{j:j+1}+i-1)}{(i-2)!}$. Similarly, upper bound of $p\big(\mathcal{L}(t_i,1) > \mathcal{L}(t_j,1),\, \forall j \in M_i\big)$ can be estimated by:

$$U(n+1) = U(n)\Big(1 - \frac{A_{i+n:i+n+1}}{n+2-i}\Big) \tag{15}$$

where $U(n)$ is upper bound of the integral defined only by the topmost $n-i+1$ equations in (10), and base case is $U(i) = 1 - A_{i:i+1} = 1 - \Phi(\frac{n_r}{2}, m_i^{i+1}, \sigma_i^{i+1})$. Solution of this recursion

is $U(T-1) = \frac{\prod_{j=i}^{T-1}(j-i+1-A_{j:j+1})}{(T-i)!}$. Let I^d be the vector points in T^d ordered by their absolute distance from s_g. Once upper bound of Eq. 11 is estimated, we can estimate the corresponding lower bound of $E(f_{\text{mis}})$ by Algorithm 1. Upper bound of $E(f_{\text{mis}})$ can be estimated similarly by the same algorithm where we use lower bound of Eq. 11 instead of its upper bound estimation in Lines $5-9$.

Algorithm 1 An algorithm for computing a lower bound for $E(f_{\text{mis}})$.

1: $r = 1$, $d = 0$ {r is the remaining probability mass, d is the expected distance}
2: Let I be an ordering of the points in T w.r.t. their distance from s_g
3: **while** $I \neq \emptyset$ **do**
4: $t_i \leftarrow$ first point in I; $I = I \setminus t_i$
5: $lb_i = 1$
6: **for** $t_j \in I$ **do**
7: $c_j = P(\mathcal{L}(t_i, 1) > \mathcal{L}(t_j, 1))$
8: $lb_i = lb_i c_j$
9: **end for**
10: $d = d + r lb_i(|t_i - s_g|)$
11: $r = r(1 - lb_i)$
12: **end while**
13: return d