Modeling temporally-regulated effects on distributions
by
Jonas Weylin Mueller
Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2015
© Massachusetts Institute of Technology 2015. All rights reserved.

Signature redacted
Department of Electrical Engineering and Computer Science
May 20, 2015
Certified by.........................
Tommi S. Jaakkola
Professor of Electrical Engineering and Computer Science
Thesis Supervisor
Certified by........
David K. Gifford
Professor of Electrical Engineering and Computer Science
Thesis Supervisor
Accepted by........
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students
Modeling temporally-regulated effects on distributions

by

Jonas Weylin Mueller

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

We present a nonparametric framework for modeling an evolving sequence of (estimated) probability distributions which distinguishes the effects of sequential progression on the observed distribution from extraneous sources of noise (i.e. latent variables which perturb the distributions independently of the sequence-index). To discriminate between these two types of variation, our methods leverage the underlying assumption that the effects of sequential-progression follow a consistent trend. Our methods are motivated by the recent rise of single-cell RNA-sequencing time course experiments, in which an important analytic goal is the identification of genes relevant to the progression of a biological process of interest at cellular resolution. As existing statistical tools are not suited for this task, we introduce a new regression model for (ordinal-value, univariate-distribution) covariate-response pairs where the class of regression-functions reflects coherent changes to the distributions over increasing levels of the covariate, a concept we refer to as trends in distributions. Through simulation study and extensive application of our ideas to data from recent single-cell gene-expression time course experiments, we demonstrate numerous strengths of our framework. Finally, we characterize both theoretical properties of the proposed estimators and the generality of our trend-assumption across diverse types of underlying sequential-progression effects, thus highlighting the utility of our framework for a wide variety of other applications involving the analysis of distributions with associated ordinal labels.

Thesis Supervisor: Tommi S. Jaakkola
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: David K. Gifford
Title: Professor of Electrical Engineering and Computer Science
Acknowledgments

I am eternally thankful to my parents and Laura for the endless love and support as well as the countless times they have gracefully beared-with my odd hours and neuroses. You are my greatest sources of happiness and inspiration. I also have the great fortune of being part of the best lab groups; from hunting for free food, Mondays at the Muddy, engaging in Coup subterfuge, ski trips, and heated neural net vs. kernels debates, the lab has always been a blast, and occasionally even productive as well. My first years at MIT have been awesome thanks to you and the numerous other great folks I can now call my good friends. Last but not least, I am deeply grateful to my advisors Tommi Jaakkola and David Gifford, whose mentorship has transformed my Master’s education into a remarkably enlightening experience. I eagerly look forward to the future research adventures which await.
## Contents

1 Introduction ................................................. 13
   1.1 Single-cell RNA-sequencing .............................. 15
   1.2 Classic regression models ............................... 17
   1.3 Existing methods for analyzing distributions ............ 21

2 Methods ......................................................... 25
   2.1 TRENDS regression ........................................ 25
       2.1.1 The Wasserstein distance ......................... 26
       2.1.2 TRENDS objective .................................. 29
   2.2 Underlying statistical model ............................. 29
   2.3 Characterizing trends in distributions .................... 31
   2.4 Measuring fit and effect size ............................ 39
   2.5 Fitting the TRENDS model ............................... 40
   2.6 Statistical Guarantees ................................... 46
   2.7 Testing for significance ................................ 51

3 Results ......................................................... 55
   3.1 Simulation study .......................................... 55
   3.2 Application to scientific data ............................ 60
       3.2.1 Kolmogorov-Smirnov method (KS) .................... 61
       3.2.2 Mutual information method (MI) ................. 61
       3.2.3 Linear TRENDS (LT) model ....................... 61
       3.2.4 Tobit model (censored regression) .............. 63
3.2.5 Human myogenesis data ........................................ 63
3.2.6 Mouse cortex data ............................................. 64
3.2.7 Results for SCRS time course data .......................... 65
3.2.8 ACS income distributions ................................. 68

4 Discussion .................................................................. 73

A Proofs and auxiliary lemmas ..................................... 81

B Supplementary Tables ............................................. 95
List of Figures

2-1 Examples of four different sequences of distributions which follow a trend. ................................................................. 27
2-2 Examples of distribution sequences which do not follow trends, along with the best-fitted sequence of distributions estimated by TRENDS regression. ...................................................................................... 38
3-1 Wasserstein errors in applications of TRENDS to simulated data . . . 57
3-2 Evaluating TRENDS against the KS and MI methods on a large simulated dataset with 600 samples drawn from four different TRENDS models with varying levels of noise .............................................. 59
3-3 Word clouds created from the terms significantly enriched in the GO annotations of the genes with significantly trending expression . . . . 66
3-4 Comparing different methods based their ability to identify genes with relevant annotations ................................................ 68
3-5 Examples of time course single-cell RNA-seq data for various known development-regulating genes ........................................ 69
3-6 The distributions of reported income in each annual ACS survey of individuals in the “other information services” industry ............. 70
List of Tables

3.1 Comparing our approximate $p$-values ($\hat{p}$) against those produced by the permutation test ($p_{perm}$) .................................................................................. 58

3.2 The top 20 industries with annual incomes most affected by temporal progression from 2007-2013 (as inferred by TRENDS) .................. 71

S1 A list of all GO annotation terms containing both the words “muscle” and “development”, used to produce the pseudo-sensitivity plots in Figure 3-4A. .................................................. 96

S2 A list of the GO annotation terms relevant to the somatosensory cortex data, used to produce the pseudo-sensitivity plots in Figure 3-4B. . . 97
Chapter 1

Introduction

An common type of data in scientific and survey settings consists of observations sampled in batches, where each batch shares a common label whose effects on the observations are the item of interest (e.g. sampling-time, which we refer to as the covariate). When the observations within a batch are treated exchangeably and per batch sample sizes are large, the empirical distribution of each batch can serve as a good approximation of the underlying distribution of the population conditioned on the value of the covariate (which is shared among all observations from the same batch). A natural goal in this setting is to quantify the covariate's effect on these conditional distributions, considering changes across all segments of the population. This may be particularly useful to identify the most interesting features when the observations are complex and high-dimensional.

Miscellaneous applications beyond the single-cell gene expression time course data focused on in this work might include: (i) comparing temporal trends in the income-distributions of different nations (from historical survey data) to identify countries with the most dramatic changes in a specific time period, (ii) quantifying the change in user-visit-time-distributions on a website vs. the number (or size) of advertisements displayed to help a company evaluate the effects of increasing advertising, (iii) measuring how much a biomarker’s distribution changes over increasing patient-age / prognostic-score / time-until-death (in years) which can be used to identify promising/unreliable biomarkers, (iv) examining how different stocks’ distribution of
daily-price-fluctuations changes subject to the number of stars assigned to them by a rating agency. In each of these examples, modeling only the mean effects is unsatisfactory and it is important to consider the full spectrum of effects across the population, particularly because various subgroups may be affected in different ways.

The primary focus in this thesis is the introduction of the TRENDS (Temporally Regulated Effects on Distribution Sequences) regression model, which may be used to infer the magnitude of these effects across the complete distributions. This is done by imposing restrictions on our regression functions based on the distinct types of covariate - conditional distribution relationships the effects of interest are expected to exhibit. TRENDS may be viewed as an extension of classic regression with a single covariate (typically of fixed-design), where one observation of the dependent variable is a batch’s entire empirical distribution rather than a scalar valued response and the condition of linear fitted-values is replaced by the condition of fitted distributions which follow a trend (an assumption characterizing the effects of interest which will be formalized). Thus, TRENDS goes beyond average-case analysis (modeling only mean-effects) to consider changes across the entire population while retaining the power of regression for distinguishing effects of interest from extraneous noise. Despite the general applicability of our framework, we motivate TRENDS with a concrete example to build intuition, centering our exposition around an important scientific application: the analysis of single-cell RNA-sequencing time course data.

This thesis is structured as follows: The first chapter develops the relevant background and motivation for the statistical problems and application domains under consideration. Chapter two describes the TRENDS model as well as the other methodological contributions of this work including:

- The mathematical notion of a trend in distributions and examples thereof
- The TRENDS regression framework and provably convergent algorithms to fit the model.
- A statistical model describing our assumptions on the underlying data generating process and theoretical bounds characterizing the finite-sample convergence
of the TRENDS estimates.

- A hypothesis test for identifying significant trends in data as well as other useful measures of effect-size based on the TRENDS estimates.

Chapter three demonstrates the application of our ideas in both simulations and in analyses of three different real-world datasets. Finally, chapter four contains a brief discussion summarizing the main ideas presented in this thesis and future research directions.

1.1 Single-cell RNA-sequencing

The recent introduction of single-cell RNA-sequencing (SCRS) techniques to obtain transcriptome-wide profiles of gene-expression levels within individual cells has generated great interest in the biological community [1, 2]. Previously only measurable in aggregate over a whole tissue-sample / culture consisting of thousands of cells, gene-expression at the single-cell level offers insight into biological phenomena at a much finer-grained resolution, and is important to quantify as even cells of the same supposed type exhibit dramatic variation in morphology and function. One promising experimental design made feasible by the advent of this technology involves sampling groups of cells at various times from tissues / cell-cultures undergoing development and applying SCRS to each of cells [2, 3]. It is hoped that this data can reveal both the genes which regulate/mark the emergence of new cell types in the developmental progression, as well as a molecular characterization of what distinct cell subtypes are present in the population based on the observed gene-expression patterns.

Furthermore, the discovery of novel gene-markers of development, i.e. genes which do not drive the developmental process but rather exhibit distinct expression patterns only at specific stages of this process, is also of great interest. Through these genes, one may better characterize cell-populations / stages-of-development and possibly identify key drivers of development by studying the regulators of these marker genes. Due to the inherent difficulty of causal inference for observational SCRS data
(in which the same cell is never profiled at different time-points), we cannot distinguish developmental marker genes from drivers without leveraging extensive prior knowledge. However, we can address the more general task of identifying the that genes influence or are-influenced-by development by searching for genes whose expression patterns exhibit biologically compelling variation over the duration of the time-course. For brevity, we characterize as developmental all genes that either affect or are affected-by development.

Contemporary physical constraints and high costs associated with SCRS technology make it infeasible to densely sample and sequence individual cells continuously across the entire time-continuum. Instead, researchers target a small number of time-points of interest, and at each of these times, a set of cells is simultaneously isolated from the population and RNA-sequencing transcriptome-profiles are subsequently generated for each individual cell. More concretely, from a cell population undergoing some biological transformation such as development, one samples $N_{\ell} \geq 1$ batches of cells (of size $n_i$) from the population at time $t_{\ell}$ where $\ell = 1, 2, \ldots, L$ indexes the time-points in the experiment and $i = 1, \ldots, N = \sum_{\ell=1}^{L} N_{\ell}$ indexes the batches. Each batch consists of cells sampled and sequenced together (the Fluidigm C1 platform, which can hold 96 individual cells, is the favored cell-capture system for SCRS experiments [1-4]), and generally, only one cell-capture plate is employed per time point due to the high costs of RNA-sequencing (thus $N_{\ell} = 1$ and $n_i \approx 90$ as some cells may be omitted from analysis due to damage during sampling or poor sequencing quality [2-5]). We can thus denote by $x^{(g)}_{i,s}$ the measured expression-value of gene $g$ in the $s$th cell ($1 \leq s \leq n_i$) of the $i$th batch, which has been captured from the population at time $t_{\ell_i}$.

Depictions of such data where $N_{\ell} = 1$ for each $\ell$ are shown in the lefthand panels of Figure 3-5. Note that this data differs from that studied in time series analysis because at each time point, one observes a different group of numerous exchangeable samples (no cell is profiled in two time points), and also the number of time points is small (generally $L < 10$). As the price of RNA-sequencing falls, multiple cell-capture plates (each producing a batch of sampled cells) are being used at each
time point to observe larger fractions of the cell population (implying $N_t > 1$) [6]. Rather than treating the cells from a single time point identically, it is desirable to retain the batch information to account for the possible presence of technical artifacts which may affect each capture plate differently. By fitting a TRENDS model to each gene's expression values, researchers can obtain a ranking of the genes based on their presumed relevance to development, as well as a powerful statistical test to determine whether the observed temporal variation in expression is biologically relevant.

1.2 Classic regression models

We now discuss why current statistical methods are inadequate for analyzing SCRS time course experiments, while gradually building up the underlying ideas behind TRENDS. Reflecting back on these arguments in our subsequent presentation of TRENDS will highlight the strengths of our proposed framework and how it addresses the shortcomings of these methods. For simplicity, we ignore the possible stratification into multiple batches per time-point in this discussion. Recall that the data in this application differs from standard time-course measurements in which the *same* subjects are repeatedly measured over time, so samples may be treated as time-curves (in SCRS experiments, *different* cells are profiled at each time point). Nevertheless, one might employ a curve-based method for analysis, studying for each gene $g$, the function of average-$g$-expression (across cells) vs. time. This provides a simple way to rank genes based on their involvement in the developmental process by positing a separate regression model for each gene $g$ of the form:

$$E\left[ x_{t,s}^{(g)} \right] = f_g(t_{u}) \quad \text{for all times } t_1, \ldots, t_L, \text{ where each } f_g \in \mathcal{F}$$

and measuring for each gene $g$ how much of the variation in the observed expression values is captured by the regression function $f_g$. Here, $\mathcal{F}$ is a (mathematically tractable) class of regression functions chosen by the researcher to capture the assumed effects temporal progression (e.g. the course of development) can have on $g$-
expression values. For example, in an analysis of tissue-sample gene-expression over
time, Bar-Joseph et al. have assumed $F$ consists of B-splines [7]. In this context,
deviations of the observed $x_{i,s}^{(g)}$ from their expectation (or the fitted values under $f_g$)
are generally interpreted either as noise or the effects of other latent variables.

Typically, the scale of these time-course experiments is chosen to be fairly large
relative to the duration of oscillations in expression that occur due to routine cellular
process such as the cell-cycle, so that the effects of development may be are not heav-
ily confounded with these standard cellular routines. However, because expression
profiles are obtained from a sparse set of time points due to cost/labor constraints,
the underlying biological rate of development might drastically differ between equally
spaced time points. Thus, changes in the expression of genes involved in different
parts of this process may be highly nonuniform over time potentially invalidating as-
sumptions such as linearity/smoothness of $f_g(\cdot)$ over $t_1, \ldots, t_L$. One common solution
for this problem in tissue-level RNA-seq data is the warping of time-scales, as used
in [7]. Because our primary focus is not on predicting gene-expression at new time-
points or interpolating between $t_t$ and $t_{t+1}$ (rather, our goal is to identify genes whose
expression changes over time in a way that is related to development) and we aim for
a procedure that respects the sequence of times without being overly sensitive to their
precise values (rates of development vary between different individuals anyway), we
simply treat sampling time as an ordinal variable, so the only information retained
about the time-points is their order $t_1 < t_2 < \cdots < t_L$ (as has been done before for
example in [8] Section 2.3.2).

In many RNA-seq studies of developmental progression and cell differentiation,
researchers entirely disregard the wall-clock time at which sequencing is done, instead
recording the time-course chronology as a sequence of stages corresponding to the
overall qualitative state of the tissue / cell-population at the time transcriptome-
profiling [9–11]. For example, in [10], Stage 1 is the oocyte, Stage 2 the zygote, \ldots,
Stage 11 the late blastocyst. Attempting to impose a common scale on the stage
numbering is a difficult task because the similarity in $g$-expression expected across
different pairs of adjacent stages might be highly diverse for different $g$. In this work,
we circumvent this issue by entirely disregarding the time-scale and the $t_t$. Thus the regression functions in (1.1) may be replaced with new $f_g : \{1, \ldots, L\} \to \mathbb{R}$ which only consider ordinal levels $\ell$ (where $\ell_1 < \ell_2 \iff t_{\ell_1} < t_{\ell_2}$), resulting the following model of $g$-expression in each cell.

$$\mathbb{E} [x_{i,s}^{(g)}] = f_g(\ell_i) \text{ for all time-levels } \ell_i \in \{1, 2, \ldots, L\}, \text{ where each } f_g \in \mathcal{F}$$

One reasonable restriction on $\mathcal{F}$ might be that it only contains monotonic functions over $\ell$. If the overall time-interval $[t_1, t_L]$ is short (i.e. the experiment focuses on a specific aspect of the development process), we expect the main developmental effects to stem from genes which are on average turned on/off over time (either gradually or suddenly) rather than genes which (on average) switch expression-mode over a brief intermediate interval and then revert back to their original state or exhibit even more complicated temporal fluctuation between high and low expression. Even if there are biologically tenable effects which are not monotonic, this restriction may nonetheless be desirable due to statistical considerations: oscillatory/wildly-fluctuating time-effects on observed $g$-expression are nearly impossible to distinguish from the profuse noise in SCRS data at the moderate sample-sizes / number of time points typically used in such studies. Thus, attempts to model more diverse effects may cause a decrease in power so drastic that this increase in generality produces inferior results than the restriction to monotonic time-effects.

Unfortunately, average-case analyses as presented above fail to adequately account for heterogeneity among cells, and results must be interpreted as the expected effect of time on $g$-expression in a "typical" cell. In developmental studies, what is generally observed is the emergence/disappearance of different types of cells over time [2, 3]. Thus, at each time point, the cells sampled might be of different types, each subtype having a different distribution of measured $g$-expression for the genes $g$ which characterize it (we call such a gene subtype-differentiating and the corresponding subtypes $g$-subtypes). It has been discovered that there is significant heterogeneity in gene expression between cells: a cell-population's average expression-level is of-
ten strongly biased by a few high-expression cells, and levels of specific transcripts have also been observed to vary as much as 1,000-fold between seemingly equivalent cells [2, 4, 5, 12, 13]. Attempting to characterize the developing cells by a single average-expression value is therefore a crude (and possibly misleading) reduction. Both Geiler-Samerotte et al. [13] and Islam et al. [5] also criticize average-based methods lamenting: "... analyzing gene expression in a tissue sample is a lot like measuring the average personal income throughout Europe – many interesting and important phenomena are simply invisible at the aggregate level'. Even when phenotypic measurements have been meticulously obtained from single cells or individual organisms, countless studies ignore the rich information in these distributions, studying the averages alone”.

In typical time-course single-cell experiments, the genes of interest are those for which there is an emerging/disappearing subpopulation of cells which exhibits different expression of these genes than the rest of the cells. For example, there might be a single type of cell present at time 1 which always expresses $g$ around a level of 5 (in some units quantifying gene-expression), but at a later time, the population has evolved into a mixture of 50% cells which highly express $g$ at a level of approximately 10 and 50% cells which do not express $g$ at all. Such a gene cannot be found by average-based methods. While it might be argued that the most important developmental genes will be those whose for which the entire $g$-expression distribution tends to shift up or down over time, quantifying this shift solely based on the change in average-expression may nonetheless remain inadequate in this case, as it fails to account for characteristics of interest such as the proportion of cells exhibiting the shift in $g$-expression. In practice, biologists will review a ranking of the genes based on time-effect to identify those which warrant further study (typically paying less attention to the particular values of inferred time-effects). Even though average-case analysis may produce roughly reasonable values for individual genes in the case of overall up/down $g$-expression-shift, the corresponding overall gene-ranking can be highly suboptimal when many genes have similar estimated mean-effects, but they differ in other ways that can only be accounted for by considering other characteristics.
of the population. To overcome the shortcomings of average-case analysis, quantile regression has been proposed as a promising alternative [14,15] and TRENDS builds on ideas from this area.

1.3 Existing methods for analyzing distributions

Evidence for the types of phenomena undiscoverable by average-case analysis can be captured by modeling the entire distribution of $g$-expression in the sampled cells over time. Letting $\hat{P}_{\ell}^{(g)}$ denote the empirical distribution of $g$-expression in the cells sampled at time point $\ell$, a standard approach would be to test for distribution-equality, i.e. $H_0 : P_1^{(g)} = \cdots = P_L^{(g)}$ against the alternative that there exists at least two time points at which the cell population has different $g$-expression distributions (the notation without the hat refers to the true expression distribution – conditioned on $\ell$ – over the entire cell population). Such nominal treatment of the different time points however fails to reflect the fact that the overall population should look more similar at adjacent time points and statistical testing fails to provide an interpretable measure of the size of the effect that time has on $g$-expression.

A natural alternative to this omnibus approach in which $\ell$-values are suboptimally treated as exchangeable is to model the $P_\ell$ as conditional distributions $\Pr(X \mid \ell)$ which follow some given structure over varying $\ell$. Work in this vein has primarily focused on accurate estimation of the conditional distributions [16–18] (or often even only of a few particular quantiles of interest [14,15]) as well as their visualization [19,20]. In particular, these goals have resulted in kernel-density based approaches and smooth nonparametric regression models [14,15,17–20] gaining great popularity in this area. While such approaches lead to estimators with great properties, the relationships they uncover may be opaque and it is not entirely clear how to best measure the effect-size of covariates on the full distribution when employing for example a regression model with functions that live in Reproducing kernel Hilbert spaces [15]. Despite this, methods to quantify the magnitude of effects are emerging in this setting such as the mutual-information based DREMI measure of [20].
Unfortunately, when noise (or latent-variable effects) are rampant, powerful measures such as mutual information become highly susceptible to the spurious effects which inevitably corrupt the observed distributions. This is a problem in applications with possible batch effects. Because cells in SCRS data are collected and sequenced in separate batches at different times, the measured gene-expression at one time point might be biased by technical artifacts which affect that batch alone [21, 22]. Other examples where this issue is problematic include studies of the evolution of demographic statistics such as income, where it is difficult to ensure that survey procedures at each time point do not suffer from bias due to location-constraints or other latent variables. Batch effects can cause powerful methods capable of identifying arbitrary differences in \( \Pr(X | \ell) \) to spuriously identify genes with noisy measurements, and we thus prefer borrowing strength approach in which a consistent change in distribution would ideally be observed in the data from multiple time points for a gene to be deemed significant.

Here we draw an analogy to regression in which restrictions of linearity are the tool of choice for measuring effect-sizes throughout the sciences, despite the long-time existence of far more expressive tractable function classes which often produce superior estimators (e.g. nonparametric regression). One possible explanation for this phenomenon is that all of the problems in which precise quantification of effects is interesting to begin with are inherently noisy for one reason or another. Thus, when analyzing effects on full distributions, we surmise that (just as for the easier problem of quantifying effects on conditional expectations studied in classic regression): transitioning to more expressive/general models, while theoretically enabling capture of a wider diversity of effects, will for the interesting problems encountered in real applications likely lead to drastic decreases in power due to the model's ability to overfit spurious variation (Figure 3-2A demonstrates how severe this issue may be). Thus we develop TRENDS with the noisy regime in mind, endowing our model with precisely enough flexibility to capture all of the types of effects we might hope to identify, while simultaneously ensuring its interpretability and robustness to noise.

Change-point analysis is a third approach statisticians might consider when pre-
sented with a sequence of distributions [23]. While a multitude of different methods have been proposed in this area, much of the work is primarily focused on detecting the precise location(s) of the change-point(s) in a long temporal sequence, whereas SCRS experiments typically span a brief time-course with \( L \leq 10 \), and the primary analytic goal is to quantify how much a gene’s expression has changed over time in a biologically interesting manner. Many change-point methods additionally require explicit parameterization of the types of distributions to consider, an undesirable necessity given the irregular skewed nature of SCRS expression measurements [22]. Moreover, assuming the cells are sampled at reasonably dense time-intervals with respect to the underlying biological processes, one expects many of the development-related genes to exhibit gradual rather than abrupt temporal changes in expression. TRENDS is capable of identifying both abruptly and gradually changing genes without favoring either type, and requires few statistical assumptions, incorporating only prior knowledge on how the biologically-relevant effects of temporal progression are expected to manifest themselves in the expression data of developmental genes.
Chapter 2

Methods

2.1 TRENDS regression

Formally, TRENDS fits a regression model to an ordered sequence of distributions, or more generally, sample pairs \( \{(\ell_i, \hat{P}_i)\}^N_{i=1} \) where each \( \ell_i \in \{1, \ldots, L\} \) is an ordinal-valued label associated with the \( i \)th batch for which we observe univariate empirical distribution \( \hat{P}_i \). Assuming a fixed-design, we require each possible level of the covariate \( 1, \ldots, L \) to be observed for at least one of the batches. In the SCRS application, \( \hat{P}_i \) is the empirical distribution of a single gene’s measured expression values over the cells captured in a single batch and the label \( \ell_i \) indicates the index of the time point at which the batch was extracted from the population for sequencing.

Unlike the typical supervised learning framework where one observes samples of \( X \) measured at different \( \ell \) and the goal is to infer some property of \( P_\ell := \Pr(X|\ell) \), in our setting, we directly observe an empirical estimate of this conditional distribution at each \( \ell: \hat{P}_i \approx \Pr(X|\ell_i) \). We thus neither seek to estimate the distributions \( P_1, \ldots, P_L \), nor test for inequality between them. Rather, the primary goal of TRENDS analysis is to infer how much of the variation in \( \Pr(X | \ell) \) across different \( \ell \) is due to increasing-\( \ell \) as opposed to the effects of other unmeasured factors. To quantify this variation, we introduce conditional effect-distributions \( Q_\ell \) for which the sequence of transformations \( Q_1 \rightarrow Q_2 \rightarrow \cdots \rightarrow Q_L \) entirely captures the effects of \( \ell \)-progression on \( \Pr(X | \ell) \), under the reasonable assumption that these underlying forces follow a trend.
Definition 1. Let $F^{-1}_\ell(p)$ denote the $p$th quantile of distribution $P_\ell$ with CDF $F_\ell$.

A sequence of distributions $P_1, \ldots, P_L$ follows a trend if:

1. For any $p \in (0, 1)$, the sequence $[F^{-1}_1(p), \ldots, F^{-1}_L(p)]$ is monotonic.

2. There exists $p^* \in [0, 1)$ and two intervals $A, B$ that partition the unit-interval at $p^*$, so one of $A$ or $B$ equals $[0, p^*]$ and the other interval equals $(p^*, 1]$, such that:
   - for all $p \in A$, the sequences $[F^{-1}_1(p), \ldots, F^{-1}_L(p)]$ are all nonincreasing, and
   - for all $q \in B$, the sequences $[F^{-1}_1(q), \ldots, F^{-1}_L(q)]$ are all nondecreasing. Note that if $p^* = 0$, then all quantiles must change in the same direction over $\ell$.

Thus, the $Q_\ell$ are not defined as estimators of the sequence of $P_\ell$, where variation is also induced by the effects of other exogenous variables, for example expression-measurement error (even if we sampled the entire population of cells, the corresponding distribution over measured-expression values would likely still change over time if the cell-population were not changing because the types of measurement error prevalent at different times varies). Rather, the $Q_\ell$ represent the distributions one would expect see in the absence of exogenous effects and random sampling variability, in the case where the underlying distributions only change due to $\ell$-progression and we observe the entire population at each $\ell$. Because we do not believe exogenous effects unrelated to $\ell$-progression are likely to follow a trend over $\ell$, we can identify the sequence of trending distributions which captures the maximal amount of variation in the $P_\ell$, and conclude that the changes in this sequence reflect the $\ell$-progression-related forces affecting $P_\ell$.

2.1.1 The Wasserstein distance

TRENDS employs the Wasserstein distance as a divergence-measure between distributions [24]. This metric (also known by the names: Kantorovich, Mallows, Dudley, or Earthmover) has the nice intuitive interpretation as the minimal amount of "work" that must be done to transform one distribution into the other (where work equals the amount of probability mass moved times distance moved). The Wasserstein
How each quantile changes for $\ell = 1, 2, 3$:

- Increasing
- Constant
- Decreasing
- Non-monotonic

Figure 2-1: Violin plots (i.e. vertically-drawn mirrored density curves) depicting four different sequences of distributions which follow a trend. The $p$th rectangle in the color bar on the righthand side indicates the monotonicity of the $p$th quantile over the sequence of distributions (for $p = 0.01, 0.02, \ldots, 0.99$): green indicates it is monotone nondecreasing as $\ell$ grows, blue indicates monotone nonincreasing, black indicates a quantile which does not change over $\ell$, and red indicates the quantile is not monotone over the sequence. Note that it follows directly from the definition that no red may appear in the color bars associated with these trending distribution-sequences.
distance provides a natural dissimilarity measure of cell populations with respect to
g-expression, because it accounts for the proportion of cells which are expressing gene
g differently as well as how different these cells’ g-expression is. Furthermore, in
accordance with Occam’s razor, it is natural to assume that the forces which drive
the change $Q_t \rightarrow Q_{t+1}$ do so in the most straightforward manner, and thus this transformation can be measured using the Wasserstein distance.

To quantify the fit of the model, we use the $L_2$ Wasserstein metric, which in the
case of univariate distributions is simply the $L_2$ distance between quantile functions [24] given by:

$$d_{L_2}(P, Q) = \left( \int_0^1 [F^{-1}(p) - G^{-1}(p)]^2 \, dp \right)^{1/2} \quad (2.1)$$

where $F, G$ are the CDFs of $P, Q$ and $F^{-1}, G^{-1}$ are the corresponding quantile functions, defined for $0 < p < 1$ as:

$$F^{-1}(p) = \inf\{x \in \mathbb{R} : p \leq F(x)\} \quad (2.2)$$

In one-dimension, the $L_2$ Wasserstein metric is easy to compute and comes equipped
with a natural metric space of quantile functions, in which the Fréchet mean (which we call the Wasserstein mean) takes the simple form stated in Lemma 2. We also employ the $L_1$ Wasserstein metric in this work in order to measure the how much the distributions change as a result of sequence-progression:

$$d_{L_1}(P, Q) = \int_0^1 |F^{-1}(p) - G^{-1}(p)| \, dp \quad (2.3)$$

To interchangeably discuss both quantile functions and distributions, we overload the $d_{L_p}(. , . )$ operators for $p = 1, 2$ to both denote Wasserstein distances between distributions (which equal the distance between the quantile functions corresponding to these distributions), as well as distances between quantile functions in $L_p$ space.
2.1.2 TRENDS objective

TRENDS finds a fitted distribution $\hat{Q}_\ell$ for each $\hat{P}_\ell$ using the Wasserstein-least-squares-fit which minimizes the following objective function:

$$
\hat{Q}_1, \ldots, \hat{Q}_L = \arg \min_{Q_1, \ldots, Q_L} \left\{ \sum_{\ell=1}^{L} \sum_{i \in I_{\ell}} d(Q_\ell, \hat{P}_i)^2 \right\}
$$

where $Q_1, \ldots, Q_L$ follow a trend

(2.4)

where $I_{\ell}$ is the set of batch-indices $i$ such that $\ell_i = \ell$, and it is assumed $N_{\ell} := |I_{\ell}| \geq 1$ for all $\ell \in \{1, \ldots, L\}$. Subsequently, an analyst can inspect changes in the $\hat{Q}_\ell$ which serve as good representations for the transformations in the underlying $P_\ell$ that are likely caused by increasing-$\ell$. Figure 2-2 shows a few examples of the fitted distributions produced by TRENDS.

If the $\hat{P}_i$ are estimated from a widely varying number of samples for different batches $i$, then it is preferable to replace the objective in (2.4) with a weighted sum as in (2.11), with weights determined based on $N_{\ell}$ and $n_i$, the number of samples used to estimate $\hat{P}_i$. This spurs the TRENDS-fit to better model the variation in the empirical distributions which are likely more accurate because of the larger sample sizes used in their estimation, while enhancing the model’s robustness to the effects of $\hat{P}_i$ which are poor approximations due to their reliance on small samples. Because the $n_i$ (number of observations in batch $i$) and $N_{\ell}$ (number of batches from level $\ell$) are fairly homogeneous in typical SCRS experiments, we do not employ nonuniform weights in this work. Nevertheless, we provide an algorithm which minimizes the more general TRENDS objective in (2.11).

2.2 Underlying statistical model

The intrinsic model behind TRENDS supposes that for each batch $i$, the observed empirical quantile function (EQF) $\hat{P}_{i-1}$ is estimated from $n_i$ exchangeable scalar observations $\{X_{i,s} \sim P_i\}_{s=1}^{n_i}$ sampled from underlying distribution $P_i = \Pr(X \mid \ell_i)$. Furthermore, the quantile function of the underlying $P_i$ is assumed to be generated
by the model:

\[ F_i^{-1} = f(\ell_i) + \mathcal{E}_i \]

where \( f : \{1, \ldots, L\} \to Q \) s.t. \( f(1), \ldots, f(L) \) follow a trend (2.5)

Here \( f \), which captures the effect of level-progression on the underlying distributions, is a mapping from ordinal levels \( \{1, \ldots, L\} \) into the space of valid quantile functions \( Q \), so \( f(\ell) \) represents the quantile function of \( Q_{\ell} \) as previously defined. The random noise functions \( \mathcal{E}_i : (0,1) \to \mathbb{R} \) represent measurement-noise as well as the effects of other unmeasured variables on the underlying batch distributions \( P_{1,h} \) from which our observations are drawn. Note that the form of the \( \mathcal{E}_i \) is implicitly constrained to ensure both \( F_i^{-1} \) and \( f(\ell_i) \) are valid quantile functions. Also, the effect of one \( \mathcal{E}_i \) may manifest itself in multiple observations \( X_{i,s} \), even under the assumption that these observations are drawn i.i.d. from \( P_i \) (e.g. in SCRS data, a batch effect may cause all of the \( g \)-expression values from one cell-capture-plate to be underestimated for some gene \( g \)). We additionally assume: \( \mathbb{E}[\mathcal{E}_i(p)] = 0 \) \( \forall p \in (0,1) \), and \( \mathcal{E}_i(p) \) is statistically independent of \( \mathcal{E}_j(q) \) \( \forall p,q,i \neq j \), so that the exogenous noise is unrelated to \( \ell \)-progression.

Our model can be viewed as the distribution-valued analog of the usual regression model, which assumes scalar \( Y_i = f(X_i) + \epsilon_i \) with \( \mathbb{E}[\epsilon_i] = 0 \) and \( \forall i \neq j : \epsilon_i \) is independent of \( \epsilon_j \). Just as the \( Y_1, \ldots, Y_N \) are rarely believed to exactly lie on the curve \( f(x) \) in the classic scalar-response model, we do not expect that the observed empirical distributions will exactly follow a trend (even if \( n_i \to \infty \) \( \forall i \) so that \( \widehat{P}_i = P_i \)). Rather our model simply encodes the assumption that the effects of level-progression on the distributions should be consistent over different \( \ell \) (i.e. the effects follow a trend). Drawing further parallels between the squared Wasserstein loss used in TRENDS and the standard squared-error-loss of scalar-valued regression, it is thus natural to seek the Wasserstein-least-squares-fit of \( \widehat{P}_1, \ldots, \widehat{P}_L \) over all trending distributions as suggested in (2.4). One can also view the \( P_i : i \in I_\ell \) as random measures concentrated around \( Q_{\ell_i} \), which serves as a sort of base distribution. Random measures have recently been used in numerous applications, primarily under the nonparamet-
ric Bayesian framework [25]. In contrast to this setting, where the base distribution is usually a pre-specified prior, the $Q_\ell$ are the primary inferential interest in TRENDS.

In SCRS data, the expression measurements are typically distorted by significant batch effects and other technical artifacts, so $\mathcal{E}_i$ is certainly non-negligible (there are even biological sources of noise due to processes unrelated to development such as transcriptional bursting and cell-cycle-modulation of expression [2]). Unlike the development-related biological variation in a developmental gene $g$'s underlying expression which we expect to follow a trend over a short time-course, both biological and technical sources of noise are highly unlikely to follow any sort of trend, and it is reasonable to assume independence of the noise functions between different batches of simultaneously captured and sequenced cells. While an arbitrary gene $g$'s measured expression distributions may be highly variable over time, it is probable that this variability is due to external factors unrelated to $g$'s role in development, unless the changes in the observed $g$-expression distribution are consistent with our notions of how a developmental relationship would manifest itself in the data (which we believe is accurately reflected by our trend definition). The TRENDS model thus provides a tool that goes beyond average-case analysis, while remaining robust to the noise rampant in these applications.

2.3 Characterizing trends in distributions

Definition 1 formally characterizes the notion of a trend and Figures 2-1 and 2-2 illustrate examples of distribution-sequences which do and do not follow trends. These examples demonstrate that this definition is consistent with our visual intuition regarding when a sequence of distributions is or isn't evolving in the same general direction/fashion. The trend assumption enforces this notion by ensuring that the temporal-forces which contribute to the transformation from $P_1$ to $P_L$ do so without reversing their effects or leading to wildly different distributions at intermediate values of $\ell$. Note that the general trend definition includes each of the three cases: (i) $F_1^{-1}(p) \leq \cdots \leq F_L^{-1}(p)$ for all $p \in [0, 1]$, (ii) $F_1^{-1}(p) \geq \cdots \geq F_L^{-1}(p)$ for all $p \in [0, 1]$, 

31
(iii) there exist \( p_1 \neq p_2 \) s.t. \( F_1^{-1}(p_1) \leq \cdots \leq F_L^{-1}(p_1) \) and \( F_1^{-1}(p_2) \geq \cdots \geq F_L^{-1}(p_2) \).

The trend assumption may be conceptually rationalized via different perspectives. One can view the quantiles as different segments of a population whose values are distributed according to \( \Pr(X \mid \ell) \) (e.g. in the case of wealth, it is has become common to differentiate between the so-called “one percent”, which correspond to the highest quantiles of the wealth-distribution, from the rest of the population). Under this viewpoint, it is reasonable to assume that while the forces of sequential progression may have different effects on the groups of individuals corresponding to different segments of the population, their effects on a single segment should be consistent over the sequence. If some segment’s values initially grow in one direction at lower levels of \( \ell \) and subsequently begin reverting in the opposite direction once \( \ell \) nears \( L \) (i.e. the corresponding quantile is non monotone), it is natural to conclude there are actually two different progression-related forces affecting this homogeneous group of individuals. However, in many settings such as the SCRS developmental experiments, the researcher has a priori interest in a specific force whose effects he would like to measure (such as a specific stage of the developmental process). In this case, the data is collected so that the range of the \( \ell \) approximately spans this force’s duration, and we thus expect the primary effects of interest to follow a trend.

This rationale is evident in the wealth-distribution analogy. The rebound from the 2007-9 recession is an example of a coherent force driven by federal policy whose effects on the “one percent” were staggeringly different than the rest of the population, with this population segment’s income rising orders of magnitude more than any other group’s [26]. Note that if one were to include earlier time points during / preceding the recession in such a study as well, then the one percent’s income falls and then rises violating our trend assumption, but this is only due to the entirely separate forces driving the recession such as the subprime mortgage crisis, whose effects should not be jointly measured with the effects of recovery policies.

The second part of the trend definition which specifies that adjacent quantiles must move in the same direction over \( \ell \) except for at \( p^* \) restricts the number of population-segments which can increase over \( \ell \) when a nearby segment of the population is
decreasing over \( \ell \). While this does limit a trend from fully capturing highly-segmented effects such as Figure 2-2C, there are few applications where the primary interest in on such phenomena, and accurate modeling of such fine-variation is impossible without immense sample-sizes. Furthermore, the restriction on the number of \( p \)-intervals between which the derivative of the quantile-sequence is allowed to change sign provides computational advantages leveraged by our estimation procedure, as well as effective regularization against spurious quantile-variation (which is especially prominent in noisy SCRS measurements or other phenomena affected by confounding variables).

Despite imposing a condition on every quantile of the distributions, the trend assumption remains very general and does not require: explicit parameterization of the distributions under consideration, specification of a precise functional form on the types of change that may be attributed to sequential progression, or reliance on the assumption that the amount of distribution-change must be smooth/relatively-constant between different sequence levels. These attributes of the trend criterion are highly desirable for modeling a developmental gene’s expression over time, where smoothness restrictions and explicit parameterizations of the form of developmental effects may be untenable. We now present examples various examples of distribution-sequences following a trend to demonstrate the broad spectrum of interesting effects which models based on the trend-assumption can detect.

**Example 1.** Any sequence of *stochastically ordered* distributions follows a trend. Recall that one considers random variable \( X_1 \sim P_1 \) less than \( X_2 \sim P_2 \) in the stochastic order (which we denote \( P_1 \preceq P_2 \)) if \( F_1(x) \geq F_2(x) \ \forall x \) (equivalently characterized as \( \Pr(X_1 > x) \leq \Pr(X_2 > x) \ \forall x \) [27, 28]. Thus, the defining characteristic of a trend – the local monotonicity restriction independently applied to each quantile – is more general than imposing a consistent *stochastic ordering/dominance* across the distribution-sequence (either \( P_1 \preceq P_2 \preceq \cdots \preceq P_L \) or \( P_1 \succeq P_2 \succeq \cdots \succeq P_L \)), as this alternative requires that local changes to each segment of the distribution *all* proceed in the same direction.
**Example 2.** Another interesting distribution-sequence which is also follows a trend is the *evolving mixture model*. In many applications, one might suppose each $P_\ell$ is a mixture of the same $K$ underlying subpopulation-specific distributions, where we let $G_k$ denote the CDF of the $k$th subpopulation-specific distribution (mixing component) with $\ell$-dependent mixing proportion $\pi^{(k)}_\ell$, so that each observed distribution can be expressed as:

$$\forall \ell \in \{1, \ldots, L\} : F_\ell = \sum_{k=1}^{K} \pi^{(k)}_\ell G_k \quad \text{where} \; \forall \ k, \ell : \pi^{(k)}_\ell \in [0, 1] , \; \pi^{(K)}_\ell = 1 - \sum_{k=1}^{K-1} \pi^{(k)}_\ell
$$

(2.6)

In this scenario, the $\ell$-effects of interest alter the mixing proportions, so that a fraction of the individuals of one subpopulation transition to become part of another as $\ell$ increases. Equivalently, this implies that the mixing proportion of one component falls while the probability assigned to the other grows by the same amount. To maximize the generality of this example, we avoid imposing a specific parameterization of the $G_k$. Rather, we merely assume these mixture components are stochastically ordered with $G_1 \preceq G_2 \preceq \cdots \preceq G_K$ because subpopulations by definition have distinct characterizations (note $G_1 \preceq G_2 \preceq \cdots \preceq G_K$ is more general than the fully distinct setting in which the $G_k$ have disjoint support).

To formalize the types of migration between subpopulations which would be classified as a trend, we conceptualize a graph $\mathcal{G}$ with vertices $1, \ldots, K$ representing each mixture component. If there is migration from subpopulation $i$ to $j > i$ in the transition between level $(\ell - 1) \rightarrow \ell$ (i.e. $\pi_{\ell-1}^{(i)} = \pi_{\ell}^{(i)} - \Delta$ and $\pi_{\ell}^{(j)} = \pi_{\ell-1}^{(j)} + \Delta$), then directed edges $i \rightarrow (i + 1), (i + 1) \rightarrow (i + 2), \ldots, (j - 1) \rightarrow j$ are added to $\mathcal{G}$ (and in the case where $j < i$, these same edges are added to $\mathcal{G}$, only their direction is reversed). The case in which multiple simultaneous migrations between subpopulations take place between $(\ell - 1) \rightarrow \ell$ is handled more delicately: First, we identify the sequence $\mathcal{S}$ of operations which produces the optimal transformation from mixing proportions vector $[\pi_{\ell-1}^{(1)}, \ldots, \pi_{\ell-1}^{(K)}] \rightarrow [\pi_{\ell}^{(1)}, \ldots, \pi_{\ell}^{(K)}]$, where the only possible operation is to select $k \in \{1, \ldots, K - 1\}$ and enact the simultaneous pair of reassignments $\pi_{\ell}^{(k)} = \pi_{\ell-1}^{(k)} - \Delta; \pi_{\ell}^{(k+1)} = \pi_{\ell-1}^{(k+1)} + \Delta$ for some $\Delta \in [-1, 1]$ whose magnitude is the
cost of this operation. Subsequently, for each operation in $S$, we introduce an edge into $G$ between the corresponding nodes $k$ and $k + 1$ whose direction is specified by the sign of $\Delta$ (edge $k \rightarrow (k + 1)$ if $\Delta > 0$, the reverse edge otherwise).

$G$ is initialized as the empty graph and for $\ell = 2, \ldots, L$, the necessary edges are added to the graph corresponding to the mixing-proportion-fluctuations between $(\ell - 1) \rightarrow \ell$ as described above. Then, the sequence of distributions $P_1, \ldots, P_L$ follows a trend if $G$ contains no cycles after step $L$ and at most one node with two incoming edges. Intuitively, this implies that a trend captures the phenomenon in which the underlying forces of progression that induce migration from one subpopulation to a larger one as $\ell$ increases, do not also cause migration in the reverse direction between these subpopulations at different values of $\ell$. Figure 2-1D depicts an example of the evolving mixture model.

**Example 3.** Our trend definition also encompasses distribution-sequences where the distributions at intermediate values of $\ell$ are monotonic quantile mixtures of $P_1$ and $P_L$, i.e.

$$\forall \ell : F_\ell^{-1} = \omega_\ell F_1^{-1} + (1 - \omega_\ell) F_L^{-1}$$

s.t. $\{\omega_\ell \in [0, 1] : \ell = 1, \ldots, L\}$ form a monotonic sequence \(2.7\)

Quantile mixtures are typically more appropriate than mixture distributions when there is no evident switching mechanism between distributions in the data-generating process [29], and thus, condition (2.7) naturally characterizes the situation in which the underlying forces of interest gradually evolve distribution $P_1$ into $P_L$ over $\ell = 1, \ldots, L$.

The above examples illustrate that our notion of a trend is quite broad encompassing many types of distribution-sequences which reflect various ways we believe forces related to $\ell$-progression might transform the population-distribution $P_\ell$ in certain domains. Despite it’s generality, our definition nonetheless guarantees the following
**Lemma 1.** If distributions $P_1, \ldots, P_L$ follow a trend, then

$$d_{L^1}(P_i, P_j) = \sum_{\ell=i+1}^j d_{L^1}(P_{\ell-1}, P_{\ell}) \quad \text{for all} \quad i < j \in \{1, \ldots, L\}$$

Lemma 1 provides an alternative justification of the trend assumption. Measuring how much the distributions are perturbed between any pair of levels via the $L_1$ Wasserstein metric, the lemma ensures that the sequence of trending distributions does not temporarily jump to some wildly different distribution for only a few intermediate $\ell$-values. If we view the underlying effects of interest as a literal force acting in distribution-space, Lemma 1 implies that this force points in the exact same direction for every $\ell$ (i.e. $Q_1, \ldots, Q_L$ lie along a line in distribution-space under the $L_1$ Wasserstein metric). However, this far more flexible than a linear restriction in the standard sense, because the magnitude of the force (i.e. how far along the line the distributions move) can vary over $\ell$.

Consider for example Figure 2-2A. It is unclear what single force might produce the sequence of distributions in the left panel, but is easy to conclude that a single underlying force which became active at $\ell = 2$ is responsible for the variation in the distributions of the right panel. Concluding that a single force is responsible for the changes in the left panel is much more plausible in Figure 2-2B, and we note that these distributions in fact lie far closer to a sequence which follows a trend. Operating under the assumption that sequential-progression-effects follow a trend, it is thus improbable that we will misattribute the effects of independent latent variables which randomly perturb the intermediate distributions to $\ell$-progression as $\ell$ grows (only in the rare event that the noise itself happens to nearly follow a trend – perhaps because it only affects $P_1$ or $P_L$ – can mis-attribution occur).

Over a single stage of development, we expect changes in the observed single-cell gene-expression distributions to stem from the gradual or sudden emergence/disappearance of different cell subtypes in the population (plus noise and random sampling effects), but we do not expect to see the emergence and subsequent disappearance (or vice versa) of a subtype in such brief durations. Furthermore, it is not evident that
there are clear distinctions between cell subtypes in the early stages of development as cells are not yet differentiated, and thus not only are the relative proportions of different subtypes changing, but the subtypes themselves may transform as well. Thus, we expect that developmental genes’ true underlying (population-valued) $g$-expression patterns are described by a combination of the previously discussed monotonic quantile mixture and mixture model scenarios. The assumption that these genes’ actual underlying $g$-expression follow a single trend over time therefore fits our intuition well, while remaining highly flexible with respect to the specific form of the trend (unlike say a linear model). The proposed trend definition also reflects our uncertainty regarding whether the changes in the cell population occur gradually over time or suddenly for different genes as well as our lack of a good parametric distribution to describe the $g$-expression values at each time.
Figure 2-2: Violin plots depicting different sequences of distributions which do not follow a trend (on left). On the right, the corresponding fitted distributions estimated by TRENDS regression are shown as well as the $R^2$ value of the fit. The $p$th rectangle in the color bar on the right indicates the monotonicity of the $p$th quantile over the sequence of distributions.
2.4 Measuring fit and effect size

Analogous to the coefficient of determination used in classic regression models, we define the Wasserstein $R^2$ to measure how much of the “variation” in the observed distributions $\tilde{P}_1, \ldots, \tilde{P}_N$ is captured by the TRENDS model’s fitted distributions $\tilde{Q}_1, \ldots, \tilde{Q}_N$:

$$R^2 := 1 - \left( \frac{1}{N} \sum_{i=1}^{N} d_{L^2}(\tilde{Q}_i, \tilde{P}_i)^2 \right) / \left( \frac{1}{N} \sum_{i=1}^{N} d_{L^2}(\tilde{P}_i, \overline{F}^{-1})^2 \right) \in [0, 1] \quad (2.8)$$

where now rather than taking squared distances between scalars, we compute squared Wasserstein distances between distributions, and the quantile function $\overline{F}^{-1} = \frac{1}{N} \sum_{i=1}^{N} \tilde{F}_i^{-1}$ is the Wasserstein mean of the observed distributions across all $i$ (note that in our overloaded notation, the summands in the denominator correspond to $L^2$ distances between the quantile function of $\tilde{P}_i$ and $\overline{F}^{-1}$). Note that by Lemma 2, the numerator and denominator in (2.8) are respectively analogous to the residuals and the overall variance used in usual scalar regression models (differences between scalars have simply been replaced by Wasserstein distances between distributions).

**Lemma 2.** The Wasserstein mean is the Fréchet mean in quantile space under the $L^2$ norm, i.e.

$$\overline{F}^{-1} := \frac{1}{N} \sum_{i=1}^{N} F_i^{-1} = \arg \min_{G^{-1} \in \mathcal{Q}} \left\{ \sum_{i=1}^{N} \int_0^1 (F_i^{-1}(p) - G^{-1}(p))^2 \, dp \right\} \quad (2.9)$$

where $\mathcal{Q}$ is space of all quantile functions.

As it is defined in (2.8), $R^2$ measures how “close” the observed $\tilde{P}_i$ are to a trending sequence of distributions, independently of the scaling of $X$. Because the trends-assumption is much more flexible than the restriction to linear functions in the classic scalar-valued regression setting, TRENDS typically produces much larger $R^2$ values than one encounters in linear regression, so this consideration must be kept in mind when interpreting TRENDS fits. Finally, if the denominator in (2.8) is zero (i.e. the $P_i$ are all identical), we trivially define $R^2 = 0$ as there is no variation to be captured.
by the TRENDS model.

In classic linear regression, the slope of the regression line is interpreted as the expected change in the response resulting from a one-unit increase in the covariate (under the linear model). While TRENDS operates on unit-less covariates, we can instead measure the overall expected Wasserstein-change under model (2.5) in the $\hat{P}_i$ over the entire ordinal progression $\ell = 1, \ldots, L$ by computing

$$\Delta := \frac{1}{L} \cdot d_{L_1}(\hat{Q}_1, \hat{Q}_L)$$  

(2.10)

where the subscript indicates we employ the $L_1$ Wasserstein distance defined in (2.3). The $L_1$ Wasserstein distance is more natural in this context, since by Lemma 1, it measures the aggregate difference over each pair of levels of $\ell$, just as the difference between the largest and smallest fitted-values in linear regression may be decomposed in terms of the covariate’s units to obtain the slope of the regression-line. Thus, $\Delta$ measures the raw magnitude of the inferred trend-effect (which depends on the scale of $X$), while $R^2$ quantifies how well the trend-effect explains the variation in the observed distributions (independently of scaling).

### 2.5 Fitting the TRENDS model

Now, we present the trend-fitting (TF) algorithm which finds distributions satisfying

$$\hat{Q}_1, \ldots, \hat{Q}_L = \arg \min_{Q_1, \ldots, Q_L} \left\{ \sum_{\ell=1}^{L} \sum_{i \in I_\ell} w_i \cdot d_{L_2}(Q_{\ell}, \hat{P}_i)^2 \right\}$$

where $Q_1, \ldots, Q_L$ follow a trend

(2.11)

recalling that we’ve defined $I_\ell := \{ i : \ell_i = \ell \}$ and $N_\ell := |I_\ell|$. To fit TRENDS to observations $\{(\ell_i, \hat{P}_{\ell_i})\}_{i=1}^N$ via our procedure, the user must first specify two things:

1. A sequence of probabilities $0 < p_1 < p_2 < \cdots < p_{P-1} < 1$ to serve as quadrature points for numerically evaluating the integral in (2.1) corresponding to the Wasserstein distance between distributions, i.e.

   the points at which we will evaluate the quantile functions of each distribution.
2. A quantile estimator \( \hat{F}^{-1}(p) \) of a CDF \( F \).

Given these two ingredients, the TF procedure solves a numerical-approximation to the constrained (distribution-valued) optimization problem in (2.11). Defining \( p_0 := 2p_1 - p_2 \) and \( p_P := 2p_{P-1} - p_{P-2} \), we employ the following midpoint-approximation of the integral

\[
\min_{G_1^{-1}, \ldots, G_L^{-1}} \left\{ \sum_{\ell=1}^{L} \sum_{i \in I_{\ell}} w_i \sum_{k=1}^{P-1} \left( \hat{F}^{-1}(p_k) - G_{i}^{-1}(p_k) \right)^2 \left[ \frac{p_{k+1} - p_{k-1}}{2} \right] \right\}
\]

where \( G_1, \ldots, G_L \) must follow a trend (2.12)

The minimizer of this problem is not unique because the fitted distributions’ quantiles can take arbitrary values in-between \( p_k \) and \( p_{k+1} \). However, all we require to numerically compute Wasserstein distances between the fitted distributions and empirical CDFs (and hence the \( R^2 \) value) is the values their quantile functions take at probabilities \( p_1, \ldots, p_{P-1} \), which are uniquely determined by (2.12). In practice, the number \( P \) (and location) of the quadrature points (probabilities) should be determined based on the amount of data available for each \( i \), properties of the specific quantile-estimator employed, and aspects of the distributions \( P_i \) of particular interest in the application.

Throughout this work, we simply use \( P - 1 \) uniformly spaced quantiles between \( \frac{1}{P} \) and \( \frac{P-1}{P} \) (with \( P = 100 \)) which provides a full overview of the distributions. Furthermore, we employ a simple quantile estimation procedure, assuming each of the \( n_i \) cells isolated at the \( i \)th capture-plate are i.i.d. samples (based on our prior knowledge of the Fluidigm C1 cell-capture techniques). If the i.i.d. assumption is untenable, then the quantile-estimation procedure should be accordingly adjusted, perhaps following [30]. Note that no unbiased minimum-variance \( \forall p \in (0, 1) \) quantile estimator has yet been proposed, so in this work, we employ the default algorithm used by the quantile function in \( R \) which provides the best approximation of the mode (Type 7 as discussed by Hyndman and Fan [31]). Other quantile estimators produced virtually identical results in our experiments, and it has been noted that there are no practical differences between the widely-used quantile estimators for sensible sample sizes.
Our estimation procedure heavily relies on the efficient Pool-Adjacent-Violators-Algorithm (PAVA) [33]. Given a scalar sequence \( y_1, \ldots, y_L \in \mathbb{R} \), PAVA finds the best-fitting nondecreasing sequence in the least-squares sense, requiring only \( O(L) \) runtime. Although PAVA is criticized for the lack of smoothness in the regression functions it produces from continuous data \( \{(x_i, y_i) \in \mathbb{R}^2\} \), this is irrelevant in its application in TRENDS, where our only interest is in the fitted distributions at each ordinal level \( \ell = 1, \ldots, L \). PAVA finds the best-fit via the following simple steps:

**Basic PAVA Algorithm:**

\[
\min_{z_{\ell}} \sum_{\ell=1}^{L} (y_{\ell} - z_{\ell})^2 \quad \text{s.t.} \quad z_1 \leq \cdots \leq z_L
\]

**Input:** A sequence of real numbers \( y_1, \ldots, y_L \)

**Output:** The minimizing sequence \( \hat{y}_1, \ldots, \hat{y}_L \) which is nondecreasing.

1. Start with the first level \( \ell = 1 \) and set the fitted value \( \hat{y}_1 = y_1 \)
2. While the next \( y_\ell > \hat{y}_{\ell-1} \), set \( \hat{y}_\ell = y_\ell \) and increment \( \ell \)
3. If the next \( \ell \) violates the nondecreasing condition, i.e. \( y_\ell < \hat{y}_{\ell-1} \), then backaverage to restore monotonicity: find the smallest integer \( k \) such that replacing \( \hat{y}_{\ell}, \ldots, \hat{y}_{\ell-k} \) by their average restores the monotonicity of the sequence \( \hat{y}_1, \ldots, \hat{y}_\ell \) and repeat Steps 2 and 3 until \( \ell = L \).

This approach is easily extendable to weighted observations, in which case the backaveraging in Step 3 is simply done in a weighted fashion [33]. If multiple \( (\ell_i, y_i) \) pairs are observed with identical covariate-levels, i.e. \( \exists \ell \) s.t. \( N_\ell := |I_\ell| > 1 \) where \( I_\ell := \{i : \ell_i = \ell\} \), we adopt the simple tertiary approach for handling predictories discussed in [34]. In this method, one defines \( \bar{y}_\ell \) as the (weighted) average of the \( \{y_i : i \in I_\ell\} \) and for each level \( \ell \) all \( y_i : i \in I_\ell \) are simply replaced with their mean-value \( \bar{y}_\ell \). Subsequently, PAVA may be applied with non-uniform weights to \( \{(\ell, \bar{y}_\ell)\}_{\ell=1}^{L} \) where the \( \ell \)th point receives weight \( N_\ell \) (or weight \( \sum_{i \in I_\ell} w_i \) if the original points are assigned non-uniform weights \( w_1, \ldots, w_N \)).

**Fact 1** (From de Leeuw [34]). Given weights \( w_1, \ldots, w_N \geq 0 \) and pairs \( (\ell_1, y_1), \ldots, (\ell_N, y_N) \) where each \( \ell \in \{1, \ldots, L\} \) appears at least once, the fitted values \( \hat{y}_1, \ldots, \hat{y}_L \) produced by
tertiary-variant of PAVA are guaranteed to be the best-fitting nondecreasing sequence in the least-squares sense, i.e.

$$\hat{y}_1, \ldots, \hat{y}_L = \arg \min_{z_1 \leq \cdots \leq z_L} \sum_{\ell=1}^{L} \sum_{i \in l_\ell} w_i (z_\ell - y_i)^2$$

By substituting nonincreasing statements in place of the nondecreasing conditions in Steps 2 and 3, the basic PAVA method can be trivially modified to find the best nonincreasing sequence in the least-squares sense. From here on, we use PAVA(((y_1, w_1), \ldots, (y_N, w_N); \delta) to refer to a more general approach than the above algorithm, which incorporates observation-weights w_i, a tertiary weighting scheme to deal with multiple y values for a single \ell, and a user-specified monotonicity condition \delta \in \{"nonincreasing", "nondecreasing"\} that determines which monotonic best-fitting sequence (in the weighted least-squares sense of Fact 1) should be returned.

Fundamentally, our TF algorithm utilizes Dykstra’s method of alternating projections [35] to jump between the space of L-length sequences of vectors which are monotone in each index over \ell and the space of L-length sequences of vectors where each vector represents a valid quantile function. Despite the iterative nature of alternating projections, we find that the TF algorithm generally converges extremely quickly in practice. Each projection only has O(LP) runtime, so our procedure has overall computational complexity O(TLP^2 + NP), which is very efficient when the total number of projections performed T is small (almost always the case in our experiments), as both P and L are limited in practice.
Trend-fitting Algorithm: Numerically solves (2.11) by optimizing (2.12)

**Input 1:** Empirical distributions and associated levels along with optional weights \( \{(\ell_i, \hat{F}_i, w_i)\}_{i=1}^N \)

**Input 2:** A grid of quantiles to work with \( 0 < p_1 < \cdots < p_{P-1} < 1 \)

**Output:** The estimated quantiles of each \( Q_\ell \) \( \{\hat{G}_\ell^{-1}(p_k) : k = 1, \ldots, P - 1\} \) for \( \ell \in \{1, \ldots, L\} \) from which these underlying trending distributions can be reconstructed.

1. \( \hat{F}_i^{-1}(p_k) := \text{quantile}(\hat{F}_i, p_k) \) for each \( i \in \{1, \ldots, N\}, k \in \{1, \ldots, P - 1\} \)
2. \( w^*:=\sum_{i \in I_\ell} w_i \) for each \( \ell \in \{1, \ldots, L\} \)
3. \( x_\ell[k]:=\frac{1}{w^*_\ell} \sum_{i \in I_\ell} w_i \hat{F}_i^{-1}(p_k) \) for each \( \ell \in \{1, \ldots, L\}, k \in \{1, \ldots, P - 1\} \)
4. for \( p^* = 0, p_1, p_2, \ldots, p_{P-1} \):
   5. \( \delta[k] := \text{"nondecreasing" if } p_k > p^*; \text{ otherwise } \delta[k] := \text{"nonincreasing"} \)
   6. \( y_1, \ldots, y_L := \text{AlternatingProjections} \left( x_1, \ldots, x_L ; \delta ; \{w^*_\ell\}_{\ell=1}^L, \{p_k\}_{k=1}^{P-1} \right) \)
   7. \( W[\delta] := \text{the value of (2.12) evaluated with } \hat{G}_\ell^{-1}(p_k) = y_\ell[k] \forall \ell, k \)
   8. Redefine \( \delta[k] := \text{"nonincreasing" if } p_k > p^*; \text{ otherwise } \delta[k] := \text{"nondecreasing"} \)
      
      and repeat Steps 6 and 7 with the new \( \delta \)
9. Identify \( \min_\delta W[\delta] \) and return \( \hat{G}_\ell^{-1}(p_k) = y^*_\ell[k] \forall \ell, k \) where \( y^* \) was produced at the Step 6 or 8 corresponding to \( \delta^* := \arg \max W[\delta] \).
Alternating Projections Algorithm: Finds the closest vector-sequence to the given vectors (in the numerical Wasserstein sense) which both represent valid quantile-functions and a trending sequence with monotonicity specified by $\delta$.

**Input 1:** Initial sequence of vectors $x_1^{(0)}, \ldots, x_L^{(0)}$

**Input 2:** Vector $\delta$ whose indices specify directions constraining the quantile-changes over $\ell$.

**Input 3:** Weights $w^*_k \in \mathbb{R}$ and quantiles to work with $0 < p_1 < \cdots < p_{P-1} < 1$

**Output:** Sequence of vectors $y_1^{(t)}, \ldots, y_L^{(t)}$ where $\forall \ell, k : y_\ell^{(t)}[k] \leq y_\ell^{(t)}[k+1]$ and the sequence $y_1^{(t)}[k], \ldots, y_L^{(t)}[k]$ is monotone nonincreasing/nondecreasing as specified by $\delta[k]$, provided that $x_\ell^{(0)}[k] \leq x_\ell^{(0)}[k+1]$ for each $\ell, k$

1. $r_\ell^{(0)}[k] := 0, s_\ell^{(0)}[k] := 0$ for each $\ell \in \{1, \ldots, L\}, k \in \{1, \ldots, P-1\}$

2. for $t = 0, 1, 2, \ldots$ until convergence:

3. $y_1^{(t)}[k], \ldots, y_L^{(t)}[k] := \text{PAVA} \left( \left( x_1^{(t)}[k] + r_1^{(t)}[k], w^*_1 \right), \ldots, \left( x_L^{(t)}[k] + r_L^{(t)}[k], w^*_L \right) ; \delta[k] \right)$
   for each $k \in \{1, \ldots, P-1\}$. PAVA computes either the least-squares nondecreasing or nonincreasing weighted fit, depending on $\delta[k]$.

4. $r_\ell^{(t+1)}[k] := x_\ell^{(t)}[k] + r_\ell^{(t)}[k] - y_\ell^{(t)}[k]$ for each $\ell, k$

5. $\forall \ell \in \{1, \ldots, L\} : x_\ell^{(t+1)}[1], \ldots, x_\ell^{(t+1)}[P-1] := \text{PAVA} \left( \left( y_\ell^{(t)}[1] + s_\ell^{(t)}[1], \frac{p_2-p_1}{2} \right), \ldots, \left( y_\ell^{(t)}[P-1] + s_\ell^{(t)}[P-1], \frac{p_{P-1}-p_{P-2}}{2} \right) ; \text{"nondecreasing"} \right)$

6. $s_\ell^{(t+1)}[k] := y_\ell^{(t)}[k] + s_\ell^{(t)}[k] - x_\ell^{(t+1)}[k]$ for each $\ell, k$

**Theorem 1.** The TF algorithm produces valid quantile-functions $\hat{G}_1^{-1}, \ldots, \hat{G}_L^{-1}$ which solve the numerical version of the TRENDS objective given in (2.12).

Because the proof of Theorem 1 provides much intuition on the TF algorithm, we do not discuss each step in great detail (all proofs are relegated to the Appendix). The basic idea of the procedure is that once we fix a configuration of $\delta$ (i.e. which quantiles are decreasing over $\ell$ and which ones are increasing), the space over which
we optimize lies at the intersection of two convex sets between which projection is easy thanks to the PAVA algorithm. Furthermore, the second statement in our trend definition prevents the set of possible $\delta$ from growing combinatorially large, so we can simply solve the convex subproblem for each possible $\delta$ to find the global solution.

2.6 Statistical Guarantees

Assuming the model given in (2.5), we now establish some finite-sample bounds regarding the quality of the $\hat{Q}_1, \ldots, \hat{Q}_L$ estimates produced by the TF algorithm. To ensure that our theoretical findings remain pragmatic, we, in addition to avoiding asymptotics, do not express our results in terms of the true Wasserstein distance, which relies on an integral that must be numerically approximated in applications as in (2.12). Rather, because both the distributions $\hat{Q}_\ell$ and Wasserstein integrals must be represented using a finite sequence of quantiles, we provide guarantees in terms of this sequence, the actual quantity one encounters in practice.

Thus in this section, $d_W(\cdot, \cdot)$ is used to refer to the midpoint-approximation of the $L_2$ Wasserstein integral illustrated in (2.12), rather than the true Wasserstein distance given in (2.1). To streamline the exposition and avoid convoluted notation, we also assume the following simplifications hold throughout this section:

- The number of batches at each level is always the same, i.e. $N_\ell := N_1 = \cdots = N_L > 0$

- The same number of samples are drawn per batch, i.e. $n := n_i \forall i \in \{1, \ldots, N\}$

- The per-batch noise functions $\mathcal{E}_i(p)$ all follow the same marginal sub-Gaussian distribution

- The same $P - 1$ quantiles are used to numerically evaluate the Wasserstein integral (2.1) across all $i$, and they are uniformly spaced, i.e. we only consider the $(k/P)$th quantiles of each distribution $P_i$ for $k = 1, \ldots, P - 1$.

- Uniform weights are employed, i.e. in (2.11): $w_i = 1 \ \forall i$. 46
In practice one should choose $P$ based on the available sample sizes, simultaneously taking into consideration that the numerical approximation $d_W(\cdot, \cdot)$ converges to the true integral as $P \to \infty$, but the estimation of extreme quantiles is highly unreliable from sparse samples. Note that all the theorems/proofs presented in this section may be trivially extended to the case where the noise at different $\ell$ follows different sub-Gaussian distributions, the $N_\ell$ are not all equal, a different number of observations $n_i$ is sampled from each $P_i$, and we use a varying number of non-uniformly-spaced quantiles to approximate integral (2.1) for different $i$. For example, to easily obtain crude but valid bounds for the heterogeneous case, one could simply redefine: $N := \min\{N_1, \ldots, N_L\}$, $n := \min\{n_1, \ldots, n_N\}$, and $\sigma^2 := \max_{i,p}\{\sigma_i^2(p) : i = 1, \ldots, N\}$ where $E_i(p)$ follows a sub-Gaussian distribution with parameter $\sigma_i^2(p)$ and $p$ ranges over the quantiles under consideration.

**Theorem 2.** Under the model specified in (2.5), assume:

(A.1) $E_i$ is constrained so that both $Q_{e_i} = f(\ell_i)$ and $F_i^{-1}$ are valid quantile functions $\forall i \in \{1, \ldots, N\}$

(A.2) $\forall p \in [0,1], \forall i : E_i(p)$ have sub-Gaussian marginal distribution with parameter $\sigma^2$. This implies $\mathbb{E}[E_i(p)] = 0$ and $\Pr(|E_i(p)| > t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right) \forall t > 0$; see [36, 37] for an in-depth discussion of sub-Gaussianity.

(A.3) $E_i(p)$ is statistically independent of $E_j(p)$ $\forall p \in [0,1], \forall i \neq j$

If the TF algorithm is applied directly to the true quantiles of $P_1, \ldots, P_N$ (i.e. we directly observe the underlying distributions rather than samples from them), then given any $\epsilon > 0$, the resulting estimates satisfy:

$$d_W(G_{i}^{-1}, G_j^{-1}) < \epsilon \quad \forall i \in \{1, \ldots, L\}$$

with probability $\geq 1 - 2PL \exp\left(-\frac{\epsilon^2 N L}{8\sigma^4 L}\right)$

Thus, Theorem 2 implies that our estimators are consistent as $N_\ell \to \infty$ and characterizes the fast finite-sample convergence-rate if we directly observe the quantiles of $P_1, \ldots, P_N$. Note that by employing the union-bound, we do not require any independence assumptions for the noise introduced at difference quantiles (i.e. between
\( E_i(p_1) \) and \( E_i(p_2) \) for \( p_1 \neq p_2 \). In fact, due to condition (A.1), the noise at different quantiles is likely to exhibit a large degree of dependence. Furthermore, the assumption of sub-Gaussian noise is fairly general, encompassing cases in which the \( E_i(p) \) are either: Gaussian, bounded, of strictly log-concave density, or any finite mixture of sub-Gaussian variables [38]. Because direct quantile-observation is unlikely in practice, we now direct our attention to the performance of TRENDS when these quantiles are instead estimated using \( n \) samples from each \( P_i \).

**Theorem 3.** Under model (2.5) and the assumptions of Theorem 2, we additionally assume:

(A.4) For each \( i \in \{1, \ldots, N\} \) : the quantiles are estimated from \( n \) i.i.d. samples \( X_{1,i}, \ldots, X_{n,i} \sim F_i \)

(A.5) There is nonzero density at each of the quantiles we estimate, i.e. \( F_i(\cdot) \) is strictly increasing around each \( F_i^{-1}(k/P) \) for \( k = 1, \ldots, P - 1 \).

(A.6) The simple quantile estimator defined below is employed for each \( k/P, k = 1, \ldots, P - 1 \)

\[ \hat{F}_i^{-1}(p) := \inf \{ x : \hat{F}_i(x) \geq p \} \]

where \( \hat{F}_i(\cdot) \) is the empirical CDF computed from \( X_{1,i}, \ldots, X_{n,i} \).

If the TF algorithm is applied to the estimated quantiles \( \hat{F}_i^{-1}(k/P), i = 1, \ldots, N, k = 1, \ldots, P - 1 \), then given any \( \epsilon > 0 \), the resulting estimated distributions satisfy:

\[ d_W(\hat{G}^{-1}_\ell, G^{-1}_\ell) < \epsilon \quad \forall \ell \in \{1, \ldots, L\} \]

with probability greater than:

\[ 1 - 2PL \left[ \exp \left( \frac{-\epsilon^2 N \ell}{32 \sigma^2 L} \right) + N \ell \exp \left( -2n \cdot R \left( \frac{\epsilon}{4 \sqrt{\ell}} \right)^2 \right) \right] \]  \hspace{1cm} (2.13)

where for \( \gamma > 0 \):

\[ R(\gamma) := \min_{i,k} \{ R(\gamma, i,k/P) : i = 1, \ldots, N, k = 1, \ldots, P - 1 \} \]  \hspace{1cm} (2.14)

\[ R(\gamma, i, p) := \min \left\{ F_i \left( F_i^{-1}(p) + \gamma \right) - p, p - F_i \left( F_i^{-1}(p) - \gamma \right) \right\} \]
Theorem 4. Under model (2.5), conditions (A.4), (A.5), and the assumptions of Theorem 2, we also assume either:

(A.7) The simple quantile-estimator defined in (A.6) is used, and the support of each \( P_i \) is bounded and connected with non-negligible density, i.e.

\[ \exists \text{ constants } B, c > 0 \text{ s.t. } \forall i: f_i(x) = 0 \text{ if } x \notin [-B, B] \text{ and } f_i(x) \geq c \forall x \in [-B, B] \]

where \( f_i \) is the density function corresponding to CDF \( F_i \).

or

(A.8) Only the following is assumed regarding the quantile-estimation procedure:

1. The quantile-estimators \( \hat{F}_i^{-1}() \) are applied to samples \( X_{1,i}, \ldots, X_{n,i} \sim P_i \) (which are not necessarily independent).
2. The quantiles of each \( P_i \) are estimated independently of the others.
3. The estimates are guaranteed to converge at a sub-Gaussian rate for each quantile of interest, i.e.

\[ \exists c > 0 \text{ s.t. for each } k = 1, \ldots, P - 1, i = 1, \ldots, N, \text{ and any } \epsilon > 0 : \]

\[ \Pr\left( \left| \hat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) \right| > \epsilon \right) \leq 2 \exp(-2nc^2\epsilon^2) \]

Then, the previous bound in (2.13) may be sharpened to ensure that \( \forall \epsilon > 0 : \)

\[ d_W(\hat{G}^{-1}_\ell, G^{-1}_i) < \epsilon \quad \forall \ell \in \{1, \ldots, L\} \]

with probability greater than:

\[ 1 - 2P\left[ L \exp\left(\frac{-c^2N\epsilon}{32\sigma^2L}\right) + \exp\left(-\frac{c^2}{8}N\epsilon^2\right)\right] \quad (2.15) \]

Theorem 5. Under model (2.5) and the assumptions of Theorems 2 and 3, we also assume:

(A.9) Each \( P_i \) has connected support with non-negligible interior density and sub-Gaussian tails at the extremes, i.e. \( \exists \text{ constants } B > b > 0, a > 0, c > 0 \text{ s.t. } \forall i : \)

1. \( F_i \) is strictly increasing,
2. \( f_i(x) \geq c \forall x \in [-B, B] \) where \( f_i \) is the density function of CDF \( F_i \).
3. \( \Pr(X_i > x) \leq \exp\left(-a [x - (B - b)]^2\right) \) if \( x > B \)
4. \( \Pr(X_i < x) \leq \exp\left(-a [x - (-B + b)]^2\right) \) if \( x < B \)
(A.10) Defining \( r := \min \left\{ 2c^2, \frac{2ab^2-1}{4PB^2} \right\} \), we have \( r > 0 \), or equivalently, \( 2ab^2 > 1 \).

(A.11) We avoid estimating extreme quantiles, i.e.

\[
F_i^{-1}(k/P) \in (-B, B) \quad \forall k \in \{1, \ldots, P-1\}
\]

Then, the previous bound in (2.13) may be sharpened to ensure that \( \forall \varepsilon > 0 \):

\[
d_W(G_{\ell}^{-1}, G_{\ell}^{-1}) < \varepsilon \quad \forall \ell \in \{1, \ldots, L\}
\]

with probability greater than:

\[
1 - 2P \left[ L \exp \left( \frac{-\varepsilon^2 N_\ell}{32\sigma^2 L} \right) + \exp \left( -\frac{\tau}{16} N_\ell n \varepsilon^2 \right) \right] \tag{2.16}
\]

Theorem 3 is the most general result applying to arbitrary distributions \( P_i \) which satisfy basic condition (A.5), but the resulting probability-bound may not converge toward to 1 if \( n \cdot R(\frac{\varepsilon}{4\sqrt{L}})^2 < O(\log N_\ell) \). Tacking on a fairly general condition that the distributions have connected support and small tails, Theorem 5 provides a much sharper finite-sample bound which decreases exponentially in both \( n \) and \( N_\ell \).

While these theorems assume the natural quantile-estimator defined in (A.6) is used, numerous superior estimators have been developed which are likely to improve convergence rates in practice [39]. Under guarantees regarding the performance of the quantile-estimation (which may be based on some combination of underlying properties of the \( P_i \) as well as the estimator employed), Theorem 4 is also able to provide a finite-sample bound for the performance of our estimators which decreases exponentially in \( n \) and \( N_\ell \). Because of the complex interplay between aspects of the probability distribution and the quality of a particular estimator, it is unclear which (estimator, distribution-family) pairs achieve the bound presumed in (A.8), although we have demonstrated that this condition is met when applying the simple quantile-estimator defined in (A.6) to distributions of bounded and connected support. Finally, note that while we specified both the distribution of noise and performance of the quantile-estimators in order to develop the theoretical guarantees presented in this section, our nonparametric significance testing methods presented in the following
section do not rely on these assumptions and remain valid regardless of the actual noise and quantile-estimation procedure.

## 2.7 Testing for significance

In this section, we introduce a test of statistical significance for the trend-effect which assesses the null hypothesis

$$H_0 : Q_1 = Q_2 = \cdots = Q_L$$

(2.17)

against the alternative that the $Q_i$ are not all equal and follow a trend. To obtain a $p$-value, we employ permutation testing on the $\ell_i$-labels of our observed distributions $\widehat{P_i}$ with test-statistic $R^2$ [40]. More specifically, one determines the null distribution of the test-statistic under $H_0$ by repeatedly executing the following steps: (i) randomly shuffle the $\ell_i$ so that each $\widehat{P_i}$ is paired with a random $\ell_i^{\text{perm}} \in \{1, \ldots, L\}$ value, (ii) fit the TRENDS model to the pairs $\{(\ell_i^{\text{perm}}, \widehat{P_i})\}_{i=1}^N$ to produce $\widehat{Q}_1^{\text{perm}}, \ldots, \widehat{Q}_L^{\text{perm}}$, (iii) use these estimated distributions to compute $R^2_{\text{perm}}$ using (2.8). Due to the quantile-noise functions $E(-)$ assumed in our model (2.5), $H_0$ allows variation in our sampling distributions $P_i$ which stems from non-$\ell$-trending forces. Thus the TRENDS test attempts to distinguish whether the effects transforming the $P_i$ follow a trend or not, but does not presume the $P_i$ will look identical under the null hypothesis.

By measuring how much further the $\widehat{P_i}$ lie from a single common distribution v.s. a sequence of trending distributions in Wasserstein-space, we note that the $R^2$ value resembles a likelihood-ratio-like test statistic between maximum-likelihood-like estimates $\overline{F}^{-1}$ and $\widehat{Q}_t$, where we operate under the Wasserstein distance rather than the Kullback-Leibler divergence from which the standard maximum likelihood framework is derived. As we do not parametrically treat the observed distributions, we find the generality of permutation testing for the significance of the trend more suitable than attempting to appeal to large-sample theory. This consideration is particularly relevant in many applications in which we only observe a single sequence of batches (e.g.
each $\ell \in \{1, \ldots, L\}$ is only associated with one distribution) and hence the effective sample size is minimal. An in-depth discussion of permutation testing and its broad applicability may be found in [41,42].

Unfortunately, in many settings of interest such as most currently existing SCRS time course data, $N$ and $L$ are both small. This limits the number of possible-permutations of distribution-labels and hence the granularity and accuracy with which we can determine $p$-values in the our test. Note that TRENDS estimation is completely symmetric with respect to a reversal of the distributions’ associated levels (i.e. replacing each $\ell_i \leftarrow L - \ell_i + 1$), so if $B$ denotes the number of possible permutations, we can only obtain $p$-values of minimum granularity $2/B$ which may be unsatisfactory in the small $N, L$ regime. In the classical tissue-level differential gene expression analyses (in which sample sizes are typically small), this problem has been dealt with by permuting the genes (of which there are many) rather than the sample labels. However, this approach is not completely valid as it discards the (often substantial) correlations between genes, and has been found to produce suboptimal results [40].

To circumvent these issues, we propose a variant of our label-permutation-based procedure to obtain finer-grained but only approximate $p$-values (where in the small $N, L$ setting, rough approximations are all one can hope for since asymptotics-derived $p$-values are also almost certainly incorrect). The underlying goal of our heuristic is to produce a richer picture of the null distribution of $R^2$ (at the cost of resorting to approximation), which is accomplished as follows:

1. Shuffle the distributions' $\ell_i$-labels as described above, but now explicitly perform all possible permutations, except for the permutations that produce a sequence $\{\ell_i^{\text{perm}}, \ldots, \ell_N^{\text{perm}}\}$ which equals either the sequence of actual labels $\{\ell_1, \ldots, \ell_L\}$ or its reverse in which each $\ell_i$ is replaced by $L - \ell_i + 1$.

2. For data in which each distribution $\hat{P}_i$ is estimated from a set of samples $\{X_{i,s}\}_{s=1}^{n_i}$, one can obtain a diverse set of $K$ null-distributed datasets from a single permutation of the labels by employing the bootstrap. For each $k = 1, \ldots, K$ and $i = 1, \ldots, N$: draw $n_i$ random samples $Z_{i,s}^{(k)}$ with replacement from
\( \{X_{i,s}\}_{s=1}^{n_i} \), compute a bootstrapped empirical distribution \( \hat{P}_i^{(k)} \) using \( \{Z_{i,s}^{(k)}\}_{s=1}^{n_i} \), and assemble the \( k \)th null-distributed dataset (under the current labels-permutation) by pairing the bootstrapped empirical distributions with the permuted labels \( \ell_i^{\text{perm}} \).

3. Apply TRENDS to each null-distributed dataset \( \{(\ell_i^{\text{perm}}, \hat{P}_i^{(k)})\}_{i=1}^{N} \) and compute a \( R^2_{\text{perm},k} \) value via (2.8) which is distributed according to the desired null (where \( K = 1 \) and \( \hat{P}_i^{(k)} = \hat{P}_i \) if bootstrapping is not performed).

4. Form a smooth approximation of the null distribution by fitting a kernel CDF estimate \( \hat{F} \) to the collection of \((B-2) \cdot K\) null samples \( \{R^2_{\text{perm},k}\} \) where \( k = 1, \ldots, K \) and perm is an index over the possible label-permutations under consideration (we use the Gaussian kernel with the plug-in bandwidth proposed by Altman and Léger, which has worked well even when only 10 samples are available [43]). Finally, the approximate \( p \)-value is computed as \( \hat{p} := 1 - \hat{F}(R^2) \), where \( R^2 \) corresponds to the fit of TRENDS on the original dataset.

Note that under the exchangeability of labels assumed in \( H_0 \), the sequence of \( \ell_i \) corresponding to the actual ordering or its reverse are equally likely a priori as any other permutation of the \( \ell_i \). Thus, Step 1 above is unbiased, despite the omission of two permutations from the set of possibilities. Producing a much richer null distribution than the empirical version based on few permutation samples, the bootstrap and kernel estimations steps enable us to obtain continuum of (approximate) \( p \)-values. Intuitively, our richer approximation is especially preferable for differentiating between significant \( p \)-values despite its sensitivity to the bandwidth setting, because the standard permutation test offers no information when the actual test statistic is greater than every permuted statistic (a common occurrence if \( B \) is small), whereas our approach assigns smaller \( p \)-values based on the distance of the actual test statistic from the set of permuted values. Finally, we remark that the kernel estimation step in our \( p \)-value approximation is similar to the approach of Tsai and Chen [44], and point out that as the number of distributions per level \( N_\ell \) grows, the approximation
factor of our procedure shrinks, as is the case for p-values based on asymptotics which are themselves only approximations.
Chapter 3

Results

3.1 Simulation study

To investigate the behavior of TRENDS on simulated data, we generate samples from the TRENDS model (2.5) under one of the following choices of the underlying trending distribution sequence $Q_1, \ldots, Q_L$ with $L = 5$:

(S1) $Q_1, \ldots, Q_5$ are Gaussian distributions with standard deviation 1 and means $0, 0.2, 0.5, 0.9, 1$

(S2) $Q_1, \ldots, Q_5$ are mean-zero Gaussian distributions with standard deviations $0.4, 0.7, 0.9, 1, 1$

(S3) Each $Q_\ell$ is a mixture distribution of $N(0, 1)$ and $N(1, 1)$ components, with the mixing proportion of the latter component ranging over $\lambda_\ell \in \{0.2, 0.4, 0.5, 0.7, 0.7\}$ for $\ell = 1, \ldots, 5$

(S4) $Q_1, \ldots, Q_5$ are identical $N(0, 1)$ distributions

Noise is introduced in the simulations by performing the following steps for the $i$th batch: we (independently of the other batches) draw a single $\mathcal{E}_i \sim N(0, \sigma^2)$ perturbation which affects all samples from the batch and $n$ i.i.d. latent values $z_{i,s} \sim Q_{\ell_i}$ ($n$ and $\sigma$ are fixed across all $i \in \{1, \ldots, N\}$). The values observed in the $i$th batch are subsequently defined as $x_{i,s} = z_{i,s} + \mathcal{E}_i$ with probability $1/2$ and $x_{i,s} = z_{i,s}$ otherwise. Thus, $\sigma$ measures the overall magnitude of the noise, which is neither Gaussian.
(it is however sub-Gaussian) nor i.i.d. between different $x_{i,s}$. We simulate numerous datasets under different settings of $n$, $\sigma$, and $N_{\ell}$, the number of batches observed per level. Each of our experiments in this section is repeated 100 times to provide better insight on the average performance and variability of the results.

We first investigate the convergence of the TRENDS estimators under each of the models $S_1$ - $S_4$, varying $n$, $\sigma$, and $N_{\ell}$ independently to identify how each of these factors affects the resulting estimates. Figure 3-1 demonstrates the rapid convergence of the TRENDS estimator under each of the underlying models we investigate and also shows that TRENDS produces a much better picture of the underlying distributions than the (noisy) observed empirical distributions when there is truly a trend. As shown in Figure 3-1(A), this phenomenon even holds in the case where there is no external noise, thanks to the additional structure of the trend-assumption exploited by our estimator. Thus, if the underlying effects follow a trend, our $\Delta$ statistic provides a more accurate measure of the effect magnitudes than Wasserstein distances between the empirical distributions, even in the noiseless regime under moderate batch sizes.

Having demonstrated the fast statistical convergence of the TRENDS estimator in practice, we turn our attention to the approximation factor of our significance testing heuristic which employs bootstrapping within each permutation as well as kernel estimation to approximate the null CDF of the test statistic. We again draw samples from each of the underlying models $S_1$ - $S_4$ with $n = 100$, $N_{\ell} = 1$, and $\sigma = 0.2$ (recall that we only advocate the use of this approximation when the number of batches is so small that the set of possible permutations is undesirably limited). To each simulated dataset, we apply the TRENDS model and then determine the significance of the TRENDS $R^2$ (which measures the model’s overall fit) via a standard permutation test utilizing all possible permutations of the batch labels (here $L = 5$ so the number of distinct possible permuted-$R^2$ values from the null is $5!/2 = 60$). We subsequently employ our $p$-value approximation to assess the significance of the same $R^2$ value using the same permutations as before, but with additional bootstrapped samples drawn under each permutation of the batch labels until the total number of null samples is enlarged to at least 1000. Subsequently, the kernel CDF procedure is applied to
Figure 3-1: The Wasserstein error (the sum over \( \ell \) of the squared Wasserstein distances between the underlying \( Q_\ell \) and the estimates thereof) of the TRENDS fitted distributions as well as the observed empirical distributions, under models \( S_1 \)-\( S_4 \) and a variety of settings of \( n, \sigma, \) and \( N_\ell \). In each plot, one of these factors is varied along the x-axis with the other factors fixed to the setting indicated at the top, and lines depict the average error over 100 experiments.

These 1000 null samples as described in the previous section to obtain an approximate p-value.

To compare our approximation with the standard permutation test p-value, we require the actual p-value of the observed \( R^2 \) describing the TRENDS fit, which is obtained as follows: a minimum of \( J = 1000 \) new datasets (i.e. batch sequences) from
the same underlying model are drawn in which $\ell$ is randomly permuted among the
different batches within a single dataset. TRENDS $R^2$ values are then computed for
each of these null datasets (which resemble the permuted data we use in practice,
but each permutation of the labels is matched with freshly sampled batches corre-
spending to a new dataset), and we can subsequently define the underlying $p$-value
as in permutation testing. Note that this approach can approximate the actual null
distribution of $R^2$ arbitrarily well as we increase $J$, and in our experiments, we be-
gin with $J = 1000$ and gradually increase up to 1,000,000 while at least 5 null-$R^2$
values greater than the one observed in the original data have not yet been observed.
Table 3.1 demonstrates that while our approximation and the permutation test have
comparable degrees of bias, the variability of the latter method can be vastly larger
than the variance of our approximate $p$-values. Because it is critical that significance
testing results remain stable to small variations in the data, we view the decrease in
variability achieved by our approximation as a highly desirable characteristic.

<table>
<thead>
<tr>
<th>Model</th>
<th>Average $p$</th>
<th>$E[\hat{p} - p]$</th>
<th>SD($\hat{p}$)</th>
<th>$E[p_{\text{perm}} - p]$</th>
<th>SD($p_{\text{perm}}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0.026</td>
<td>-0.0012</td>
<td>0.011</td>
<td>0.0032</td>
<td>0.019</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.045</td>
<td>-0.017</td>
<td>0.023</td>
<td>-0.0076</td>
<td>0.038</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0.134</td>
<td>-0.031</td>
<td>0.031</td>
<td>-0.022</td>
<td>0.047</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0.481</td>
<td>0.0017</td>
<td>0.023</td>
<td>0.026</td>
<td>0.077</td>
</tr>
</tbody>
</table>

Table 3.1: Comparing our approximate $p$-values ($\hat{p}$) against those produced by the permutation test ($p_{\text{perm}}$). The second column contains the average of the actual true $p$-values (over 100 runs of the experiment for each underlying model $S_1$-$S_4$), the third and fifth columns contain the bias of the methods, and the fourth and final columns contain the standard deviation of the significance values produced by each approach.

Finally, we investigate how severe the drop in statistical power may be if one
applies a method sensitive to arbitrary types of temporal effects in the TRENDS
settings. Fixing $N_\ell = 1$ for $\ell = 1, \ldots, 5$ and $n_i = 100$ for all $i$, we generate 600 sets
of samples (empirical distribution - $\ell_i$ pairs) from the different underlying TRENDS
models described above (100 from each of $S_1$, $S_2$, $S_3$ and 300 from $S_4$ to ensure a
balance between the examples with and without non-constant trend effect, where
additional diversity under the null is introduced in the $S_4$ samples by choosing the standard deviation randomly between 0 and 2 rather than fixing it at 1. For each sampled sequence, the standard deviation of the quantile noise $\sigma$ is identical across all distributions $P_i$ and is set to one of from one of ten values between 0 and 1. We apply TRENDS to each sample to obtain a $p$-value for the significance of the observed $R^2$ (using all possible possible permutations, enlarging the set of different null statistics to 1000 via the bootstrap, and finally applying the kernel CDF heuristic). Additionally, we apply the KS and MI methods (see next section) which make no underlying assumption and can capture arbitrary deviations between distributions with different $\ell$. Figure 3-2 depicts a receiver operating characteristic (ROC) plot evaluating how well each method identifies the trending samples, and the superiority of TRENDS is evident. This results also reinforces the utility of the $p$-values produced through our approximation.

Figure 3-2: Evaluating TRENDS against the KS and MI methods on a large simulated dataset with 600 samples drawn from four different TRENDS models with varying levels of noise. (A) ROC curves depicting how well the TRENDS $p$-value, KS $p$-value, and mutual information are able to distinguish between samples from $S_4$ vs. the other three underlying models with non-constant trend. (B) TRENDS $p$-values and $R^2$ for each of the 600 simulated sequences, colored by the underlying model from which they are drawn.
3.2 Application to scientific data

While TRENDS exhibits desirable properties both in theory and in our simulation experiments, it remains unclear that the TRENDS assumption is valid for real data. To demonstrate how useful insights can be derived from the TRENDS framework in practice, we apply the method to two SCRS time course datasets and compare its results against those of other approaches. In each analysis, the expression values of each gene are represented in (log-transformed) Fragments Per Kilobase of transcript per Million mapped reads (FPKM) [3]. Other expression-measures such as the raw read count or the posterior for a zero-inflated negative-binomial rate have also been employed [21, 22], but these typically require delicate statistical procedures tailored to the values' range. While TRENDS is a fully nonparametric general-purpose model and can thus be applied to any representation of gene-expression (and normalized variant thereof), we find FPKM values favorable due to their interpretability and direct comparability between different genes. Similarly, the methods we compare TRENDS against are all well suited for (log-transformed) FPKM values (i.e. they do not hinge on count-specific distributional assumptions, as is required by many of the popular differential expression methods for tissue-level RNA-seq data [21]).

Implicitly assuming that temporal-progression effects on expression are likely larger in genes more important in development (since the FPKM-scale is directly comparable between genes), we measure the size of these effects using the $\Delta$ statistic of (2.10). By fitting a TRENDS model to each gene's cellular expression distribution over the time course, we can obtain a ranking of the genes' presumed developmental importance as measured by $\Delta$. If rather than measuring the relative importance of different genes in developmental processes, the goal in some settings may be to simply to pinpoint a set of high-confidence candidate genes relevant at all in development (ignoring the degree to which their expression transforms in the developmental progression). In this case, one may apply our permutation test to establish which genes exhibit strong statistical evidence of an underlying nonconstant TREND effect. We find that both the $\Delta$ effect and significance testing approaches produce highly
informative results in our applications.

We now briefly describe the other methods that TRENDS is compared against. Note that the methods which model the full conditional distributions may be ordered based on increasing generality of the underlying assumption (and types of effects that can be captured) as follows: Linear TRENDS $\rightarrow$ TRENDS $\rightarrow$ KS / MI. By selecting a model later in this ordering, one can capture a wider diversity of effects at the cost of a decrease in statistical power.

3.2.1 Kolmogorov-Smirnov method (KS)

This approach performs an omnibus test of the hypothesis that there exist $\ell_1$ and $\ell_2$ such that $\Pr(X | \ell_1) \neq \Pr(X | \ell_2)$. As a test statistic and measure of effect-size, we use the maximum Kolmogorov-Smirnov test statistic between these empirical conditional distributions over all possible pairs $\ell_1 < \ell_2 \in \{1, \ldots, L\}$. To assess statistical significance, permutation testing is employed (permuting the $\ell$ labels of each individual cell), since the asymptotic distribution is no longer valid after maximization.

3.2.2 Mutual information method (MI)

In this approach, one computes the mutual information between $\ell$ and $X$ to estimate the size of the effect. Because we operate in the fixed-design setting, $\ell$ is technically not a random variable, so we instead employ a conditional variant of the mutual information in which the marginal distribution of $\ell$ is disregarded, similar to the DREMI method presented in [20]. This is achieved by simply reweighing our samples to ensure the marginal distribution of $\ell$ is uniform over $\{1, \ldots, L\}$, and subsequently obtaining kernel density estimates of the reweighed joint $(X, \ell)$ distribution as well as each conditional $\Pr(X | \ell_1)$ from which the mutual information can be calculated.

3.2.3 Linear TRENDS (LT) model

The Linear TRENDS (LT) method is very similar to our approach, except it relies on the stringent assumption that each quantile evolves according to its own ordinary
linear regression model over the values of the covariate. This approach thus operates on real-valued rather than ordinal covariates (e.g. the actual values of the time points $t_\ell$ when available in our SCRS context). Therefore, the primary difference between Linear TRENDS and our proposed method is the fact that the former accounts for the scaling of the covariate with the added assumption that effects manifest linearly across this scale. This regression framework has been previously proposed in numerous contexts, although it is primarily only used to simultaneously estimate a few specific quantiles of the conditional distribution [14, 15]. Linear quantile regression (with non crossing) can nonetheless be employed to model the full distribution by simply selecting a grid of quantiles spanning $(0, 1)$ as is done in TRENDS. In [14, 15], the model is fit by optimizing the standard quantile regression loss function jointly over the quantiles of interest in a quadratic programming formulation with constraints that ensure the quantiles do not cross (only enforced over the covariate’s range so that the lines reflecting the expected quantiles of the response are not restricted to be parallel).

In our setting, the empirical quantiles of each conditional distribution are available, so one can directly employ the usual squared error loss on the fitted quantiles themselves rather than relying on the quantile regression loss function used in [14, 15]. Analogous to the proof of Theorem 1, one can easily show that optimizing the squared error loss (on each quantile) implies the distributions constructed from the set of fitted quantiles are the Wasserstein least-squares fit under the restriction that each quantile evolves linearly over $t_\ell$, the time at which the distribution is sampled. By replacing the PAVA step (over $\ell$) of the TF algorithm with standard linear-regression (with $t_\ell$ as the sole covariate, and omitting the search for the split between increasing and decreasing quantiles), our alternating projections method can be trivially adapted to fit the set of non-crossing quantile linear regressions under the squared-loss, which we refer to as the LT estimator. In the case where we estimate around 100 quantiles representing the entire distributions, we find that our linearized TF algorithm is orders of magnitude faster than the quadratic programs introduced in [14, 15], which have difficulty dealing with the large number of constraints required in this setting. We
thus employ the linearized TF algorithm to fit the linear quantile regressions in our applications (efficiency is particularly important when fitting the model thousands of times in our gene-expression analyses), noting that besides the vast decrease in runtime, this variant produces nearly identical results as the methods of [14,15] despite the change in loss function. Just as in TRENDS, one can measure the overall size of the effect of $\ell$ on the observed distributions under the Linear TRENDS model by computing a $\Delta$ value via (2.10), and we can also test the significance of the effect (simultaneously across all quantiles) using the same permutation methodology.

### 3.2.4 Tobit model (censored regression)

Trapnell et al. introduce an average-case method as part of their Monocle framework which is specifically tailored for the analysis of single-cell gene expression over time [3], but does not attempt to model the entire distribution of expression across the cell population. This approach evaluates the genes based on the significance of the regression coefficients in a Tobit-family generalized additive model fit to the FPKM values. It is thus assumed the expression values follow a log-normal distribution, and the Tobit link function is introduced to deal with the scarcity of observed reads from some genes expected to be significantly expressed (which plagues SCRS data due to the small amount of RNA that can be isolated from individual cells). In this approach, we try both directly regressing $X$ against $t_\ell$ (and refer to this generalized linear model as the linear Tobit), as well as initially performing a B-spline basis expansion of the $t_\ell$ values before fitting the Tobit regression in order to capture a diverse set of nonlinear effects, as done in [3].

### 3.2.5 Human myogenesis data

Trapnell et al. recently studied the single-cell transcriptome dynamics of skeletal myoblast cells during differentiation to identify the genes which orchestrate the morphological/functional changes observed in this process (myoblasts are embryonic progenitor cells which subsequently become muscle cells in myogenesis) [3]. Primary
human myoblast cells were cultured in high-serum medium and differentiation was induced by switching to low-serum medium. Subsequently, the cells were dissociated and individually captured at four 24-h intervals (where time is measured starting from the serum-switch), and a RNA-Seq library was prepared and sequenced for each cell. Following Trapnell et al., we measure gene-expression in each cell by applying the \( \log_{10}(x + 1) \) transform to the normalized FPKM values submitted to the Gene Expression Omnibus by the authors (the data are available under accession GSE52529).

While the Fluidigm C1 microfluidic system used in this experiment is capable of profiling 96 cells in each capture plate (one plate is used per time point), some of the wells in the capture plate contain visible debris and thus cannot be confirmed to come from a (whole) single cell. In addition to discarding the data from these wells, Trapnell et al. also stringently omit all cells whose libraries were not sequenced deeply (\( \geq 1 \) million reads) because they all apply high-dimensional manifold-like methods in their analysis which are sensitive to noise. Since the TRENDS regression model is designed to distinguish biological effects from noise, we retain these cells embracing the additional (albeit noisy) insight on the underlying \( g \)-expression distributions they provide. Omitting only the wells annotated for debris, the data we analyze contains the following number of cells at each time point: 0h: 93 cells, 24h: 93 cells, 48h: 93 cells, 72h: 76 cells.

3.2.6 Mouse cortex data

In a large-scale SCRS experiment, Zeisel et al. isolated 1,691 cells from the somatosensory cortex (i.e. the sensory system located in the forebrain) of juvenile CD1 mice aged P22-P32 [6]. We thus treat the age (in postnatal days) of the mice as our covariate, whose ordinal representation takes one of \( L = 10 \) possible values. Numerous Fluidigm C1 runs of captured cells were employed for some of the age-values, implying the number of batches per covariate-level \( N_\ell > 1 \) for many \( \ell \). From these cells, nonzero expression measurements are obtained for 14,575 genes, and we again measure expression by applying the \( \log_{10}(x + 1) \) transform after first computing FPKM values from the sequencing read counts available in the Gene Expression Omnibus.
3.2.7 Results for SCRS time course data

As the myoblast data only contains four \( \ell \)-levels and one batch from each, the TRENDS permutation test stringently identifies merely 76 out of the 17,341 genes analyzed with significant non-constant trend at the 0.05 level. A search for terms which are statistically overrepresented in the Gene Ontology (GO) annotations of these significant genes (using ConsensusPathDB [45]) indicates known developmental relevance of a large subset (see Figure 3-3A), including enriched terms such as “stem cell development”, “tissue development”, and “anatomical structure development”. In contrast, the cortex data is much richer, and TRENDS accordingly finds far stronger statistical evidence of trending genes, identifying 1,351 as significant at the 0.05 level. Again, we search for enriched terms in the GO annotation of these significant genes, and find a large subset to be developmentally relevant (see Figure 3-3B), with significantly overrepresented terms such as “neurogenesis”, “nervous system development”, “neuron fate commitment”, and “positive regulation of developmental process”.

Furthermore, out of the top ten most developmentally important genes inferred by TRENDS \( \Delta \) value in the myogenesis experiment, the following nine have each been previously shown to play a significant role in myogenesis and some are even already employed as standard markers to distinguish different stages of the differentiation process (see the work cited for each gene, which is listed along with its ranking by \( \Delta \) value in parentheses): (1) Metallothionein 2A [46], (2) Alpha-actin-2 [47], (3) metallothionein 1L [46], (4) troponin I type 1 [48], (5) myosin light chain, phosphorylatable, fast skeletal muscle [48], (6) myosin, heavy chain 3, skeletal muscle, embryonic [3], (7) metallothionein 1E [46], (9) fatty acid binding protein 3, muscle and heart [49], (10) dickkopf WNT signaling pathway inhibitor 1 [50]. From the cortex analysis, the following genes with top ten \( \Delta \) values have been previously implicated in brain development (again with \( \Delta \)-based ranks shown in parentheses): (1) Somatostatin, (3) Prostaglandin D2 synthase (brain), (4) Myelin proteolipid protein, (5) Myelin Oligodendrocyte Glycoprotein, (6) Neuropeptide Y, (9) Apolipoprotein-D, (10) Ermin,
ERM-Like Protein. Thus, TRENDS accurately assigns the largest inferred effects to highly developmentally-relevant genes, a property of great importance when studying a less characterized developmental process than the brain and myoblast examples presented here, particularly because considerable effort is required to experimentally probe putative developmental candidates.

Figure 3-5A demonstrates that TRENDS infers that the gene identified as most developmentally relevant in myogenesis, Metallothionein 2A, is universally down-regulated with the developmental progression across all cells in the population. Interestingly, the vast majority of cells express Metallothionein 2A at a uniformly high level of $\geq 3$ log FPKM at the time of serum-switch, but almost no cell exhibits such high expression of the gene 24 hours later and its expression has become much more heterogenous throughout the population, with some cells retaining significantly nonzero Metallothionein 2A expression for the remainder of the 72 hour time course. As a highly expressive model, TRENDS is able to account for all of these different sorts of changes in its quantification of the underlying temporal effects.

Figure 3-3: ConsensusPathDB word clouds [45] created from the terms significantly enriched (at the default 0.01 level) in the GO annotations of the genes with significantly trending expression in the myoblast data (A) and in the somatosensory cortex cells (B).

Because myogenesis is the focus of the first experiment, any gene which has been previously implicated in muscle development is of interest, and we can thus form a lower-bound approximation of the fraction of “true positives” discovered by each method by counting the genes whose annotations contain a term in which both the words “muscle” and “development” are found (e.g. “muscle organ development”, “skeletal muscle tissue development”, etc. Appendix Table S1 contains the full list of GO annotations which meet this criterion). Figure 3-4A depicts a pseudo-sensitivity plot.
based on this approximation over the genes with the highest presumed developmental importance inferred under each method. It is clear that a much larger fraction of the top genes discovered by TRENDS and Linear TRENDS have been previously linked with muscle development than the top genes under the other methods. We repeat the same analysis for the cortex data, only with a different set of “ground truth” annotation terms (listed in Appendix Table S2), and again find that TRENDS outperforms all other approaches (Figure 3-4B). This illustrates the necessity in SCRS time course analysis to account for both the full expression distribution and the temporal ordering of the experiment, much more important considerations in this setting than the particular distinction between monotone-quantile vs. linear trends, which only leads to an improvement of approximately two additional true positives uncovered in the top 250 genes. Because it is unrealistic for researchers to study more than a few hundred individual genes in greater detail, it is important that a method for developmental gene discovery produces many higher ranking true positives which might subsequently be verified through further experimentation.

This data also provides another example demonstrating the importance of treating the full expression distributions rather than just modeling mean-effects. It has been shown that the Nestin gene plays an essential role in myogenesis, determining the onset and pace of myoblast differentiation [51]. Nestin overexpression can also bring differentiation to a halt, a process possibly underway in the high-expression cells from the later time points depicted in Figure 3-5B. Thus, this is certainly a gene of great interest in a developmental study of myoblasts. Under TRENDS, Nestin is the 35th most significantly trending gene in the myoblast data, but it fails to be identified by the average-based methods, only ranking 3,291 and 5,094 in the results of the linear / B-spline Tobit models (with q-value = 1 under both of these models). Although Figure 3-5B depicts a clear temporal effect on mean Nestin expression, the average-based methods fail to identify this gene because they do not properly account for the other types of changes affecting different segments of the cell population in the multitude of other genes with similar mean-effects. Furthermore, although the closely-related Linear TRENDS model appears to do nearly as well as TRENDS in our pseudo-
sensitivity analysis (Figure 3-4), we find the linearity assumption overly restrictive, causing for example the Linear TRENDS model to fail to identify TSPYL5, a nuclear transcription factor which suppresses levels of well-known myogenesis regulator p53 [52, 53]. Based on the existing findings, TSPYL5 is of great potential interest in this experiment, but the Linear TRENDS model only assigns this gene a p-value of 0.11 because time is not linearly related to the quantiles of TSPYL5 expression, while TRENDS identifies it as significant ($p = 0.04$) since TSPYL5 expression follows a trend fairly closely ($R^2 = 0.95$).

### 3.2.8 ACS income distributions

To demonstrate the utility of TRENDS beyond SCRS analysis, we present a brief study the long-term impacts of the 2007 recession on incomes in various industries. American Consensus Survey (ACS) reported income data from 12,020,419 individuals across the USA in the years 2007-2013 were obtained from the Integrated Public Use
Figure 3-5: Examples of time course single-cell RNA-seq data for various known development-regulating genes. The empirical cellular gene expression distributions are depicted on the left, along with the corresponding TRENDS fitted distributions on the right. Each dot denotes an individual cell's expression value, and the shaded regions represent kernel density estimates of the expression distribution across the cell population.

Microdata Series [54]. After filtering out individuals with missing or $\leq$1 and under reported income, the data consists of 257 industries from which at least 100 people were surveyed in each of the years under consideration. We fit TRENDS to the data
from each industry individually, treating the observations from each year as a single batch and the year-index in our time series as the label \((\ell = 1, \ldots, 7)\). Table 3.2 lists the industries which according to TRENDS are subject to the largest temporal effects in income distribution over the post-recession period. The table contains numerous industries from the business/financial and manufacturing sectors, which were known to be particularly affected by the recession. Interestingly, many industries from the energy sector are also included in the table, a reflection of the enactment of the Energy Independence and Security Act of 2007, which sought to move the U.S. toward greater energy efficiency and reduce reliance on imported oil. The other industries in which income distributions were subject to the largest temporal progression effects are predominantly technology-related, representing an overall trend which has been active over the last decade. Of particular note is the “other information services” industry (includes web search, online media, and news syndicates), in which incomes have generally risen over the years and we observe the emergence of a distinct subgroup of individuals with reported incomes in the hundreds of thousands.

![Graph](image)

Figure 3-6: The distributions of reported income in each annual ACS survey of individuals in the “other information services” industry. (A) depicts kernel density estimates applied to the survey results from each year while (B) depicts the TRENDS fitted distributions.
<table>
<thead>
<tr>
<th>Industry</th>
<th>(R^2)</th>
<th>p-value</th>
<th>(\Delta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other information services</td>
<td>0.97</td>
<td>0.02</td>
<td>5465</td>
</tr>
<tr>
<td>Software publishers</td>
<td>0.78</td>
<td>0.10</td>
<td>2991</td>
</tr>
<tr>
<td>Electronic auctions</td>
<td>0.86</td>
<td>0.04</td>
<td>2584</td>
</tr>
<tr>
<td>Oil and gas extraction</td>
<td>0.78</td>
<td>0.12</td>
<td>2454</td>
</tr>
<tr>
<td>Miscellaneous petroleum and coal products</td>
<td>0.52</td>
<td>0.38</td>
<td>2415</td>
</tr>
<tr>
<td>Other telecommunication services</td>
<td>0.80</td>
<td>0.07</td>
<td>2414</td>
</tr>
<tr>
<td>Pharmaceutical and medicine manufacturing</td>
<td>0.98</td>
<td>0.04</td>
<td>2220</td>
</tr>
<tr>
<td>Management of companies and enterprises</td>
<td>0.66</td>
<td>0.12</td>
<td>2194</td>
</tr>
<tr>
<td>Metal ore mining</td>
<td>0.89</td>
<td>0.02</td>
<td>2074</td>
</tr>
<tr>
<td>Support activities for mining</td>
<td>0.88</td>
<td>0.03</td>
<td>1915</td>
</tr>
<tr>
<td>Electric and gas, and other combinations</td>
<td>0.82</td>
<td>0.03</td>
<td>1910</td>
</tr>
<tr>
<td>Non-depository credit and related activities</td>
<td>0.92</td>
<td>0.06</td>
<td>1860</td>
</tr>
<tr>
<td>Sound recording industries</td>
<td>0.51</td>
<td>0.38</td>
<td>1731</td>
</tr>
<tr>
<td>Electronic component and product manufacturing</td>
<td>0.99</td>
<td>0.02</td>
<td>1719</td>
</tr>
<tr>
<td>Securities, commodities, funds, trusts, and other financial investments</td>
<td>0.57</td>
<td>0.23</td>
<td>1665</td>
</tr>
<tr>
<td>Agricultural chemical manufacturing</td>
<td>0.77</td>
<td>0.09</td>
<td>1635</td>
</tr>
<tr>
<td>Communications, and audio and video equipment manufacturing</td>
<td>0.72</td>
<td>0.09</td>
<td>1628</td>
</tr>
<tr>
<td>Pipeline transportation</td>
<td>0.70</td>
<td>0.14</td>
<td>1620</td>
</tr>
<tr>
<td>Coal mining</td>
<td>0.90</td>
<td>0.04</td>
<td>1573</td>
</tr>
<tr>
<td>Natural gas distribution</td>
<td>0.69</td>
<td>0.11</td>
<td>1546</td>
</tr>
</tbody>
</table>

Table 3.2: The top 20 industries with annual incomes most affected by temporal progression from 2007-2013 (as inferred by TRENDS). The coloring delineates broader sectors: manufacturing in red, business/finance in green, energy in blue, technology in magenta.
Chapter 4

Discussion

While well-established methods exists for quantifying how much a series of probability distributions changes over a sequence (or levels taken by an associated ordinal variable), TRENDS addresses the more scientifically relevant task of quantifying how much of the change in the observed distributions may be attributed to the sequence/level progression rather than exogenous latent variables. Inspired by the way in which usual regression methods can distinguish mean-changes due to covariates of interest from spurious mean-differences through the introduction of an assumed model, TRENDS extends this capability from average-effect analysis to the entire distributions by leveraging a model in which the effects of level-progression follow a trend, an assumption we have demonstrated to be very relevant to the effects of interest in many domains.

The TRENDS model introduces many theoretical questions which we did not address in this work (as they are less relevant in our applications), such as a more comprehensive examination of the interplay between the convergence rates of our estimators and assumptions regarding: the distributions (e.g. parametric families), the noise (e.g. smoothness of the noise functions $\mathcal{E}_t(\cdot)$), and the quantile-estimation procedure. To improve the method’s practical performance, we envision a number of possible developments. Firstly, it is desirable to investigate additional forms of regularization (for example, encouraging smoothness in the evolution of adjacent quantiles over increasing $\ell$, or borrowing strength between the TRENDS fits across different
variables). Another promising extension is the incorporation of TRENDS into a Bayesian framework with (possibly infinite) priors over the size and number of the blocks of adjacent quantiles which must jointly shift in the same direction over \( \ell \) in our trend definition (our setting of unrestricted block-sizes with and most two different blocks appears to work well in practice but may be improved with suitable priors). Finally, Lemma 1 provides an alternative definition of a trend which also holds for multidimensional distributions. Methods based on this notion would be promising for tasks like geospatial modeling over time, although computing the Wasserstein-least-squares-fit is more complex in higher dimensions.

Nonetheless, we find the basic TRENDS model we have presented here can extract to a plethora of valuable insights when the goal is effect-size estimation rather than prediction. In the case of SCRS data, numerous technological innovations have been proposed which will allow biologists to profile the transcriptomes of thousands of individual cells such as the massively parallel RNA single-cell sequencing framework [55] or the currently under development Drop-Seq protocol [56]. As the number of cells and time points profiled in these experiments grows, significant discoveries may be made by analyzing the evolution of expression distributions across cell populations, and TRENDS provides a principled method to conduct this analysis.
Bibliography


Appendix A

Proofs and auxiliary lemmas

Proof of Lemma 1. For any $i < j \in \{1, \ldots, L\}$:

$$d_{L_i}(P_i, P_j) = \int_0^1 |F_{i}^{-1}(p) - F_{j}^{-1}(p)| \, dp = \int_0^1 \sum_{\ell=i+1}^{j} |F_{\ell}^{-1}(p) - F_{\ell-1}^{-1}(p)| \, dp = \sum_{\ell=i+1}^{j} d_{L_\ell}(P_{\ell-1}, P_{\ell})$$

where the second equality follows from the fact that $F_{i}^{-1}(p), F_{i+1}^{-1}(p), \ldots, F_{j}^{-1}(p)$ is assumed to be monotone for each $p$.  \hfill \Box

Proof of Lemma 2. Given any $G^{-1} \in Q$, we can define function $H : [0, 1] \to \mathbb{R}$ such that $G^{-1} = H + \frac{1}{N} \sum_{i=1}^{N} F_{i}^{-1}$. We have:

$$\sum_{j=1}^{N} \int_0^1 \left( F_{j}^{-1}(p) - G^{-1}(p) \right)^2 \, dp$$

$$= \int_0^1 \sum_{j=1}^{N} \left( F_{j}^{-1}(p) - H(p) - \frac{1}{N} \sum_{i=1}^{N} F_{i}^{-1}(p) \right)^2 \, dp$$

$$\geq \int_0^1 \sum_{j=1}^{N} \left( F_{j}^{-1}(p) - \frac{1}{N} \sum_{i=1}^{N} F_{i}^{-1}(p) \right)^2 \, dp$$

regardless of the value taken by $H(p)$ for each $p \in [0, 1]$.  \hfill \Box
Proof of Theorem 1. We have:

$$
\argmin_{G_1^{-1}, \ldots, G_L^{-1}} \left\{ \sum_{\ell=1}^L \sum_{i \in I_\ell} \sum_{k=1}^{P-1} \left( \tilde{F}_{\ell}^{-1}(p_k) - G_{\ell}^{-1}(p_k) \right)^2 \right\}
$$

where $G_1, \ldots, G_L$ follow a trend

$$
\argmin_{v^{(1)}, \ldots, v^{(L)}} \left\{ \sum_{k=1}^{P-1} \left( \frac{p_{k+1} - p_k}{2} \right) \sum_{\ell=1}^L \sum_{i \in I_\ell} w_i \left( \tilde{F}_{\ell}^{-1}(p_k) - v^{(\ell)}_k \right)^2 \right\}
$$

for $v^{(\ell)} \in \mathbb{R}^{P-1}$ with entry $v_k^{(\ell)}$ at $k$th index

s.t. $\forall k < k' \in \{1, \ldots, P-1\}$: \{ $\forall \ell : v_k^{(\ell)} < v_{k'}^{(\ell)}$ since $G_{\ell}^{-1}$ must be a valid quantile function \ $v_k^{(1)}, \ldots, v_k^{(L)}$ is a monotone sequence whose direction $= \delta[k]$ for one of the $\delta$ constructed in Step 6 or 8 of the procedure.

This is because the set of all $\delta$ considered by the TF algorithm contains every possible increasing/decreasing configuration (a mapping from $k \in \{1, \ldots, P-1\} \rightarrow \{\text{"nonincreasing", "nondecreasing"}\}$) whose corresponding quantile-sequence satisfies the second condition of the trend definition.

$$
= \argmin_{v^{(1)}, \ldots, v^{(L)}} \left\{ \sum_{k=1}^{P-1} \left( \frac{p_{k+1} - p_k}{2} \right) \sum_{\ell=1}^L w_{\ell}^* \left( \tilde{F}_{\ell}^{-1}(p_k) - v^{(\ell)}_k \right)^2 \right\}
$$

(A.1)

s.t. $\forall k < k' \in \{1, \ldots, P-1\}$: \{ $\forall \ell : v_k^{(\ell)} < v_{k'}^{(\ell)}$ since $G_{\ell}^{-1}$ must be a valid quantile function \\ $v_k^{(1)}, \ldots, v_k^{(L)}$ is a monotone sequence whose direction $= \delta[k]$ for one of the $\delta$ constructed in Step 6 or 8 of the procedure.

where we defined $w_{\ell}^* := \sum_{i \in I_\ell} w_i, \quad \tilde{F}_{\ell}^{-1}(p) := \frac{1}{w_{\ell}^*} \sum_{i \in I_\ell} w_i \tilde{F}_{\ell}^{-1}(p_k)$

We will now show that for any $\delta$ constructed in Step 6 or 8, the corresponding $y_{\ell}$ produced by the AlternatingProjections algorithm are the optimal valid quantile-functions if we impose the additional constraint that for any $k$, the $p_k$th quantile-sequence must be increasing/decreasing as specified by $\delta[k]$. Establishing this fact completes the proof because the trends-condition is simply the union of $2P$ such constraints, each of which is tested by the TF procedure. Therefore, one of corresponding $y_1, \ldots, y_L$ sequences must be the global minimum.

Having fixed an increasing/decreasing configuration $\delta$, let $\mathcal{H}$ denote the Hilbert space of all $L \times (P - 1)$ matrices, and $\mathcal{X}$ be the vector-space of all sequences (a.k.a. $L \times (P - 1)$ matrices) [v^{(1)}, \ldots, v^{(L)}] s.t. $\forall \ell \in \{1, \ldots, L\}, k \in \{1, \ldots, P-1\} : v^{(\ell)} \in \mathbb{R}^{P-1}$ and $v_1^{(\ell)}, \ldots, v_{P-1}^{(\ell)}$ is a nondecreasing sequence. Similarly, define $\mathcal{Y}$ to be the vector-space of all sequences [v^{(1)}, \ldots, v^{(L)}] s.t. $\forall \ell, k : v^{(\ell)} \in \mathbb{R}^{P-1}$ and $v_k^{(1)}, \ldots, v_k^{(L)}$ is a monotone sequence which is increasing if and only if $\delta[k]$ specifies it. Finally, we
also define the following metric over these sequences

\[
d_W ([v^{(1)}, \ldots, v^{(L)}], [w^{(1)}, \ldots, w^{(L)}]) = \sum_{k=1}^{P-1} \left( \frac{p_{k+1} - p_{k-1}}{2} \right) \sum_{\ell=1}^{L} w^*_\ell \left( v^{(\ell)}_k - w^{(\ell)}_k \right)^2
\]

(A.2)

Lemmas 3 and 4 show that our AlternatingProjections algorithm is equivalent to Dykstra’s method of alternating projections [35] between \(\mathcal{X}\) and \(\mathcal{Y}\) under metric \(d_W\). Furthermore, both \(\mathcal{X}\) and \(\mathcal{Y}\) are closed and convex, and the initial point (i.e. sequence) \([x^{(1)}, \ldots, x^{(L)}]\) must lie in \(\mathcal{X}\) because \(\forall \ell, k\): the TF algorithm initializes \(x^{(\ell)}\) as a (weighted) average of valid quantile-functions (assuming the quantile-estimators do not produce invalid quantile-functions), and thus itself must be nondecreasing in \(k\). Therefore, we can apply the celebrated result stated in [35, 57] which implies that Dykstra’s algorithm must converge to the projection of the initial-sequence onto \(\mathcal{X}\cap \mathcal{Y}\). By construction, this projection (under metric \(d_W\)) exactly corresponds to the solution of the constrained optimization in (2.12) under the additional constraint imposed by \(\delta\).

Lemma 3. Recall the definitions from the TF algorithm and the proof of Theorem 1. Given any \([x^{(1)}, \ldots, x^{(L)}] \in \mathcal{X}\), its projection onto \(\mathcal{Y}\) under metric \(d_W\), \([y^{(1)}, \ldots, y^{(L)}]\), may be computed \(\forall k \in \{1, \ldots, P - 1\}\) as

\[
y^{(1)}_k, \ldots, y^{(L)}_k = \text{PAVA}\left((x^{(1)}_k, w_1), \ldots, (x^{(L)}_k, w_L); \delta[k]\right)
\]

Proof of Lemma 3. Choose any \([z^{(1)}, \ldots, z^{(L)}] \in \mathcal{Y}\). By consequence of Fact 1

\[
\text{PAVA}\left((x^{(1)}_k, w_1^*), \ldots, (x^{(L)}_k, w_L^*); \delta[k]\right)
\]

\[
= \arg\min_{\text{monotone } \lambda_1, \ldots, \lambda_L} \left\{ \sum_{\ell=1}^{L} w^*_\ell \left( x^{(\ell)}_k - \lambda_\ell \right)^2 \right\}
\]

where the \(\lambda_\ell\) are only increasing if specified by \(\delta[k]\)

\[
\Rightarrow \sum_{\ell=1}^{L} w^*_\ell \left( y^{(\ell)}_k - x^{(\ell)}_k \right)^2 \leq \sum_{\ell=1}^{L} w^*_\ell \left( z^{(\ell)}_k - x^{(\ell)}_k \right)^2 \quad \forall k
\]

since \(z^{(1)}_k, \ldots, z^{(L)}_k\) have monotonicity specified by \(\delta\)

\[
\Rightarrow \sum_{k=1}^{P-1} \left( \frac{p_{k+1} - p_{k-1}}{2} \right) \sum_{\ell=1}^{L} w^*_\ell \left( y^{(\ell)}_k - x^{(\ell)}_k \right)^2 \leq \sum_{k=1}^{P-1} \left( \frac{p_{k+1} - p_{k-1}}{2} \right) \sum_{\ell=1}^{L} w^*_\ell \left( z^{(\ell)}_k - x^{(\ell)}_k \right)^2
\]

\(\square\)
Lemma 4. Recall the definitions from the TF algorithm and the proof of Theorem 1. Given any \([y^{(1)}, \ldots, y^{(L)}] \in \mathcal{Y}\), its projection onto \(\mathcal{X}\) under metric \(d_W\), \([x^{(1)}, \ldots, x^{(L)}]\), may be computed \(\forall \ell \in \{1, \ldots, L\}\) as

\[
x^{(\ell)}_1, \ldots, x^{(\ell)}_{P-1} = \text{PAVA} \left( \left( y^{(\ell)}_1, \frac{p_2 - p_0}{2} \right), \ldots, \left( y^{(\ell)}_{P-1}, \frac{p_P - p_{P-2}}{2} \right); \text{“nondecreasing”} \right)
\]

Proof of Lemma 4. Choose any \([z^{(1)}, \ldots, z^{(L)}] \in \mathcal{X}\). By Fact 1

\[
\text{PAVA} \left( \left( y^{(\ell)}_1, \frac{p_2 - p_0}{2} \right), \ldots, \left( y^{(\ell)}_{P-1}, \frac{p_P - p_{P-2}}{2} \right); \text{“nondecreasing”} \right) = \arg\min_{\lambda_1 \leq \cdots \leq \lambda_{P-1}} \left\{ \sum_{k=1}^{P-1} \left( \frac{p_{k+1} - p_{k-1}}{2} \right) \left( y^{(\ell)}_k - \lambda_k \right)^2 \right\} \text{ for each } \ell
\]

\[
\implies \sum_{k=1}^{P-1} \left( \frac{p_{k+1} - p_{k-1}}{2} \right) \left( x^{(\ell)}_k - y^{(\ell)}_k \right)^2 \leq \sum_{k=1}^{P-1} \left( \frac{p_{k+1} - p_{k-1}}{2} \right) \left( z^{(\ell)}_k - y^{(\ell)}_k \right)^2 \quad \forall \ell
\]

since \([z^{(1)}, \ldots, z^{(L)}] \in \mathcal{X} \implies \forall \ell: z^{(\ell)}_1 \leq \cdots \leq z^{(\ell)}_{P-1}\)

\[
\implies \sum_{k=1}^{P-1} \left( \frac{p_{k+1} - p_{k-1}}{2} \right) \sum_{\ell=1}^{L} w_\ell^* \left( x^{(\ell)}_k - y^{(\ell)}_k \right)^2 \leq \sum_{k=1}^{P-1} \left( \frac{p_{k+1} - p_{k-1}}{2} \right) \sum_{\ell=1}^{L} w_\ell^* \left( x^{(\ell)}_k - z^{(\ell)}_k \right)^2
\]

\[\square\]
Proof of Theorem 2. Recalling that \( G^{-1}(p) \) denotes the \( p \)th quantile of \( Q_\ell \equiv f(\ell) \), we also define:

\[
\bar{F}_\ell^{-1}(p) := \frac{1}{N_\ell} \sum_{i \in I_\ell} F_i^{-1}(p)
\]  

(A.3)

By a standard application of the Chernoff bound [36, 37]:

\[
\Pr\left( |\bar{F}_\ell^{-1}(p) - G_\ell^{-1}(p)| > \eta \right) = \Pr\left( \left| \frac{1}{N_\ell} \sum_{i \in I_\ell} \xi_i(p) \right| > \eta \right) \leq 2 \exp\left( -\frac{\eta^2 N_\ell}{2 \sigma^2} \right) \quad \forall \eta > 0
\]

Recall that we compute the Wasserstein integral using \( P - 1 \) equally-spaced quantiles and the midpoint approximation, so

\[
d \left( \bar{F}_\ell^{-1}, G_\ell^{-1} \right)^2 \approx d_W \left( \bar{F}_\ell^{-1}, G_\ell^{-1} \right)^2 = \sum_{k=1}^{P-1} \frac{1}{P} (\bar{F}_\ell^{-1}(k/P) - G_\ell^{-1}(k/P))^2
\]

\[
\Pr\left( \sum_{\ell=1}^{L} d_W \left( \bar{F}_\ell^{-1}, G_\ell^{-1} \right)^2 > \eta \right) \leq \sum_{\ell=1}^{L} \sum_{k=1}^{P-1} \Pr\left( \frac{1}{P} (\bar{F}_\ell^{-1}(k/P) - G_\ell^{-1}(k/P))^2 > \frac{\eta}{PL} \right)
\]

by a union-bound

\[
= L \cdot P \cdot \Pr\left( |\bar{F}_\ell^{-1}(k/P) - G_\ell^{-1}(k/P)| > \sqrt{\frac{\eta}{L}} \right)
\]

\[
\leq 2PL \exp\left( -\frac{\eta N_\ell}{2 \sigma^2 L} \right) 
\]

(A.4)

Note that \( \hat{G}_1^{-1}, \ldots, \hat{G}_L^{-1} \) form the best trending approximation to the \( F_1^{-1} \) by Theorem 1, and since \( G_1^{-1}, \ldots, G_L^{-1} \) are valid quantile functions which also follow a trend, this implies:

\[
\sum_{\ell=1}^{L} \sum_{i \in I_\ell} d_W \left( F_i^{-1}, \hat{G}_\ell^{-1} \right)^2 \leq \sum_{\ell=1}^{L} \sum_{i \in I_\ell} d_W \left( F_i^{-1}, G_\ell^{-1} \right)^2
\]

\[
\Rightarrow \sum_{\ell=1}^{L} d_W \left( \bar{F}_\ell^{-1}, \hat{G}_\ell^{-1} \right)^2 \leq \sum_{\ell=1}^{L} d_W \left( \bar{F}_\ell^{-1}, G_\ell^{-1} \right)^2 \quad \text{by Lemma 2}
\]

\[
\Rightarrow \forall \ell : d_W \left( \bar{F}_\ell^{-1}, \hat{G}_\ell^{-1} \right)^2 \leq \sum_{\ell=1}^{L} d_W \left( \bar{F}_\ell^{-1}, G_\ell^{-1} \right)^2
\]

Thus, by the triangle-inequality:

\[
d_W \left( \hat{G}_\ell^{-1}, G_\ell^{-1} \right) \leq d_W \left( \bar{F}_\ell^{-1}, G_\ell^{-1} \right) + d_W \left( \bar{F}_\ell^{-1}, \hat{G}_\ell^{-1} \right) \leq 2 \left[ \sum_{\ell=1}^{L} d_W \left( \bar{F}_\ell^{-1}, G_\ell^{-1} \right)^2 \right]^{1/2} \quad \forall \ell
\]

85
which implies \( \forall \epsilon > 0 \) we can combine this result with (A.4) setting \( \eta := \epsilon^2/4 \) to get:

\[
\Pr \left( \exists \ell : d_W(\hat{G}_\ell^{-1}, G_\ell^{-1}) > \epsilon \right) \leq \Pr \left( \sum_{\ell=1}^{L} d_W(\tilde{F}_\ell^{-1}, G_\ell^{-1})^2 > \frac{\epsilon^2}{4} \right) \leq 2PL \exp \left( -\frac{\epsilon^2 N_\ell}{8\sigma^2 L} \right)
\]

Proof of Theorem 3. We proceed similarly as in the proof of Theorem 2. Defining

\[
\overline{F}_\ell^{-1}(p) := \frac{1}{N_\ell} \sum_{i \in I_\ell} \hat{F}_i^{-1}(p)
\]

by Theorem 2.12 and Lemma 2, we have:

\[
\sum_{\ell=1}^{L} d_W \left( \hat{G}_\ell^{-1}, \overline{F}_\ell^{-1} \right)^2 \leq \sum_{\ell=1}^{L} d_W \left( \hat{G}_\ell^{-1}, \overline{F}_\ell^{-1} \right)^2 \Rightarrow d_W \left( \hat{G}_\ell^{-1}, \overline{F}_\ell^{-1} \right)^2 \leq \sum_{\ell=1}^{L} d_W \left( \hat{G}_\ell^{-1}, \overline{F}_\ell^{-1} \right)^2 \quad \forall \ell
\]

since \( G_1^{-1}, \ldots, G_L^{-1} \) are valid quantile functions which follow a trend. Thus:

\[
\forall \ell : d_W \left( \hat{G}_\ell^{-1}, G_\ell^{-1} \right) \leq d_W \left( \hat{G}_\ell^{-1}, \overline{F}_\ell^{-1} \right) + d_W \left( \overline{F}_\ell^{-1}, G_\ell^{-1} \right) \quad \text{by the triangle-inequality}
\]

\[
\leq 2 \left[ \sum_{\ell=1}^{L} d_W \left( \overline{F}_\ell^{-1}, G_\ell^{-1} \right)^2 \right]^{1/2}
\]

\[
\leq 2 \left[ \sum_{\ell=1}^{L} \left( d_W \left( \overline{F}_\ell^{-1}, G_\ell^{-1} \right) + d_W \left( \overline{F}_\ell^{-1}, \overline{F}_\ell^{-1} \right) \right) \right]^{1/2} \quad \text{by the triangle-inequality}
\]

\[
\leq 2\sqrt{2} \left[ \sum_{\ell=1}^{L} d_W \left( \overline{F}_\ell^{-1}, G_\ell^{-1} \right)^2 + \sum_{\ell=1}^{L} d_W \left( \overline{F}_\ell^{-1}, \overline{F}_\ell^{-1} \right)^2 \right]^{1/2} \quad \text{by Cauchy-Schwartz}
\]

Therefore \( \forall \epsilon > 0 \):

\[
\Pr \left( \exists \ell : d_W \left( \hat{G}_\ell^{-1}, G_\ell^{-1} \right) > \epsilon \right) \leq \Pr \left( \sum_{\ell=1}^{L} d_W \left( \overline{F}_\ell^{-1}, G_\ell^{-1} \right)^2 + \sum_{\ell=1}^{L} d_W \left( \overline{F}_\ell^{-1}, \overline{F}_\ell^{-1} \right)^2 > \frac{\epsilon^2}{8} \right)
\]

\[
\leq \Pr \left( \sum_{\ell=1}^{L} d_W \left( \overline{F}_\ell^{-1}, G_\ell^{-1} \right)^2 > \frac{\epsilon^2}{16} \right) + \Pr \left( \sum_{\ell=1}^{L} d_W \left( \overline{F}_\ell^{-1}, \overline{F}_\ell^{-1} \right)^2 > \frac{\epsilon^2}{16} \right) \quad \text{by the union-bound}
\]

and we can use (A.4) to bound the first summand, resulting in the following bound

86
Finally, Lemma 6 implies:

\[
\Pr \left( \sum_{\ell=1}^{L} d_W \left( \frac{\tilde{G}_\ell^{-1} \cdot G_\ell^{-1}}{2N_\ell P L \exp} \right) > \frac{\epsilon^2}{16} \right) \leq 2N_\ell P L \exp \left( -2nR \left( \frac{\epsilon}{4\sqrt{L}} \right)^2 \right)
\]

which produces the desired bound when combined with (A.6).

Proof of Theorem 4. By Lemma 7, (A.7) ⇒ (A.8), so we only need to show the result assuming (A.8) holds. Lemma 8 then implies:

\[
\Pr \left( \sum_{\ell=1}^{L} d_W \left( \frac{\tilde{F}_\ell^{-1} \cdot F_\ell^{-1}}{2N_\ell} \right) > \frac{\epsilon^2}{16} \right) \leq 2P \exp \left( -\frac{c^2}{8} n L n \epsilon^2 \right)
\]

Note that the bound in (A.6) only requires the assumptions from Theorem 2, so we can combine it with the above expression to obtain the desired bound.
Proof of Theorem 5.

Consider \( \Pr \left( \hat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon \right) \)

= \( \Pr \left( \hat{F}_i \left( F_i^{-1}(k/P) + \epsilon \right) \leq \frac{k}{P} \right) \)

= \( \Pr \left( \sum_{j=1}^{n} \mathbb{1} \left[ X_{ij} \leq F_i^{-1}(k/P) + \epsilon \right] \leq \frac{n \epsilon}{P} \right) \) (A.7)

This is the CDF evaluated at \( \tilde{x} := \frac{n \epsilon}{P} \) of a binomial random variable with success probability \( \tilde{p} := F_i \left( F_i^{-1}(k/P) + \epsilon \right) \) in \( n \) trials.

Now assume \( \epsilon + F_i^{-1}(k/P) \geq B > 0 \), which implies \( n \tilde{p} \geq \tilde{x} \).

Letting \( D(\alpha \parallel \beta) \) denote the relative entropy between the Bernoulli(\( \alpha \)) and Bernoulli(\( \beta \)) distributions, we can thus apply a tail-inequality for the binomial CDF which [58] derived from the Chernoff bound to upper-bound (A.7) by

\[
\leq \exp \left( -n D \left( \frac{\tilde{x}}{n} \parallel \tilde{p} \right) \right)
\]

= \( \exp \left( -n \left[ \frac{\tilde{x}}{n} \log \left( \frac{\tilde{x}/n}{\tilde{p}} \right) + \left( 1 - \frac{\tilde{x}}{n} \right) \log \left( \frac{1 - \tilde{x}/n}{1 - \tilde{p}} \right) \right] \) \)

= \( \exp \left( -n \left[ \frac{k}{P} \log \left( \frac{k/P}{F_i \left( F_i^{-1}(k/P) + \epsilon \right)} \right) + \left( 1 - \frac{k}{P} \right) \log \left( \frac{1 - k/P}{1 - F_i \left( F_i^{-1}(k/P) + \epsilon \right)} \right) \right] \) \)

\leq \( \exp \left( -n \left[ \frac{k}{P} \log \left( \frac{k}{P} \right) + \left( 1 - \frac{k}{P} \right) \log \left( \frac{1 - k/P}{1 - F_i \left( F_i^{-1}(k/P) + \epsilon \right)} \right) \right] \) since \( F_i(\cdot) \leq 1 \)

= \( e^{-nC(k)} \cdot \exp \left( n \left( 1 - \frac{k}{P} \right) \log \left( 1 - F_i \left( F_i^{-1}(k/P) + \epsilon \right) \right) \right) \)

where \( C(k) := \frac{k}{P} \log \left( \frac{k}{P} \right) + \left( 1 - \frac{k}{P} \right) \log \left( 1 - \frac{k}{P} \right) \geq -1 \)

\[
\leq e^n \cdot \exp \left( n \left( 1 - \frac{k}{P} \right) \log \left( 1 - F_i \left( F_i^{-1}(k/P) + \epsilon \right) \right) \right)
\]

since the fact \( \log x \geq \frac{x - 1}{x} \ \forall x > 0 \) implies \( C(k) \geq -1 \ \forall k \in \{1, \ldots, P - 1\} \)

\[
\leq e^{-n} \cdot \exp \left( n \left( 1 - \frac{k}{P} \right) \log \left( 1 - z \right) \right) \text{ where } z := 1 - \exp \left( -a \left( F_i^{-1}(k/P) + \epsilon - B + b \right)^2 \right) \]

because \( 1 - k/P > 0 \) and by (A.9): \( F_i \left( F_i^{-1}(k/P) + \epsilon \right) \geq z \)

since we’ve assumed \( F_i^{-1}(k/P) + \epsilon \geq B \).
\[
\begin{align*}
&= e^{-n} \cdot \exp \left( -2an \left( 1 - \frac{k}{P} \right) \left( F_i^{-1}(k/P) + \epsilon - B + b \right)^2 \right) \\
&\leq e^{-n} \cdot \exp \left( -2an \left( 1 - \frac{k}{P} \right) \min \left\{ b^2, \frac{(B - F_i^{-1}(k/P))^2}{(B - F_i^{-1}(k/P))^2} \right\} \epsilon^2 \right)
\end{align*}
\]

because \( \epsilon \geq B - F_i^{-1}(k/P) \) implies
\[
\min \left\{ b^2, \frac{(B - F_i^{-1}(k/P))^2}{(B - F_i^{-1}(k/P))^2} \right\} \epsilon^2 \leq (F_i^{-1}(k/P) + \epsilon - B + b)^2
\]

\[
= \exp \left( -n \left[ 2a \left( 1 - \frac{k}{P} \right) \min \left\{ b^2, \frac{(B - F_i^{-1}(k/P))^2}{(B - F_i^{-1}(k/P))^2} \right\} \epsilon^2 - 1 \right] \right)
\]

\[
\leq \exp \left( -n \left( 2a \left( 1 - \frac{k}{P} \right) \min \left\{ b^2, \frac{(B - F_i^{-1}(k/P))^2}{(B - F_i^{-1}(k/P))^2} \right\} - 1 \right) \epsilon^2 \right)
\]

since we assumed \( \epsilon \geq B - F_i^{-1}(k/P) \)

\[
\leq \exp \left( -n \left( 2a \left( 1 - \frac{k}{P} \right) \frac{b^2 - 1}{4B^2} \right) \epsilon^2 \right)
\]

because by (A.9) and (A.11):

\[
-F_i^{-1}(k/P) \leq B \text{ and } 0 < b \leq B
\]

And finally, we can use the fact that \( k \leq P - 1 \) to obtain the following bound

\[
\Pr \left( \tilde{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon \right) \leq \exp \left( -n \left( \frac{2ab^2 - 1}{4PB^2} \right) \epsilon^2 \right) \tag{A.8}
\]

Following the proof of Lemma 7, one can show that (A.9) implies

\[
\Pr \left( \tilde{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon \right) \leq \exp(-2nc^2\epsilon^2) \text{ if } 0 < \epsilon < B - F_i^{-1}(k/P) \tag{A.9}
\]

Combining (A.9) with (A.8), we thus have

\[
\Pr \left( \tilde{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon \right) \leq \exp \left( -nre^2 \right) \forall \epsilon > 0
\]

where \( r := \min \left\{ 2c^2, \frac{2ab^2 - 1}{4PB^2} \right\} > 0 \) by (A.10).

One can show by an identical argument that

\[
\Pr \left( F_i^{-1}(k/P) - \tilde{F}_i^{-1}(k/P) > \epsilon \right) \leq \exp \left( -nre^2 \right) \forall \epsilon > 0
\]
and therefore
\[ \Pr \left( \left| \hat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) \right| > \epsilon \right) \leq 2 \exp \left( -n \epsilon^2 \right) \quad \forall \epsilon > 0 \quad (A.10) \]

\( \hat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) \) is thus sub-Gaussian with parameter \( \frac{1}{2nr} \) and independent of \( \hat{F}_j^{-1}(k/P) - F_j^{-1}(k/P) \) \( \forall j \neq i \) because we assumed the simple quantile-estimator defined in (A.6) is used. Following the proof of Lemma 8, \( \forall \gamma > 0 \):

\[ \Pr \left( \sum_{\ell=1}^{L} d_W \left( \hat{F}_\ell^{-1}, \bar{F}_\ell^{-1} \right)^2 > \frac{\epsilon^2}{16} \right) \leq 2P \exp \left( - \frac{\tau}{16} N \epsilon^2 \right) \quad (A.11) \]

Note that the bound in (A.6) only requires the assumptions from Theorem 2, so we can combine it with the above inequality to obtain the desired bound. \( \square \)

**Lemma 5** (From Serfling [59]: Theorem 2.3.2 on page 75). For \( p \in (0, 1) \): if \( \exists \) unique \( x \) s.t. \( F(x) = p \) and \( \hat{F}^{-1}(p) \) is estimated using \( n \) i.i.d. samples from CDF \( F_i \), then \( \forall \gamma > 0 \):

\[ \Pr \left( \left| \hat{F}_i^{-1}(p) - F_i^{-1}(p) \right| > \gamma \right) \leq 2 \exp \left( -2n R(\gamma, i, p)^2 \right) \]

where \( R(\gamma, i, p) := \min \left\{ F_i \left( F_i^{-1}(p) + \gamma \right) - p, \ p - F_i \left( F_i^{-1}(p) - \gamma \right) \right\} \)
**Lemma 6.** Under the assumptions of Theorem 3 and definitions (2.14), (A.3), (A.5)

\[ \forall \gamma > 0 : \quad \Pr \left( \sum_{\ell=1}^{L} d_{W} \left( \tilde{F}_{\ell}^{-1}, \tilde{F}_{\ell}^{-1} \right)^{2} > \gamma \right) \leq 2N_{\ell}PL \exp \left( -2nR \left( \sqrt{\gamma/L} \right)^{2} \right) \]

**Proof of Lemma 6.**

\[
\Pr \left( \sum_{\ell=1}^{L} d_{W} \left( \tilde{F}_{\ell}^{-1}, \tilde{F}_{\ell}^{-1} \right)^{2} > \gamma \right) \\
= \Pr \left( \sum_{\ell=1}^{L} \frac{1}{N_{\ell}} \sum_{i \in \ell} \sum_{k=1}^{P-1} \left( \tilde{F}_{i}^{-1}(k/P) - F_{i}^{-1}(k/P) \right)^{2} > \gamma \right) \\
\leq N_{\ell}L \sum_{k=1}^{P-1} \Pr \left( \left| \tilde{F}_{i}^{-1}(k/P) - F_{i}^{-1}(k/P) \right| > \sqrt{\gamma/L} \right) \quad \text{by the union-bound} \\
\leq 2N_{\ell}L \sum_{k=1}^{P-1} \exp \left( -2nR \left( \sqrt{\gamma/L}, i, k/P \right)^{2} \right) \quad \text{by (A.5) and Lemma 5} \\
\leq 2N_{\ell}LP \exp \left( -2nR \left( \sqrt{\gamma/L} \right)^{2} \right) \quad \text{by definition (2.14)}
\]

\[ \square \]

**Lemma 7.** If we assume (A.4) and (A.5), then condition (A.7) implies condition (A.8).

**Proof of Lemma 7.** Assume WLOG that \( F_{i}^{-1}(k/P) \geq 0 \) and note that \( F_{i}^{-1}(k/P) \leq B \) by (A.7).

Then, by a bound established in the proof of Lemma 5 given in [59], \( \forall \epsilon > 0 : \)

\[ \Pr \left( \tilde{F}_{i}^{-1}(k/P) - F_{i}^{-1}(k/P) > \epsilon \right) \leq \exp \left( -2nR(\epsilon, i, k/P)^{2} \right) \quad (A.12) \]

and

\[ \Pr \left( F_{i}^{-1}(k/P) - \tilde{F}_{i}^{-1}(k/P) > \epsilon \right) \leq \exp \left( -2nR(\epsilon, i, k/P)^{2} \right) \quad (A.13) \]

By (A.7): \( f_{i}(x) = \frac{d}{dx} F_{i}(x) \geq c \forall x \in (-B, B) \) which implies

\[ R(\gamma, i, p) \geq c\gamma > 0 \quad \text{if} \quad F_{i}^{-1}(p) \pm \gamma \in (-B, B) \quad (A.14) \]

because recall that we defined \( R(\gamma, i, p) := \min \{ F_{i} \left( F_{i}^{-1}(p) + \gamma \right) - p, p - F_{i} \left( F_{i}^{-1}(p) - \gamma \right) \} \).

Together with (A.14), (A.12) and (A.13) imply

\[ \Pr \left( \tilde{F}_{i}^{-1}(k/P) - F_{i}^{-1}(k/P) > \epsilon \right) \leq \exp \left( -2nc^{2}\epsilon^{2} \right) \quad \text{if} \quad F_{i}^{-1}(k/P) + \epsilon < B \quad (A.15) \]

91
and

\[
\Pr \left( F_i^{-1}(k/P) - \hat{F}_i^{-1}(k/P) > \epsilon \right) \leq \exp(-2nc^2\epsilon^2) \quad \text{if } F_i^{-1}(k/P) - \epsilon > -B \quad (A.16)
\]

Note that because \( f_i(x) = 0 \ \forall x \geq B \), we have

\[
\Pr \left( \hat{F}_i^{-1}(k/P) > F_i^{-1}(k/P) + \epsilon \right) = 0 \quad \text{if } \epsilon \geq B - F_i^{-1}(k/P)
\]

\[
\implies \Pr \left( \hat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon \right) = 0 \quad \text{if } \epsilon \geq B - F_i^{-1}(k/P) \quad (A.17)
\]

as well as

\[
\Pr \left( \hat{F}_i^{-1}(k/P) < F_i^{-1}(k/P) - \epsilon \right) = 0 \quad \text{if } \epsilon \geq B + F_i^{-1}(k/P)
\]

\[
\implies \Pr \left( F_i^{-1}(k/P) - \hat{F}_i^{-1}(k/P) > \epsilon \right) = 0 \quad \text{if } \epsilon \geq B + F_i^{-1}(k/P) \quad (A.18)
\]

Putting together (A.15), (A.16), (A.17), and (A.18), we thus have

\[
\Pr \left( \hat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon \right) \leq \exp(-2nc^2\epsilon^2) \quad \forall \epsilon > 0
\]

and

\[
\Pr \left( F_i^{-1}(k/P) - \hat{F}_i^{-1}(k/P) > \epsilon \right) \leq \exp(-2nc^2\epsilon^2) \quad \forall \epsilon > 0
\]

which implies

\[
\Pr \left( \left| F_i^{-1}(k/P) - \hat{F}_i^{-1}(k/P) \right| > \epsilon \right) \leq 2\exp(-2nc^2\epsilon^2) \quad \forall \epsilon > 0
\]

\( \Box \)
Lemma 8. Under condition (A.8) and definitions (2.14), (A.3), (A.5)

\[ \forall \gamma > 0 : \quad \Pr \left( \sum_{\ell=1}^{L} d_{W} \left( \hat{F}^{-1}_{\ell}, \bar{F}^{-1}_{\ell} \right)^{2} > \gamma \right) \leq 2P \exp \left( -2nc^{2}N_{\ell}\gamma \right) \]

Proof of Lemma 8.

\[
\begin{align*}
\Pr \left( \sum_{\ell=1}^{L} d_{W} \left( \hat{F}^{-1}_{\ell}, \bar{F}^{-1}_{\ell} \right)^{2} > \gamma \right) &= \Pr \left( \frac{1}{LN_{\ell}} \sum_{\ell=1}^{L} \sum_{i \in I_{\ell}} \sum_{k=1}^{P-1} \frac{1}{P} \left( \hat{F}^{-1}_{i}(k/P) - F^{-1}_{i}(k/P) \right)^{2} > \frac{\gamma}{L} \right) \\
&\leq \sum_{k=1}^{P-1} \Pr \left( \left| \frac{1}{LN_{\ell}} \sum_{\ell=1}^{L} \sum_{i \in I_{\ell}} \hat{F}^{-1}_{i}(k/P) - F^{-1}_{i}(k/P) \right| > \sqrt{\frac{\gamma}{L}} \right) \quad \text{by the union-bound} \\
&\leq 2 \sum_{k=1}^{P-1} \exp \left( -2nc^{2}LN_{\ell}\sqrt{\frac{\gamma}{L}} \right) = 2P \exp \left( -2nc^{2}N_{\ell}\gamma \right)
\end{align*}
\]

where in the last inequality, we have used the fact that (A.8) implies the \( \hat{F}^{-1}_{i}(k/P) - F^{-1}_{i}(k/P) \) are independent sub-Gaussian random variables with parameter \( \frac{1}{4nc^{2}} \), so the inequality follows from a standard application of the Chernoff bound [36,37]. \( \square \)
Appendix B

Supplementary Tables
<table>
<thead>
<tr>
<th>Gene Ontology ID</th>
<th>Annotation Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 GO:0048745</td>
<td>smooth muscle tissue development</td>
</tr>
<tr>
<td>2 GO:0048747</td>
<td>muscle fiber development</td>
</tr>
<tr>
<td>3 GO:0048742</td>
<td>regulation of skeletal muscle fiber development</td>
</tr>
<tr>
<td>4 GO:0048739</td>
<td>cardiac muscle fiber development</td>
</tr>
<tr>
<td>5 GO:0048635</td>
<td>negative regulation of muscle organ development</td>
</tr>
<tr>
<td>6 GO:0007517</td>
<td>muscle organ development</td>
</tr>
<tr>
<td>7 GO:0007519</td>
<td>skeletal muscle tissue development</td>
</tr>
<tr>
<td>8 GO:0048743</td>
<td>positive regulation of skeletal muscle fiber development</td>
</tr>
<tr>
<td>9 GO:0048738</td>
<td>cardiac muscle tissue development</td>
</tr>
<tr>
<td>10 GO:0055013</td>
<td>cardiac muscle cell development</td>
</tr>
<tr>
<td>11 GO:0048741</td>
<td>skeletal muscle fiber development</td>
</tr>
<tr>
<td>12 GO:0055014</td>
<td>atrial cardiac muscle cell development</td>
</tr>
<tr>
<td>13 GO:0055015</td>
<td>ventricular cardiac muscle cell development</td>
</tr>
<tr>
<td>14 GO:0048643</td>
<td>positive regulation of skeletal muscle tissue development</td>
</tr>
<tr>
<td>15 GO:0097084</td>
<td>vascular smooth muscle cell development</td>
</tr>
<tr>
<td>16 GO:0060948</td>
<td>cardiac vascular smooth muscle cell development</td>
</tr>
<tr>
<td>17 GO:0055001</td>
<td>muscle cell development</td>
</tr>
<tr>
<td>18 GO:0055026</td>
<td>negative regulation of cardiac muscle tissue development</td>
</tr>
<tr>
<td>19 GO:0045843</td>
<td>negative regulation of striated muscle tissue development</td>
</tr>
<tr>
<td>20 GO:0016202</td>
<td>regulation of striated muscle tissue development</td>
</tr>
<tr>
<td>21 GO:0048642</td>
<td>negative regulation of skeletal muscle tissue development</td>
</tr>
<tr>
<td>22 GO:0055024</td>
<td>regulation of cardiac muscle tissue development</td>
</tr>
<tr>
<td>23 GO:0061049</td>
<td>cell growth involved in cardiac muscle cell development</td>
</tr>
<tr>
<td>24 GO:0014706</td>
<td>striated muscle tissue development</td>
</tr>
<tr>
<td>25 GO:0007525</td>
<td>somatic muscle development</td>
</tr>
<tr>
<td>26 GO:0061052</td>
<td>negative regulation of cell growth involved in cardiac muscle cell development</td>
</tr>
<tr>
<td>27 GO:0045844</td>
<td>positive regulation of striated muscle tissue development</td>
</tr>
<tr>
<td>28 GO:0014707</td>
<td>branchiomereric skeletal muscle development</td>
</tr>
<tr>
<td>29 GO:0007522</td>
<td>visceral muscle development</td>
</tr>
<tr>
<td>30 GO:0048641</td>
<td>regulation of skeletal muscle tissue development</td>
</tr>
<tr>
<td>31 GO:1901863</td>
<td>positive regulation of muscle tissue development</td>
</tr>
<tr>
<td>32 GO:0072208</td>
<td>metanephric smooth muscle tissue development</td>
</tr>
<tr>
<td>33 GO:0003229</td>
<td>ventricular cardiac muscle tissue development</td>
</tr>
<tr>
<td>34 GO:0060538</td>
<td>skeletal muscle organ development</td>
</tr>
<tr>
<td>35 GO:0061050</td>
<td>regulation of cell growth involved in cardiac muscle cell development</td>
</tr>
<tr>
<td>36 GO:0055020</td>
<td>positive regulation of cardiac muscle fiber development</td>
</tr>
<tr>
<td>37 GO:0061061</td>
<td>muscle structure development</td>
</tr>
<tr>
<td>38 GO:0061051</td>
<td>positive regulation of cell growth involved in cardiac muscle cell development</td>
</tr>
<tr>
<td>39 GO:0055002</td>
<td>striated muscle cell development</td>
</tr>
<tr>
<td>40 GO:0060537</td>
<td>muscle tissue development</td>
</tr>
<tr>
<td>41 GO:0007527</td>
<td>adult somatic muscle development</td>
</tr>
<tr>
<td>42 GO:0002074</td>
<td>extraocular skeletal muscle development</td>
</tr>
</tbody>
</table>

Table S1: A list of all GO annotation terms containing both the words “muscle” and “development”, used to produce the pseudo-sensitivity plots in Figure 3-4A.
<table>
<thead>
<tr>
<th>Gene Ontology ID</th>
<th>Annotation Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 GO:0007420</td>
<td>brain development</td>
</tr>
<tr>
<td>2 GO:0030900</td>
<td>forebrain development</td>
</tr>
<tr>
<td>3 GO:0021987</td>
<td>cerebral cortex development</td>
</tr>
<tr>
<td>4 GO:0048666</td>
<td>neuron development</td>
</tr>
<tr>
<td>5 GO:0030182</td>
<td>neuron differentiation</td>
</tr>
<tr>
<td>6 GO:0021895</td>
<td>cerebral cortex neuron differentiation</td>
</tr>
<tr>
<td>7 GO:0048854</td>
<td>brain morphogenesis</td>
</tr>
<tr>
<td>8 GO:0048853</td>
<td>forebrain morphogenesis</td>
</tr>
<tr>
<td>9 GO:0022008</td>
<td>neurogenesis</td>
</tr>
</tbody>
</table>

**Table S2:** A list of the GO annotation terms relevant to the somatosensory cortex data, used to produce the pseudo-sensitivity plots in Figure 3-4B.