Modeling Persistent Trends in Distributions

Jonas Mueller, Tommi Jaakkola, and David Gifford

MIT Computer Science & Artificial Intelligence Laboratory

Cambridge, MA 02139

Abstract

We present a nonparametric framework to model a short sequence of probability distributions that vary both due to underlying effects of sequential progression and confounding noise. To distinguish between these two types of variation and estimate the sequential-progression effects, our approach leverages an assumption that these effects follow a persistent trend. This work is motivated by the recent rise of single-cell RNA-sequencing experiments over a brief time course, which aim to identify genes relevant to the progression of a particular biological process across diverse cell populations. While classical statistical tools focus on scalar-response regression or order-agnostic differences between distributions, it is desirable in this setting to consider both the full distributions as well as the structure imposed by their ordering. We introduce a new regression model for ordinal covariates where responses are univariate distributions and the underlying relationship reflects consistent changes in the distributions over increasing levels of the covariate. This concept is formalized as a *trend* in distributions, which we define as an evolution that is linear under the Wasserstein metric. Implemented via a fast alternating projections algorithm, our method exhibits numerous strengths in simulations and analyses of single-cell gene expression data.

KEY WORDS: Wasserstein distance, batch effect, quantile regression, pool adjacent violators algorithm, single cell RNA-seq.

1. Introduction

A common type of data in scientific and survey settings consists of real-valued observations sampled in batches, where each batch shares a common label (this numerical/ordinal value is the *covariate*) whose effects on the observations are the item of interest. When each batch consists of a large number of i.i.d. observations, the empirical distribution of the batch may be a good approximation of the underlying population distribution conditioned on the value of the covariate. 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. In the case of high-dimensional observations, one can measure this effect separately for each variable to identify which are the most interesting. However, it may often occur that, in addition to random sampling variability, there exist unmeasured confounding variables (unrelated to the covariate) that affect the observations in a possibly dependent manner within the same batch (cf. *batch effects* in Risso et al. 2014).

The primary focus of this paper is the introduction of the TRENDS (Temporally Regulated Effects on Distribution Sequences) regression model, which infers the magnitude of these covariate-effects across entire distributions. TRENDS is an extension of classic regression with a single covariate (typically of fixed-design), where one realization of our dependent variable is a batch's entire empirical distribution (rather than a scalar) and the condition that fitted-values are smooth/linear in the covariate is replaced by the condition that fitted distributions follow a *trend*. Formally defined in §5, a trend describes a sequence of distributions where the pth quantile evolves monotonically for all $p \in (0, 1)$, though not necessarily in the same direction for different p, and there are at most two partitions of the quantiles that move in opposite directions. Thus, TRENDS extends scalar-valued regression to full distributions while retaining the ability to distinguish effects of interest from extraneous noise.

Despite the generality of our ideas, we motivate TRENDS with a concrete scientific

1

application: the analysis of single-cell RNA-sequencing time course data (see §S7 for a different application to income data; references preceded by 'S' are in the Supplementary Material).

The recent introduction of single-cell RNA-seq (scRNA-seq) techniques to obtain transcriptome-wide gene expression profiles from individual cells has drawn great interest (Geiler-Samerotte et al. 2013). 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 scRNA-seq to each sampled cell (Trapnell et al. 2014, Buettner et al. 2015). It is hoped that these data can reveal which *developmental* genes regulate/mark the emergence of new cell types over the course of development.

Current scRNA-seq cost/labor constraints prevent dense sampling of cells continuously across the entire time-continuum. Instead, researchers target a few time-points, simultaneously isolating sets of cells at each time and subsequently generating RNA-seq transcriptome profiles for each individual cell that has been sampled. More concretely, from a cell population undergoing some biological process like development, one samples $N_\ell \geqslant 1$ batches of cells from the population at time t_ℓ 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 n_i cells sampled and sequenced together. We denote by $x_{i,s}^{(g)} \in \mathbb{R}$ the measured expression of gene g in the sth cell of the ith batch ($1 \leqslant s \leqslant n_i$), sampled at time t_{ℓ_i}.

Because expression profiles are restricted to a sparse set of time points in current scRNA-seq experiments, the underlying rate of biological progression can drastically dif-

fer between equidistant times. Thus, changes in the expression of genes regulating differ-ent parts of this process may be highly nonuniform over time, invalidating assumptions like linearity or smoothness. One common solution in standard tissue-level RNA-seq time course analysis is time-warping (Bar-Joseph et al. 2003). Since our interest lies not in pre-dicting gene-expression at new time-points, we instead aim for a procedure that respects the sequence of times without being sensitive to their precise values. In fact, researchers commonly disregard the wall-clock time at which sequencing is done, instead recording the experimental chronology as a sequence of stages corresponding overall qualitative states of the biological sample. For example, in Deng et al. (2014): Stage 1 is the oocyte, Stage 2 the zygote, ..., Stage 11 the late blastocyst. Attempting to impose a common scale on the stage numbering is difficult because the similarity in expression expected across different pairs of adjacent stages might be highly diverse for different genes. In this work, we circumvent this issue by disregarding the time-scale and t_ℓ values, instead working only with the ordinal levels ℓ (so the only information retained about the times is their order $t_1 < t_2 < \cdots < t_L$), as done by Bijleveld et al. (1998) (Section 2.3.2).

Depictions of such data from two genes (where $N_\ell = 1$ for each ℓ) are shown in the lefthand panels of Figure 1. Lacking longitudinal measurements, these data differ from those studied in time series analysis: 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 a result of falling RNA-seq costs, multiple cell-capture plates (each producing a batch of sampled cells, i.e. $N_\ell > 1$) are being used at each time point to observe larger fractions of the cell population (Zeisel et al. 2015). Because the cells in a batch are simultaneously collected and sequenced (independently of other batches), the measured gene-expression values are often biased by *batch effects*: technical artifacts that perturb observed values in a possibly correlated fashion between cells of the same batch (Risso et al. 2014, Kharchenko et al. 2014). Rather than treating the cells from a single time point identically, it is desirable to retain batch information and
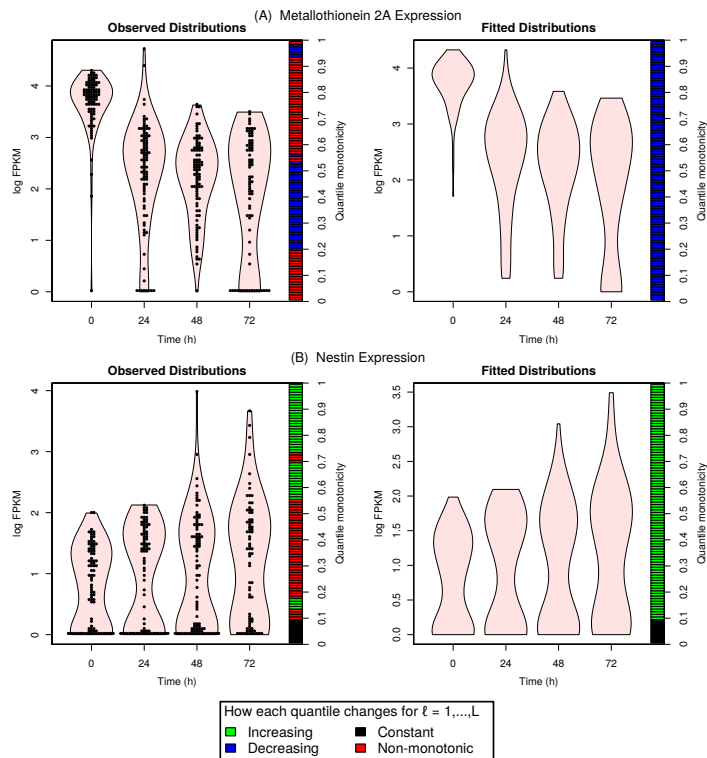
3

Figure 1: Violin plots (kernel density estimates) depicting the empirical distribution of known developmental genes' expression measured in myoblast cells (on left), and the corresponding TRENDS fitted distributions (on right). Each point shows a sampled cell.

account for this nuisance variation. Batch effects are also prevalent in other applications including temporal studies of demographic statistics, where a simultaneously-collected group of survey results may be biased by latent factors like location.

Furthermore, cell populations can exhibit enormous heterogeneity, particularly in developmental or in vivo settings (Trapnell et al. 2014, Buettner et al. 2015). A few high-expression cells often bias a population's average expression, and transcript levels can vary 1,000-fold between seemingly equivalent cells (Geiler-Samerotte et al. 2013)[1]. By fitting a TRENDS model (which accounts for both batch effects and the full distribution of expression across cells) to each gene's expression values, researchers can rank genes based

---

[1]Geiler-Samerotte et al. lament: "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".

on their presumed developmental relevance or employ hypothesis testing to determine whether observed temporal variation in expression is biologically relevant.

2. Related Work

To better motivate the ideas subsequently presented in this paper, we first describe why existing methods are not suited for scRNA-seq time course experiments and similar ordered-batched data lacking longitudinal measurements. As an alternative to time-series techniques, regression models might be applied in this setting, such as the Tobit generalized linear model of Trapnell et al. (2014). However, these models rely on linearity/smoothness assumptions, which can be inappropriate for sporadic processes such as development. More importantly, classic regression models scalar values such as conditional expectations, for which results must be interpreted as the effects in a hypothetical "average cell".

Rather than focusing only on (conditional) expectations or a few quantiles, it is often more appropriate to model the full (conditional) distribution of values in a heterogeneous population (Geiler-Samerotte et al. 2013, Buettner et al. 2015). Let P_ℓ denote the underlying distribution of the observations from covariate-level ℓ. An omnibus test for distribution-equality ($H_0 : P_1 = \cdots = P_L$ vs. the alternative that they are not all identical, cf. the Komogorov-Smirnov method described in §S3) can capture arbitrary changes, but fails to reflect sequential dynamics. Significance tests also do not quantify the size of effects, only the evidence for their existence. Krishnaswamy et al. (2014) have proposed a mutual-information based measure (DREMI) to quantify effects, which could be applied to our setting. However, under systematic noise caused by batch effects, measures of general statistical dependence between the batch-values and label ℓ (e.g. mutual information or hypothesis testing) become highly susceptible to the spurious variation present in the observed distributions (resulting in false positives). We thus prefer borrowing strength in the sense that a consistent change in distribution should ideally be observed across multiple time points for an effect to be deemed significant.

Instead of these general approaches, we model the P_ℓ as conditional distributions $\Pr(X \mid \ell)$ which follow some assumed structure as ℓ increases. Work in this vein has focused on modeling only a few particular quantiles of interest (Bondell et al. 2010) or accurate estimation of the conditional distributions using smooth nonparametric regression techniques (Fan et al. 1996, Hall et al. 1999). While such estimators possess nice theoretical properties and good predictive-power, the relationships they describe may be opaque and it is unclear how to quantify the covariate's effect on the entire distribution. Note that in the case of classic regression, interpretable linear methods remain favored for measuring effects throughout the sciences, despite the availability of flexible nonlinear function families. Our TRENDS framework retains this interpretability while modeling effects across full distributions.

Change-point analysis can also be applied to sequences of distributions, but is designed for detecting the precise locations of change-points over long intervals. However, scRNA-seq experiments only span a brief time-course (typically $L \leqslant 10$), and the primary analytic goal is rather to quantify how much a gene's expression has changed in a biologically interesting manner. Many change-point methods require explicit parameterization of the types of distributions, an undesirable necessity given the irregular nature of scRNA-seq expression measurements (Kharchenko et al. 2014). Moreover, some development-related genes exhibit gradual rather than abrupt temporal temporal changes in expression. Requiring few statistical assumptions, TRENDS models changes ordinally rather than only considering effects that are either smooth or instantaneous, and this method can therefore accurately quantify both abrupt or gradual effects.

3. Methods

Formally, TRENDS fits a regression model to an ordered sequence of distributions, or more broadly, sample pairs $\{(\ell_i, \widehat{P}_i)\}_{i=1}^N$ where each $\ell_i \in \{1, \dots, L\}$ is an ordinal-valued label associated with the ith batch, for which we have univariate empirical distribution

6

$\widehat{P}_i$. Here, it is supposed that for each batch i: a (empirical) quantile function $\widehat{F}_i^{-1}$ is estimated from n_i scalar observations $\{X_{i,s}\}_{s=1}^{n_i} \sim P_i$ sampled from underlying distribution $P_i = \Pr(X \mid \ell_i)$, which may be contaminated by different batch effects for each i. We assume a fixed-design where each level of the covariate $1, \ldots, L$ is associated with at least one batch. In scRNA-seq data, $\widehat{P}_i$ is the empirical distribution of one gene's measured expression values over the cells captured in the same batch and ℓ_i indicates the index of the time point at which the batch was sampled from the population for sequencing.

Unlike the supervised learning framework where one observes samples of X measured at different ℓ and the goal is to infer some property of $P_\ell := \Pr(X|\ell)$, in our setting, we can easily obtain $\widehat{P}_i$ as an empirical estimate of $\Pr(X|\ell_i)$. 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 $\Pr(X \mid \ell)$ across different ℓ may be attributed to changes in ℓ as opposed to the effects of other unmeasured confounding factors. To quantify this variation, we introduce conditional effect-distributions Q_ℓ for which the sequence of transformations $Q_1 \to Q_2 \to \cdots \to Q_L$ entirely captures the effects of ℓ-progression on $\Pr(X \mid \ell)$, under the assumption that these underlying forces follow a *trend* (defined in §5). We emphasize that the Q_ℓ themselves are not our primary inferential interest, rather it is the variation in these conditional-effect distributions that we attribute to increasing-ℓ rather than batch effects.

Thus, the Q_ℓ are *not* estimators of the sequence of P_{ℓ_i}. Rather, the Q_ℓ 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 ℓ-progression and we observe the entire population at each ℓ. Because we do not believe exogenous effects unrelated to ℓ-progression are likely to follow a trend over ℓ, we can identify the sequence of trending distributions which best models the variation in $\{\widehat{P}_{\ell_i}\}_{i=1}^N$ and reasonably conclude that changes in this sequence reflect the ℓ-progression-related forces affecting P_ℓ.

4. Wasserstein Distance

TRENDS employs the Wasserstein distance to measure divergence between distributions. Intuitively interpreted as the minimal amount of "work" that must be done to transform one distribution into the other, this metric has been successfully applied in many domains (Levina & Bickel 2001, Mueller & Jaakkola 2015). The Wasserstein distance is a natural dissimilarity measure of populations because it accounts for the proportion of individuals that are different as well as *how* different these individuals are. For univariate distributions, the L_q Wasserstein distance is simply the L_q distance between quantile functions given by:

$$d_{L_q}(P, Q) = \left(\int_0^1 \left| F^{-1}(p) - G^{-1}(p) \right|^q \, \mathrm{d}p \right)^{1/q} \tag{1}$$

where F, G are the CDFs of P, Q and F^{-1}, G^{-1} are the corresponding *quantile* functions. Slightly abusing notation, we use $d_{L_q}(\cdot, \cdot)$ to denote both Wasserstein distances between distributions or the corresponding quantile functions' L_q-distance (both $q = 1, 2$ are used in this work). In addition to being easy to compute (in 1-D), the L_2 Wasserstein metric is equipped with a natural space of quantile functions, in which the Fréchet mean takes the simple form stated in Lemma 1. Calling this average the *Wasserstein mean*, we note its implicit use in the popular quantile normalization technique (Bolstad et al. 2003).

Lemma 1. *Let $\mathcal{Q}$ denote the space of all quantile functions. The Wasserstein mean is the Fréchet mean in $\mathcal{Q}$ under the L_2 norm:*

$$\overline{\mathbf{F}}^{-1} := \frac{1}{N} \sum_{i=1}^N F_i^{-1} = \underset{G^{-1} \in \mathcal{Q}}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \int_0^1 \left(F_i^{-1}(p) - G^{-1}(p) \right)^2 \mathrm{d}p \right\} \tag{2}$$

5. Characterizing trends in distributions

Definition 1. Let $F_\ell^{-1}(p)$ denote the pth quantile of distribution P_ℓ with CDF F_ℓ. A sequence of distributions $P_1, \ldots, P_L$ follows a ***trend*** if:

1. For any $p \in (0, 1)$, the sequence $[F_1^{-1}(p), \ldots, F_L^{-1}(p)]$ is monotonic.

2. There exists $p^* \in [0, 1)$ and two intervals A, B that partition the unit-interval at p^* (one of A or B equals $(0, p^*)$ and the other equals $[p^*, 1)$) such that: for all $p \in A$, the sequences $[F_1^{-1}(p), \ldots, F_L^{-1}(p)]$ are all nonincreasing, and for all $q \in B$, the sequences $[F_1^{-1}(q), \ldots, F_L^{-1}(q)]$ are all nondecreasing. Note that if $p^* = 0$, then all quantiles must change in the same direction as ℓ grows.
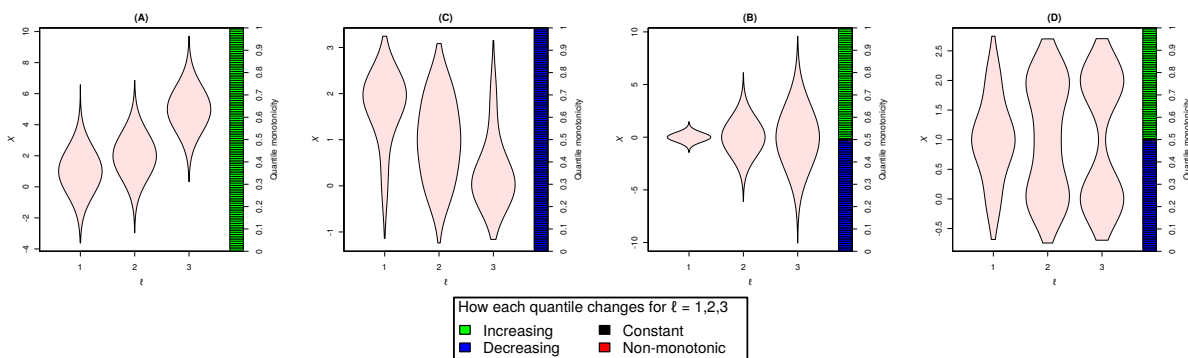

Figure 2: Violin plots depicting four different sequences of distributions which follow a trend. The pth rectangle in the color bar on the righthand side indicates the monotonicity of the pth quantile over the sequence of distributions (for $p = 0.01, 0.02, \ldots, 0.99$).

Our formal definition of a trend applies to distributions which evolve in a consistent fashion, ensuring that the temporal-forces that drive the transformation from P_1 to P_L do so without reversing their effects or leading to wildly different distributions at intermediate ℓ values. While the second condition of our definition technically subsumes the first, Condition 1 contains our key idea and is therefore separated from Condition 2, a subtler additional assumption that does not significantly alter results in practice. Note that the trend definition employed in this paper is intended for relatively short sequences and does not include cyclic/seasonal patterns studied in time-series modeling.

Lemma 2. *If distributions* $P_1, \ldots, P_L$ *follow a trend, then*

$$d_{L_1}(P_i, P_j) = \sum_{\ell=i+1}^{j} d_{L_1}(P_{\ell-1}, P_\ell) \quad for \ all \ \ i < j \in \{1, \ldots, L\}$$

9

Measuring how much the distributions are perturbed between each pair of levels via the L_1 Wasserstein metric, Lemma 2 shows the trend criterion as an instance of Occam's razor, where the underlying effects of interest are assumed to transform the distribution sequence in the simplest possible manner (recall that the Wasserstein distance is interpreted as the minimal work required for a given transformation). If one views the underlying effects of interest as a literal force acting in the space of distributions, Lemma 2 implies that this force points the same direction for every ℓ (i.e. $P_1, \ldots, P_L$ lie along a line in the L_1 Wasserstein metric space of distributions). A trend is more flexible than a linear restriction in the standard sense, because the magnitude of the force (how far along the line the distributions move) can vary over ℓ. Thus, we have formally extended the colloquial definition of a trend ("a general direction in which something is developing or changing") to probability distributions.

To further conceptualize the trend idea, one can view quantiles as different segments of a population whose values are distributed according to $\Pr(X \mid \ell)$ (e.g. for wealth-distributions, it has become popular to highlight the "one percent"). From this perspective, it is reasonable to assume that while the force of sequential progression may have different effects on the groups of individuals corresponding to different segments of the population, its effects on a single segment should be consistent over the sequence. If some segment's values initially change in one way at lower levels of ℓ and subsequently revert in the opposite direction over larger ℓ (i.e. this quantile is non-monotone), it is natural to conclude there are actually multiple different progression-related forces affecting this homogeneous group of individuals. It is therefore natural to assume a trend if we only wish to measure the effects of a single primary underlying force. Often in settings such as scRNA-seq developmental experiments, the researcher has a priori interest in a specific effect (such as how each gene contributes to a specific stage of the developmental process). Therefore, data are collected over a short ℓ-range such that the primary effects of interest should follow a trend.

The second condition in the trend definition specifies that adjacent quantiles must move in the same direction over ℓ except at most a single p^*. This restricts the number of population-segments which can increase over ℓ when a nearby segment of the population is decreasing. Intuitively, Condition 2 forces us to borrow strength across adjacent quantiles when estimating effects that follow a trend. The main effect of the additional restriction imposed by this condition prevents a trend from completely capturing extremely-segmented effects (such as the example depicted in Figure 3C). However, applications involving such complex phenomena are uncommon (it is difficult to imagine a setting where the primary effects-of-interest push more than two adjacent segments of a population in different directions), and such nuanced changes can be reasonably attributed to spurious nuisance variation. We note that a trend can still roughly approximate the major overall effects even when the actual distribution-evolution violates Condition 2 (as seen in Figure 3C). In practice, the results of our method are not significantly affected by this second restriction, but it provides nice theoretical properties ensuring our estimation procedure (presented in §8) efficiently finds a globally optimal solution, as well as additional robustness against spurious quantile-variation in the data (possibly due to estimation-error given limited samples per batch).

Figure 2 depicts simple examples of trending distribution-sequences. In each example, it is visually intuitive that the evolution of the distributions proceeds in a single consistent fashion. To highlight the broad spectrum of interesting effects TRENDS can detect, we present three conceptual examples in §S1 of distribution-sequences that follow a trend, which includes consistent changes in location/scale and the growth/disappearance of modes. Despite imposing conditions on every quantile, the trend criterion does not require: explicit parameterization of the distributions, specification of a precise functional form of the ℓ-effects, or reliance on a smooth or constant amount of change between different levels. This generality is desirable for modeling developmental gene expression and other enigmatic phenomena where stronger assumptions may be untenable.
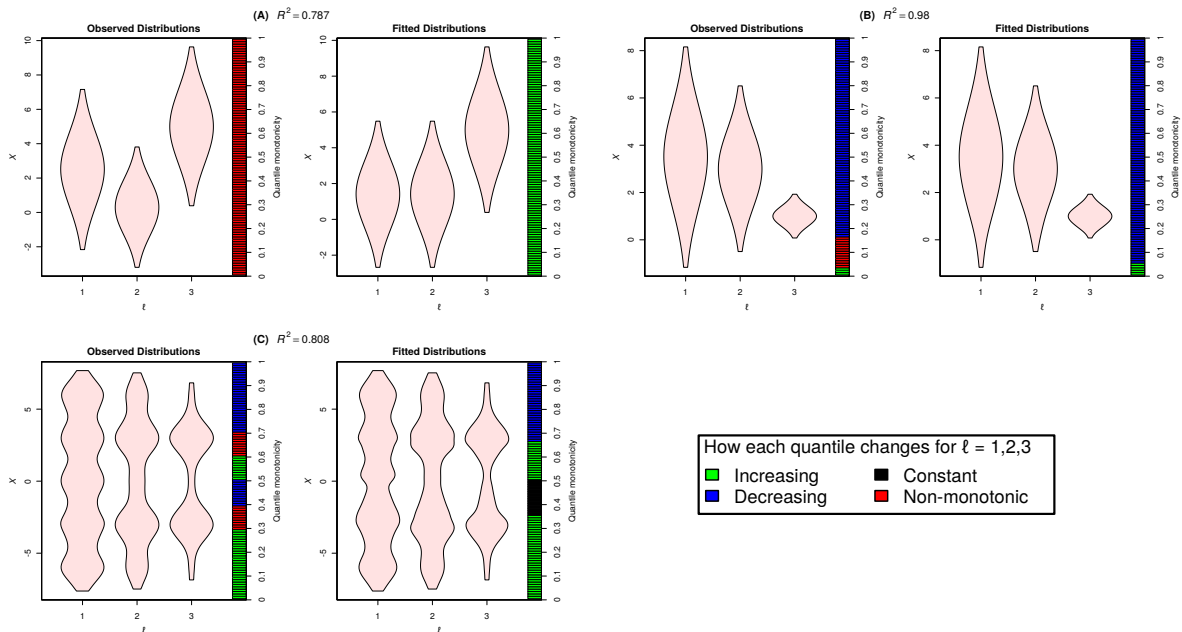
11

Figure 3: Violin plots depicting sequences of distributions which do *not* follow a trend (Observed Distributions in lefthand panels). Shown to the right of each example are the corresponding fitted distributions estimated by TRENDS (with the TRENDS R^2 value).

The lefthand panels of Figure 3 depict three examples of sequences which do not follow a trend for different reasons. To the right of each example, we show the "best-fitting" sequence that does follow a trend (formally defined in (4)), each distribution of which corresponds to our estimate of Q_ℓ (introduced in §3). We reiterate that the Q_ℓ are not by themselves of interest, but are merely used to quantify the sequential-progression effects (as will be described in §7). Nonetheless, the visual depiction of the trending Q_ℓ provides insight regarding what sort of changes a trend can accurately approximate. Whereas the evolution of the (trending) fitted distributions in Figure 3A (on right) can intuitively be attributed to one consistent force, multiple are required to explain the variation in the original non-trending sequence of distributions on the left. Identifying a single consistent effect responsible for the changes in the left panel of Figure 3B is far more plausible, and we note that these distributions in fact are much closer to following a trend (while hard to visually discern, the $0.04^{\text{th}} - 0.16^{\text{th}}$ quantiles of the observed distribution sequence

increase between $\ell = 1$ to 2 and decrease slightly from $\ell = 2$ to 3, thus violating a trend).

During specific stages of development, changes in the observed cellular gene-expression distributions generally stem from the emergence/disappearance of different cell subtypes (plus batch and random sampling effects). Clear subtype distinctions may not exist in early stages where cells remain undifferentiated, and thus not only are the relative proportions of different subtypes changing, but the subtypes themselves may transform as well. Therefore, developmental genes' underlying expression patterns are likely described by Examples 2 and 3 (of specific conceptual types of trends) in §S1. The trend criterion fits our a priori knowledge well, while remaining flexible with respect to the precise nature of expression changes.

6. TRENDS regression model

Recall that in our setting, even the underlying batch distributions P_i (from which the observations $X_{i,s}$ are sampled) may be contaminated by latent confounding effects. We assume the quantile functions of each P_i are generated from the model below:

$$F_i^{-1} = G_{\ell_i}^{-1} + \mathcal{E}_i \ \ \text{such that } G_1^{-1}, \ldots, G_L^{-1} \text{ follow a trend, and the following hold:} \tag{3}$$

(A.1) $\mathcal{E}_i : (0,1) \to \mathbb{R}$ is constrained so that $G_{\ell_i}^{-1}$ and F_i^{-1} are valid quantile functions.

(A.2) For all $p \in (0,1)$ and i: $\mathcal{E}_i(p)$ follows a sub-Gaussian(σ) distribution (Honorio & Jaakkola 2014), so $\mathbb{E}[\mathcal{E}_i(p)] = 0$ and $\Pr(|\mathcal{E}_i(p)| > t) \leqslant 2 \exp\left(-\frac{t^2}{2\sigma^2}\right)$ for any $t > 0$.

(A.3) For all $p \in (0,1)$ and $i \neq j$: $\mathcal{E}_i(p)$ is statistically independent of $\mathcal{E}_j(p)$.

In this model, G_ℓ^{-1} is the quantile function of the conditional effect-distribution Q_ℓ, whose evolution captures the underlying effects of level-progression. The random noise functions $\mathcal{E}_i : (0,1) \to \mathbb{R}$ can represent measurement-noise or the effects of other unobserved variables which contaminate a batch. Note that the form of $\mathcal{E}_i$ is implicitly constrained to ensure all $F_i^{-1}, G_{\ell_i}^{-1}$ are valid quantile functions. Because $\mathcal{E}_i(p_1)$ and $\mathcal{E}_i(p_2)$ are allowed to be dependent for $p_1 \neq p_2$, the effect of one $\mathcal{E}_i$ may manifest itself in multiple

observations $X_{i,s}$, even if these observations are drawn i.i.d. from P_i (for example, a batch effect can cause all of the observed values from a batch to be under-measured). In fact, condition (A.1) encourages significant dependence between the noise at different quantiles for the same batch. The assumption of sub-Gaussian noise is quite general, encompassing cases in which the $\mathcal{E}_i(p)$ are either: Gaussian, bounded, of strictly log-concave density, or any finite mixture of sub-Gaussian variables (Honorio & Jaakkola 2014). Although condition (A.3) stringently ensures all dependence between observations from different ℓ arises due to the trend, similar independence assumptions are required in general regression settings where one cannot reasonably a priori specify a functional form of dependence in the noise. Real batch effects are likely to satisfy (A.3) since they typically have the same chance of affecting any given batch in a certain manner (because the same experimental procedure is repeated across batches, as in the case of the cell-capture and library preparation in scRNA-seq). Nonetheless, we note that assumption (A.2) can be immediately generalized (with trivial changes to our proofs) in order to allow heteroscedasticity in the batch effects $\mathcal{E}_i$ (endowing each batch with a different σ_i sub-Gaussian parameter), but we opt for simplicity in this theoretical exposition.

Model (3) is a distribution-valued analog of the usual regression model, which assumes scalars $Y_i = f(X_i) + \epsilon_i$ where $\epsilon_i \sim$ sub-Gaussian(σ^2) and ϵ_i is independent of ϵ_j for $i \neq j$. In (3), an analogous f maps each ordinal level $\{1, \ldots, L\}$ to a quantile function, $f(\ell_i) = G_{\ell_i}^{-1}$, and the class of functions is restricted to those which follow a trend. Our assumption of mean-zero $\mathcal{E}_i$ that are independent between batches is a straightforward extension of the scalar error-model to the batch-setting, and ensures that the exogenous noise is unrelated to ℓ-progression under (3). Just as the $Y_1, \ldots, Y_N$ are rarely expected to exactly lie on the curve $f(x)$ in the classic scalar-response model, we do not presume that the observed distributions $\widehat{P}_i$ will exactly follow a trend (even as $n_i \to \infty$ $\forall i$ so that $\widehat{P}_i \to P_i$). Rather our model simply encodes the assumption that the effects of level-progression on the distributions should be consistent over different ℓ (i.e. the effects follow a trend).

For each ℓ, TRENDS finds a fitted distribution $\widehat{Q}_\ell$ using the *Wasserstein least-squares* fit which minimizes the following objective:

$$\widehat{Q}_1, \ldots, \widehat{Q}_L = \operatorname*{argmin}_{Q_1,\ldots,Q_L} \left\{ \sum_{\ell=1}^{L} \sum_{i \in I_\ell} d_{L_2}(Q_\ell, \widehat{P}_i)^2 \right\} \quad \text{where } Q_1, \ldots, Q_L \text{ follow a trend} \qquad (4)$$

where I_ℓ is the set of batch-indices i such that $\ell_i = \ell$, and we require $N_\ell := |I_\ell| \geqslant 1$ for all $\ell \in \{1, \ldots, L\}$. Subsequently, one can inspect changes in the $\widehat{Q}_\ell$ which should reflect the transformations in the underlying P_ℓ that are likely caused by increasing ℓ. Figure 3 shows some examples of fitted distributions produced by TRENDS regression. The objective in (4) bears great similarity to the usual least-squares loss used in scalar regression, the only differences being: scalars have been replaced by distributions, squared Euclidean distances are now squared Wasserstein distances, and the class of regression functions is defined by a trend rather than linearity/smoothness criteria.

Expression measurements in scRNA-seq are distorted by significant batch effects, so the $\mathcal{E}_i$ are likely to be large. In addition to technical artifacts, Buettner et al. (2015) find biological sources of noise due to processes such as transcriptional bursting and cell-cycle modulation of expression. Unlike development-driven changes in the underlying expression of a developmental gene, other biological/technical sources of variation are unlikely to follow any sort of trend. TRENDS thus provides a tool for modeling full distributions, while remaining robust to the undesirable variation rampant in these applications by leveraging independence of the noise between different batches of simultaneously captured and sequenced cells.

7. Measuring fit, effect size, and statistical significance

Analogous to the coefficient of determination used in classic regression, we define the Wasserstein R^2 to measure how much of the variation in the observed distributions

$\widehat{P}_1, \ldots, \widehat{P}_N$ is captured by the TRENDS model's fitted distributions $\widehat{Q}_1, \ldots, \widehat{Q}_L$:

$$R^2 := 1 - \left(\frac{1}{N} \sum_{i=1}^{N} d_{L_2}(\widehat{Q}_{\ell_i}, \widehat{P}_i)^2 \right) \Big/ \left(\frac{1}{N} \sum_{i=1}^{N} d_{L_2}(\widehat{P}_i, \overline{\mathbf{F}}^{-1})^2 \right) \quad \in [0, 1] \tag{5}$$

Here, squared distances between scalars in the classic R^2 are replaced by squared Wasserstein distances between distributions, and the quantile function $\overline{\mathbf{F}}^{-1} = \frac{1}{N} \sum_{i=1}^{N} \widehat{F}_i^{-1}$ is the *Wasserstein mean* of all observed distributions. By Lemma 1, the numerator and denominator in (5) are respectively analogous to the residuals and the overall variance from usual scalar regression models.

In classic linear regression, the regression line slope is interpreted as the expected change in the response resulting from a one-unit increase in the covariate. While TRENDS operates on unit-less covariates, we can instead measure the overall *expected Wasserstein-change* under model (3) in the $\widehat{P}_i$ over the full ordinal progression $\ell = 1, \ldots, L$ using:

$$\Delta := \frac{1}{L} \cdot d_{L_1}(\widehat{Q}_1, \widehat{Q}_L) \tag{6}$$

The L_1 Wasserstein distance is a natural choice, since by Lemma 2, it measures the aggregate difference over each pair of adjacent ℓ levels (just as the difference between the largest and smallest fitted-values in linear regression may be decomposed in terms of covariate units to obtain the regression-line slope). Thus, Δ measures the raw magnitude of the inferred trend-effect (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). Note that if the TRENDS model is fit to the distributions from the example in Figure 3B, the TRENDS-inferred effect of sequential-progression is nearly as large as the overall variation in this sequence, which agrees with our visual intuition that the observed distributions already evolve in a fairly consistent fashion.

Finally, we introduce a test to assess statistical significance of the trend-effect. We compare the null hypothesis $H_0 : Q_1 = Q_2 = \cdots = Q_L$ 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 ℓ_i-labels of our observed distributions $\widehat{P}_i$ with test-statistic R^2 (Good 1994). More specifically, the null distribution is determined by repeatedly executing the following steps: (i) randomly shuffle the ℓ_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_{perm} using (5). Due to the quantile-noise functions $\mathcal{E}_i(\cdot)$ assumed in our model (3), H_0 allows variation in our sampling distributions P_i which stems from non-ℓ-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 one distribution vs. a sequence of trending distributions in Wasserstein-space, we note that our R^2 resembles a likelihood-ratio-like test statistic between maximum-likelihood-like estimates $\overline{\mathbf{F}}^{-1}$ and $\widehat{Q}_\ell$ (where we operate under the Wasserstein distance rather than Kullback-Leibler which underlies the maximum likelihood framework).

As we do not parametrically treat the distributions, we find permutation testing more suitable than relying on asymptotic approximations. Unfortunately, N and L may be small, undesirably limiting the number of possible label-permutations. In §S2, we overcome the granularity problem that arises in such settings by developing a more intricate permutation procedure akin to the smoothed bootstrap of Silverman & Young (1987).

To determine whether our model is reasonable when working with real data, it is best to rely on prior domain knowledge regarding whether or not the effects of primary interest should follow a trend. When this fact remains uncertain, then (as in the case of classical regression) the question is not properly answered using just our Wasserstein R^2 values (which we caution tend to be much larger than the familiar R^2 values from linear regression, due to the heightened flexibility of our TRENDS model). §S6 demonstrates a simple method for model checking based on plotting empirically-estimated residual functions $\widehat{\mathcal{E}}_i$ against the sequence-level ℓ. Similar plots of scalar residuals are the most

common diagnostic employed in standard regression analysis. While this model-checking procedure is able to clearly delineate simulated deviations from our assumptions, it shows little indication that the TRENDS assumptions are inappropriate for the real scRNA-seq data from major known developmentally-relevant genes. Our simulation in §S6 also empirically demonstrates that despite its restrictive assumptions, the TRENDS model can provide superior estimates of severely-misspecified effects than the initial empirical distributions.

8. Fitting the TRENDS model

We propose the trend-fitting (TF) algorithm which finds distributions satisfying

$$\widehat{Q}_1, \ldots, \widehat{Q}_L = \arg\min_{Q_1, \ldots, Q_L} \left\{ \sum_{\ell=1}^{L} \sum_{i \in I_\ell} w_i \cdot d_{L_2}(Q_\ell, \widehat{P}_i)^2 \right\} \text{ where } Q_1, \ldots, Q_L \text{ follow a trend} \quad (7)$$

If $\widehat{P}_i$ (the empirical per-batch distributions) are estimated from widely varying sample sizes n_i for different batches i, then it is preferable to replace the objective in (4) with the weighted sum in (7). Given weights w_i chosen based on n_i and N_ℓ, TRENDS can better model the variation in the empirical distributions that are likely more accurate due to larger sample size. As n_i and N_ℓ are fairly homogeneous in scRNA-seq experiments, we use uniform weights here (but provide an algorithm for the general formulation). To fit TRENDS to data $\{(\ell_i, \widehat{P}_{\ell_i}, w_i)\}_{i=1}^N$ via our procedure, the user must first specify:

- Numerical quadrature points $0 < p_1 < p_2 < \cdots < p_{P-1} < 1$ for evaluating the Wasserstein distance integral in (1), i.e. which $P - 1$ quantiles to use for each batch

- a quantile estimator $\widehat{F}^{-1}(p)$ for empirical CDF $\widehat{F}$

Given these two specifications, the TF procedure solves a numerical-approximation of the constrained distribution-valued optimization problem in (7). Defining $p_0 := 2p_1 - p_2$ and

18

$p_P := 2p_{P-1} - p_{P-2}$, we employ the following midpoint-approximation of the integral

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

where $G_1, \ldots, G_L$ must follow a trend   (8)

While this problem is unspecified between the p_kth and p_{k+1}th quantiles, all we require to numerically compute Wasserstein distances (and hence R^2 or Δ) is the values of the quantile functions at $p_1, \ldots, p_{P-1}$, which are uniquely determined by (8). Although our algorithm operates on a discrete set of quantiles like techniques for quantile regression (Bondell et al. 2010), this is only for practical numerical reasons; the goal of our TRENDS framework is to measure effects across an entire distribution. Throughout this work, we use $P - 1$ uniformly spaced quantiles between $\frac{1}{P}$ and $\frac{P-1}{P}$ (with $P = 100$) to comprehensively capture the full distributions while ensuring computational efficiency. In settings with limited data per batch, one might alternatively select fewer quadrature points (quantiles), avoiding tail regions of the distributions for increased stability (our results were robust to the precise number of quadrature points employed).

Since no unbiased minimum-variance $\forall p \in (0, 1)$ quantile estimator is known, we simply use the default setting in R's quantile function, which provides the best approximation of the mode (Type 7 of Hyndman & Fan (1996)). Other quantile estimators perform similarly in our experiments, and Keen (2010) have found little practical difference between estimation procedures for sample sizes ≥ 30. Here, we assume the n_i cells sampled in the ith batch are i.i.d. samples (reasonable for cell-capture techniques). If this assumption is untenable in another domain, then the quantile-estimation should be accordingly adjusted (cf. Heidelberger & Lewis 1984).

Our procedure uses the Pool-Adjacent-Violators-Algorithm (PAVA), which given an input sequence $y_1, \ldots, y_L \in \mathbb{R}$, finds the least-squares-fitting nondecreasing sequence in only $O(L)$ runtime (de Leeuw 1977). The basic PAVA procedure is extended to weighted

---

Basic PAVA Algorithm: $\min_{z_\ell} \sum_{\ell=1}^{L} (y_\ell - z_\ell)^2$ s.t. $z_1 \leqslant \cdots \leqslant z_L$

Input: A sequence of real numbers $y_1, \ldots, y_L$
Output: The minimizing sequence $\widehat{y}_1, \ldots, \widehat{y}_L$ which is nondecreasing.

1. Start with the first level $\ell = 1$ and set the fitted value $\widehat{y}_1 = y_1$
2. While the next $y_\ell \geqslant \widehat{y}_{\ell-1}$, set $\widehat{y}_\ell = y_\ell$ and increment ℓ
3. If the next ℓ violates the nondecreasing condition, i.e. $y_\ell < \widehat{y}_{\ell-1}$, 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 $\hat{y}_1, \ldots, \hat{y}_\ell$. Repeat Steps 2 and 3 until $\ell = L$.

---

observations by performing weighted backaveraging in Step 3. When multiple (ℓ_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 predictor-ties (de Leeuw 1977). Here, one defines $\bar{y}_\ell$ as the (weighted) average of the $\{y_i : i \in I_\ell\}$ and for each level ℓ all $y_i : i \in I_\ell$ are simply replaced with their mean-value $\bar{y}_\ell$. Subsequently, PAVA is applied with non-uniform weights to $\{(\ell, \bar{y}_\ell)\}_{\ell=1}^{L}$ where the ℓth point receives weight N_ℓ (or weight $\sum_{i \in I_\ell} w_i$ if the original points are assigned non-uniform weights $w_1, \ldots, w_N$). By substituting "nonincreasing" in place of "nondecreasing" in Steps 2 and 3, the basic PAVA method can be trivially modified to find the least-squares *nonincreasing* sequence. From here on, we use $\text{PAVA}((y_1, w_1), \ldots, (y_N, w_N); \delta)$ to refer to a more general version of basic PAVA, which incorporates observation-weights w_i (for multiple y values at a single ℓ), and a user-specified monotonicity condition $\delta \in \{\text{"nonincreasing", "nondecreasing"}\}$ that determines which monotonic best-fitting sequence to find.

Theorem 1. *The Trend-Fitting algorithm produces valid quantile-functions $\widehat{G}_1^{-1}, \ldots, \widehat{G}_L^{-1}$ which solve the numerical version of the TRENDS objective given in (8).*

Fundamentally, our TF algorithm utilizes Dykstra's method of alternating projections (Boyle & Dykstra 1986) to project between the set of L-length sequences of vectors which are monotone in each index over ℓ and the set 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 converges extremely quickly in practice. This

20

Trend-Fitting Algorithm: Numerically solves (7) by optimizing (8)

Input 1:   Empirical distributions and associated levels (and optional weights) $\{(\ell_i, \widehat{F}_i, w_i)\}_{i=1}^N$
Input 2:   A grid of quantiles to work with $0 < p_1 < \cdots < p_{P-1} < 1$
Output: The estimated quantiles of each Q_ℓ $\{\widehat{G}_\ell^{-1}(p_k) : k = 1, \ldots, P-1\}$ for $\ell \in \{1, \ldots, L\}$
   from which these underlying trending distributions can be reconstructed.

1. $\widehat{F}_i^{-1}(p_k) := \mathbf{quantile}(\widehat{F}_i, p_k)$ for each $i \in \{1, \ldots, N\}, k \in \{1, \ldots, P-1\}$

2. $w_\ell^* := \sum_{i \in I_\ell} w_i$ for each $\ell \in \{1, \ldots, L\}$

3. $x_\ell[k] := \dfrac{1}{w_\ell^*} \sum_{i \in I_\ell} w_i \widehat{F}_i^{-1}(p_k)$ 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. $\delta[k] :=$ "nondecreasing" if $p_k > p^*$; otherwise $\delta[k] :=$ "nonincreasing"

6. $y_1, \ldots, y_L := \mathbf{AlternatingProjections}\left(x_1, \ldots, x_L \; ; \; \delta \; ; \; \{w^*\}_{\ell=1}^L, \{p_k\}_{k=1}^{P-1} \right)$

7. $W[\delta] :=$ the value of (8) evaluated with $G_\ell^{-1}(p_k) = y_\ell[k]$ $\forall \ell, k$

8.   Redefine $\delta[k] :=$ "nonincreasing" if $p_k > p^*$; otherwise $\delta[k] :=$ "nondecreasing"
       and repeat Steps 6 and 7 with the new δ

9. Identify $\min\limits_{\delta} W[\delta]$ and return $\widehat{G}_\ell^{-1}(p_k) = y_\ell^*[k]$ $\forall \ell, k$ where y^* was produced at the
    Step 6 or 8 corresponding to $\delta^* := \arg\max W[\delta]$.

AlternatingProjections Algorithm: Finds the Wasserstein-least-squares sequence of vectors which represent valid quantile-functions and a trend whose monotonicity is specified by δ.

Input 1:   Initial sequence of vectors $x_1^{(0)}, \ldots, x_L^{(0)}$
Input 2:   Vector δ whose indices specify directions constraining the quantile-changes over ℓ.
Input 3:   Weights $w_\ell^* \in \mathbb{R}$ 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 $\forall \ell, k:$ $y_\ell^{(t)}[k] \leqslant y_\ell^{(t)}[k+1]$ and the sequence
   $y_1^{(t)}[k], \ldots, y_L^{(t)}[k]$ is monotone nonincreasing/nondecreasing as specified by $\delta[k]$,
      provided that $x_\ell^{(0)}[k] \leqslant x_\ell^{(0)}[k+1]$ for each ℓ, 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. $y_1^{(t)}[k], \ldots, y_L^{(t)}[k] := \mathbf{PAVA}\left(\left(x_1^{(t)}[k] + r_1^{(t)}[k], w_1^* \right), \ldots, \left(x_L^{(t)}[k] + r_L^{(t)}[k], w_L^* \right) ; \delta[k] \right)$
       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 ℓ, k

5. $\forall \ell \in \{1, \ldots, L\} :$ $x_\ell^{(t+1)}[1], \ldots, x_\ell^{(t+1)}[P-1] :=$
    $\mathbf{PAVA}\left(\left(y_\ell^{(t)}[1] + s_\ell^{(t)}[1], \frac{p_2 - p_0}{2} \right), \ldots, \left(y_\ell^{(t)}[P-1] + s_\ell^{(t)}[P-1], \frac{p_P - p_{P-2}}{2} \right) ; \text{"nondecreasing"} \right)$

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

---

procedure has overall computational complexity $O(TLP^2 + NP)$, which is efficient when T (the total number of projections performed) is small, since both P and L are limited. The proof of Theorem 1 provides much intuition on the TF algorithm (all proofs are relegated to §S8). Essentially, once we fix a δ configuration (specifying which quantiles are decreasing over ℓ and which are increasing), our feasible set becomes the intersection of two convex sets between which projection is easy via PAVA. Furthermore, the second statement in our trend definition limits the number of possible δ configurations, so we simply solve one convex subproblem for each possible δ to find the global solution.

9. Theoretical results

Under the model given in (3), we establish some results regarding the quality of the $\hat{Q}_1, \ldots, \hat{Q}_L$ estimates produced by the TF algorithm. To develop pragmatic theory, we use finite-sample bounds defined in terms of quantities encountered in practice rather than the true Wasserstein distance (1), which relies on an integral that must be numerically approximated. Thus, in this section, $d_W(\cdot, \cdot)$ is used to refer to the midpoint-approximation of the L_2 Wasserstein integral illustrated in (8). In addition to the conditions of model (3), we make the following simplifications throughout for ease of exposition:

(A.4) The number of batches at each level is the same, i.e. $N_\ell := N_1 = \cdots = N_L \geqslant 1$

(A.5) The same number of samples are drawn per batch, i.e. $n := n_i$ for all $1 \leqslant i \leqslant N$

(A.6) For $k = 1, \ldots, P - 1$: the (k/P)th quantiles of each distribution are considered

(A.7) Uniform weights are employed, i.e. in (7): $w_i = 1$ for all i

Theorem 2. *Under model (3) and additional conditions (A.4)-(A.7), suppose the TF algorithm is applied directly to the true quantiles of $P_1, \ldots, P_N$. Then, given any $\epsilon > 0$, the resulting estimates satisfy:* $\quad d_W(\hat{G}_\ell^{-1}, G_\ell^{-1}) < \epsilon \quad$ *for each $\ell \in \{1, \ldots, L\}$*

with probability greater than: $\quad 1 - 2PL \exp\left(-\dfrac{\epsilon^2 N_\ell}{8\sigma^2 L}\right)$ $\qquad\qquad$ (9)

Thus, Theorem 2 implies that our estimators are consistent with asymptotic rate $O_P(1/\sqrt{N_\ell})$ if we directly observe the true per-batch quantiles $P_1, \ldots, P_N$ (which are contaminated by $\mathcal{E}_i$ under our model). By using the union-bound, our proof does not require any independence assumptions for the noise introduced at different quantiles of the same batch. Because direct quantile-observation is unlikely in practice, we now examine the performance of TRENDS when these quantiles are instead estimated using n samples from each P_i. Here, we additionally assume:

(A.8) For $i = 1, \ldots, N$: quantiles are estimated from n i.i.d. samples $X_{1,i}, \ldots, X_{n,i} \sim P_i$

(A.9) There is nonzero density at each of the quantiles we estimate, i.e. CDF F_i is strictly increasing around each $F_i^{-1}(k/P)$ for $k = 1, \ldots, P - 1$.

(A.10) The simple quantile estimator defined below is used for each $k/P, k = 1, \ldots, P - 1$

$$\widehat{F}_i^{-1}(p) := \inf\{x : \widehat{F}_i(x) \geq p\}$$

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

Theorem 3. *Under the assumptions of Theorem 2 and (A.8)-(A.10), suppose the TF algorithm is applied to estimated quantiles $\widehat{F}_i^{-1}(k/P)$ for $i = 1, \ldots, N, k = 1, \ldots, P - 1$. Then, given any $\epsilon > 0$, the resulting estimates satisfy:* $d_W(\widehat{G}_\ell^{-1}, G_\ell^{-1}) < \epsilon$ *for each $\ell \in \{1, \ldots, L\}$ with probability greater than:*

$$1 - 2PL\left[\exp\left(\frac{-\epsilon^2 N_\ell}{32\sigma^2 L}\right) + N_\ell \exp\left(-2n \cdot R\left(\frac{\epsilon}{4\sqrt{L}}\right)^2\right)\right] \tag{10}$$

where for $\gamma > 0$:

$$R(\gamma) := \min_{i,k}\{R(\gamma, i, k/P) : i = 1, \ldots, N, k = 1, \ldots, P - 1\}$$

$$R(\gamma, i, p) := \min\left\{F_i\left(F_i^{-1}(p) + \gamma\right) - p , \; p - F_i\left(F_i^{-1}(p) - \gamma\right)\right\} \tag{11}$$

Theorem 3 is our most general result applying to arbitrary distributions P_i that satisfy basic condition (A.9). However, the resulting probability-bound may not converge toward

to 1 if $n \cdot R(\frac{\epsilon}{4\sqrt{L}})^2 < O(\log N_\ell)$, which occurs if few samples are available per batch (because then the P_i are can be very poorly estimated). Thus, TRENDS is in general only designed for applications with large per-batch sample sizes. The bounds obtained under the extremely broad setting of Theorem 3 may be significantly improved by instead adopting one of the following stronger assumptions:

(A.11) The simple quantile-estimator defined in (A.10) is used, and the support of each P_i is bounded and connected with non-neglible density, i.e. $\exists$ constants $B, c > 0$ s.t. $\forall i$:

$$f_i(x) = 0 \; \forall x \notin [-B, B] \quad \text{and} \quad f_i(x) \geqslant c \; \forall x \in [-B, B] \quad (f_i \text{ is density for CDF } F_i).$$

(A.12) The following is known regarding the quantile-estimation procedure:

1. The quantiles of each P_i are estimated independently of the others.

2. The quantile-estimates converge at a sub-Gaussian rate for each quantile of interest, i.e. there exists $c > 0$ such that for each k, i and any $\epsilon > 0$:

$$\Pr\left(\left|\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P)\right| > \epsilon\right) \leqslant 2\exp(-2nc^2\epsilon^2)$$

Theorem 4. *Under the assumptions of Theorem 2, conditions (A.8), (A.9), and one of either (A.11) or (A.12), the bound in (10) may be sharpened to ensure that for any $\epsilon > 0$:*

$$d_W(\widehat{G}_\ell^{-1}, G_\ell^{-1}) < \epsilon \quad \text{for each } \ell \in \{1, \dots, L\}$$

with probability greater than:

$$1 - 2P\left[L\exp\left(\frac{-\epsilon^2 N_\ell}{32\sigma^2 L}\right) + \exp\left(-\frac{c^2}{8}N_\ell \, n\epsilon^2\right)\right] \tag{12}$$

In Theorem 4, the additional assumption of bounded/connected underlying distributions results in a much better finite sample bound that is exponential in both n and N_ℓ (implying asymptotic $O_P(N_\ell^{-1/2} + n^{-1/2})$ convergence). While this condition and the result of Theorem 3 assume use of the simple quantile-estimator from (A.10), numerous superior procedures have been developed which can likely improve practical convergence

rates (Zielinski 2006). Assuming guaranteed bounds for the quantile-estimation error (which may be based on both underlying properties of the P_i as well as the estimation procedure), one can also obtain the same exponential bound. In fact, condition (A.11) is an example of a distribution and quantile-estimator combination which achieves the error required by (A.12). Because the boundedness assumption is undesirably limiting, we also derive a similar result under weaker assumptions:

(A.13) Each P_i has connected support with non-neglible interior density and sub-Gaussian tails, i.e. there are constants $B > b > 0, a > 0, c > 0$ such that for all i :

    (1) F_i is strictly increasing,

    (2) $f_i(x) \geqslant c \ \forall x \in [-B, B]$ where f_i is the density function of CDF F_i.

    (3) $\Pr(X_i > x) \leqslant \exp\left(-a\left[x - (B-b)\right]^2\right)$ if $x > B$

         and $\Pr(X_i < x) \leqslant \exp\left(-a\left[x - (-B+b)\right]^2\right)$ if $x < -B$

(A.14) Defining $r := \min\left\{2c^2 , \frac{2ab^2 - 1}{4PB^2}\right\}$, we have $r > 0$, or equivalently, $2ab^2 > 1$.

(A.15) We avoid estimating extreme quantiles, i.e. $F_i^{-1}(k/P) \in (-B, B) \ \forall k \in \{1, \ldots, P-1\}$

Theorem 5. *Under the assumptions of Theorems 2 and 3 as well as conditions (A.13)- (A.15), the previous bound in (10) may be sharpened to ensure that for all $\epsilon > 0$:*

$$d_W(\widehat{G}_\ell^{-1}, G_\ell^{-1}) < \epsilon \ \ \text{for each } \ell \in \{1, \ldots, L\}$$

with probability greater than:

$$1 - 2P\left[L\exp\left(\frac{-\epsilon^2 N_\ell}{32\sigma^2 L}\right) + \exp\left(-\frac{r}{16}N_\ell n\epsilon^2\right)\right] \tag{13}$$

Theorem 5 again provides an exponential bound in both n and N_ℓ under a realistic setting where the distributions are small tailed with connected support, and the simple quantile estimator of (A.10) is applied at non-extreme quantiles. Note that while we specified properties of the distributions, noise, and quantile estimation in order to develop this theory, our nonparametric significance tests do not rely on these assumptions.

10. Simulation study

We perform a simulation which realistically reflects various properties of scRNA-seq data, based on assumptions similar to those explicitly relied upon by the model of Kharchenko et al. (2014). Samples are generated from one of the following choices of the underlying trending distribution sequence $Q_1, \ldots, Q_L$ with $L = 5$ (additional details in §S4):

(S_1) $Q_\ell \sim \mathrm{NB}(r_\ell, p_\ell)$ with $r_\ell = 5$ and $p_\ell = 0.3, 0.3, 0.4, 0.5, 0.8$ for $\ell = 1, \ldots, 5$.

(S_2) Q_ℓ is a mixture of $\mathrm{NB}(r = 5, p = 0.3)$ and $\mathrm{NB}(r = 5, p = 0.7)$ components, with the mixing proportion of the latter ranging over $\lambda_\ell = 0.1, 0.4, 0.8, 0.8, 0.8$ for $\ell = 1, \ldots, 5$.

(S_3) $Q_\ell \sim \mathrm{NB}(r = 5, p = 0.5)$ for $\ell = 1, \ldots, 5$.

$\mathrm{NB}(r, p)$ denotes the negative binomial distribution parameterized by r (target number of successful trials) and p (probability of success in each trial). To capture various types of noise affecting scRNA-seq measurements (e.g. dropout, PCR amplification bias, transcriptional bursting), noise for the ith batch is introduced (independently of the other batches) via the following steps: rather than sampling from Q_{ℓ_i}, we instead sample from $P_{\ell_i} \sim \mathrm{NB}(\widetilde{r}_\ell, \widetilde{p}_\ell)$, where $\widetilde{r}_\ell = r_\ell + r_{\mathrm{noise}}$ and $\widetilde{p}_\ell = p_\ell + p_{\mathrm{noise}}$. Here, $p_{\mathrm{noise}}, r_{\mathrm{noise}}$ are independently drawn from centered Gaussian distributions with standard deviations σ, $10 \cdot \sigma$ respectively (σ thus controls the degree of noise). For the mixture-models in S_2, we sample from P_{ℓ_i} which is also a mixture of negative binomials (with the same mixing proportions as Q_{ℓ_i}) where the parameters of both mixing components are perturbed by noise variables r_{noise}, p_{noise}. To the observations sampled from P_{ℓ_i}, we finally apply a $\log_{10}(x+1)$ transform (also applied to the scRNA-seq data in §11) before proceeding with our analysis.

We first investigate the convergence of TRENDS estimates under each of the models S_1, S_2, and S_3, varying n, N_ℓ, and the amount of noise independently. Figure 4 shows the Wasserstein error (sum over ℓ of the squared Wasserstein distances between the underlying Q_ℓ and estimates thereof) of our TRENDS estimates vs. the error of the empirical

distributions. The plot demonstrates rapid convergence of the TRENDS estimator (as guaranteed by our theory in §9) and shows that TRENDS can produce a much better picture of the underlying distributions than the (noisy) observed empirical distributions. As shown in Figure 4A, this may occur even in the absence of noise, thanks to the additional structure of the trend-assumption exploited by our estimator. Thus, when the underlying effects follow a trend, our Δ statistic provides a much more accurate measure of their magnitude than distances between the empirical distributions. These results indicate that the largest benefit of our TRENDS approach is for small to moderate sized samples.
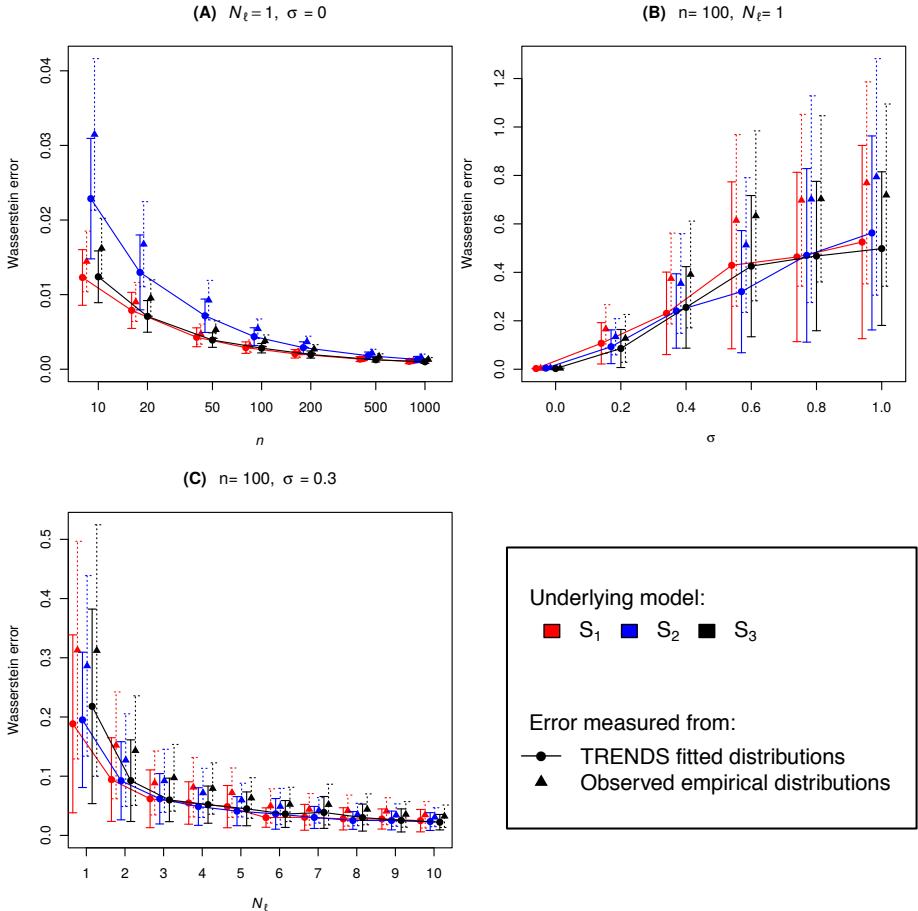

Figure 4: The Wasserstein error of the TRENDS fitted distributions vs. the observed empirical distributions, under models S_1 - S_3 with various settings of n, σ, and N_ℓ. Depicted is the average error (and standard deviation) over 100 repetitions.

To compare performance, we evaluate TRENDS against alternative methods under our models S_1-S_3 with substantial batch-noise ($\sigma = 0.1$). Fixing $N_\ell = 1, n_i = 1000$ for all ℓ, i, we generate 400 datasets from the different underlying trending models described above (100 from each of S_1, S_2, and 200 from S_3). TRENDS is applied to each dataset to obtain a p-value (via the permutation procedure described in §S2). In this analysis, we also apply the following alternative methods (detailed in §S3): a linear variant of our TRENDS model (where quantiles are restricted to evolve linearly rather than monotonically), an omnibus-testing approach (using the maximal Kolmogorov-Smirnov (KS) statistic between any pair of distributions), and a measure of the (marginally-normalized) mutual information (MI) between ℓ and the values in each batch. The latter two alternative methods make no underlying assumption and capture arbitrary variation in distributions over ℓ. We employ the same approach to ascertain statistical significance (at the 0.05 level) under each method. All p-values are obtained via permutation-testing (with 1000 permutations). To correct these p-values for multiple comparisons, we employ the step-down minP adjustment algorithm of Ge et al. (2003), which cleverly avoids double permutations to remain computationally efficient.

Method	FPR	TPR	AUROC
TRENDS	0.02	0.35	0.87
Linear-TRENDS	0.03	0.32	0.85
KS	1.0	1.0	0.44
MI	1.0	1.0	0.53

Table 1: False-positive rate (FPR) and true-positive rate (TPR) produced by different methods, as well as AUROC values. FPR is determined by the fraction of datasets generated under model S_3 deemed statistically significant (or S_1, S_2 for TPR).

Table 1 demonstrates that methods sensitive to arbitrary differences in distributions are highly susceptible to spurious batch effects (both the KS and MI identify all 400 datasets as statistically significant), whereas our TRENDS method has the lowest false-positive rate, only incorrectly rejecting its null hypothesis for 4 out of the 200 datasets from S_3. TRENDS also exhibits the greatest power in these experiments. To ascertain how well

these methods distinguish the trending data from the non-trending samples, we computed area under the ROC curve (AUROC) by generating ROC curves for each method using its p-values (ties broken using test statistics) as a classification-rule for determining which simulated datasets the method would correctly distinguish from constant model S_3 at each possible cutoff value. The results of Table 1 show that TRENDS is superior at drawing this distinction in these simulations.

11. Single cell RNA-seq analysis

To evaluate the practical utility of our method, we analyze two scRNA-seq time course experiments and compare TRENDS against the alternative approaches described in §S3. The first dataset is from Trapnell et al. (2014) who profiled single-cell transcriptome dynamics of skeletal myoblast cells at 4 time-points during differentiation (myoblasts are embryonic progenitor cells which undergo myogenesis to become muscle cells). In a second larger-scale scRNA-seq experiment, Zeisel et al. (2015) isolated 1,691 cells from the somatosensory cortex (the brain's sensory system) of juvenile CD1 mice aged P22-P32. We treat age (in postnatal days) as our batch-labels, with $L = 10$ possible levels. §S5 contains detailed descriptions of the data and our analysis.

Assuming that trending temporal-progression effects on expression reflect each gene's importance in development, we measure the size of these effects using our Δ statistic (6). Fitting a separate TRENDS model to each gene's measurements, we thus produce a ranking of the genes' presumed developmental importance. If instead, one's goal is simply to pinpoint high-confidence candidate genes relevant at all in development (ignoring the degree to which their expression transforms in the developmental progression), then our permutation test can be applied to establish which genes exhibit strong statistical evidence of an underlying nonconstant TREND effect. For all methods, p-values are obtained using the same procedure as in the simulation study (1000 permutations with step-down minP multiple-testing correction). In these analyses, significance testing (which identifies high-

confidence effects) and the Δ statistic (which identifies very large effects) both produce informative results.

As the myoblast data only contains four ℓ-levels and one batch from each, the TRENDS permutation test stringently identifies only 20 genes with significant non-constant trend at the 0.05 level (with multiple-testing correction). Terms which are statistically overrepresented in the Gene Ontology (GO) annotations of these significant genes (Kamburov et al. 2011), indicate the known developmental relevance of a large subset (see Figure 5A). Enriched biological process annotations include "anatomical structure development" and "cardiovascular system development" (Table S2A). In contrast, the cortex data are much richer, and TRENDS accordingly finds far stronger statistical evidence of trending genes, identifying 212 as significant (at the 0.05 level with multiple testing correction). A search for GO enriched terms in the annotations of these genes shows a large subset to be developmentally relevant (Figure 5B), with enriched terms such as "neurogenesis" and "nervous system development" (Table S2B). Due to the limited batches in these scRNA-seq data (each of which may be corrupted under our model), the TRENDS significance-tests act conservatively (a desirable property given the pervasive noise in scRNA-seq data), identifying small sets of genes we have high-confidence are primarily developmentally relevant.

Figure 5: Word clouds of terms significantly enriched (at the 0.01 level) in GO annotations of the genes with significantly trending expression in each analysis (Kamburov et al. 2011).

Ranking the genes by their TRENDS-inferred developmental effects (using Δ), 9 of the top 10 genes in the myoblast experiment have been previously discovered as significant regulators of myogenesis and some are currently employed as standard markers for different stages of differentiation (see Table S3A). Also, 7 of the top 10 genes in the cortex analysis have been previously implicated in brain development, particularly in sensory

regions (Table S3B). Thus, TRENDS accurately assigns the largest inferred effects to clearly developmental genes (see also Table S4). Since experiments to probe putative candidates require considerable effort, this is a very desirable feature for studying less well-characterized developmental systems than our cortex/myoblast examples. Figure 1A shows TRENDS predicts that MT2A (the gene with the largest Δ-inferred effect in myogenesis and a known regulator of this process) is universally down-regulated in development across the entire cell population. Interestingly, the majority of cells express MT2A at a uniformly high level of $\geqslant 3$ log FPKM just before differentiation is induced, but almost no cell exhibits this level of expression 24 hours later. MT2A expression becomes much more heterogenous with some cells retaining significant MT2A expression for the remainder of the time course while others have stopped expressing this gene entirely by the end. TRENDS accounts for all of these different changes via the Wasserstein distance which appropriately quantifies these types of effects across the population.

Because any gene previously implicated in muscle development is of interest in the myoblast analysis, we can form a lower-bound approximation of the fraction of "true positives" discovered by different methods by counting the genes with a GO annotation containing both the words "muscle" and "development" (e.g. "skeletal muscle tissue development"). Table S5 contains all GO annotations meeting this criterion. Figure 6A depicts a pseudo-sensitivity plot based on this approximation over the genes with the highest presumed developmental importance inferred under different methods. Here, the Tobit models are censored regressions specifically designed for scRNA-seq data, which solely model conditional expectations rather than the full distribution of expression across the cells (see §S3). A larger fraction of the top genes found by TRENDS and our closely-related Linear TRENDS method have been previously annotated for muscle development than top candidates produced by the other methods.

We repeat this analysis for the cortex data using a different set of "ground truth" GO annotations (listed in Table S6), and again find that TRENDS produces higher sensitivity

than the other approaches (Figure 6B) based on this crude measure. As researchers cannot practically probe a large number of genes in greater detail, it is important that a computational method for developmental gene discovery produces many high ranking true positives which can be verified through limited additional experimentation. While TRENDS appears to display greater sensitivity than other methods, we note that it is difficult to evaluate other performance-metrics (e.g. specificity) using the scRNA-seq data, since the complete set of genes involved in relevant developmental processes remains unknown.
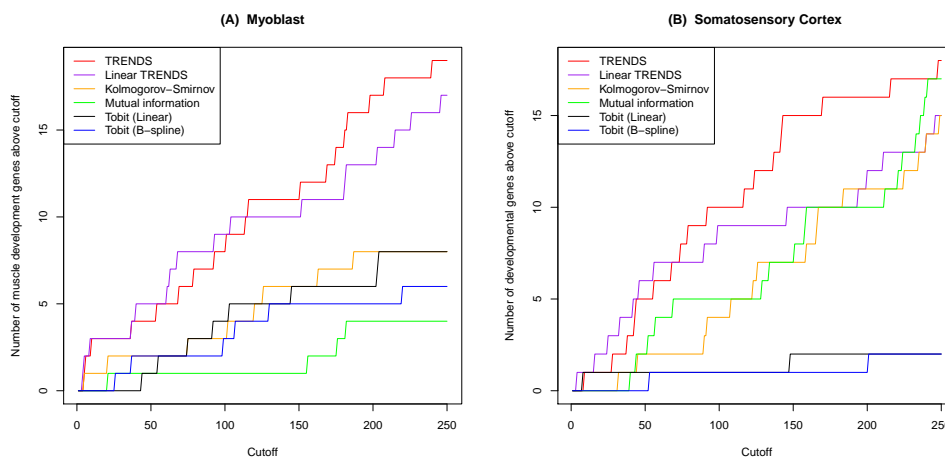

Figure 6: Pseudo-sensitivity of various methods based on their ability to identify known developmental genes. (A) the number of genes with a GO annotation containing both "muscle" and "development" found in the top K genes (ranked by the different methods for the myoblast data), over increasing K. (B) similar plot for the cortex data, where developmental genes are now those annotated with a relevant GO term from Table S6.

The Nestin gene in the myoblast data provides one example demonstrating the importance of treating full expression distributions rather than just mean-effects. Nestin plays an essential role in myogenesis, determining the onset and pace of myoblast differentiation, and its overexpression can also bring differentiation to a halt (Pallari et al. 2011), a process possibly underway in the high-expression cells from the later time points depicted in Figure 1B. TRENDS ranks Nestin 35th in terms of inferred developmental effect-size (with TRENDS p-value $= 0.02$ before multiple-testing correction and 0.09 after), but

this gene is overlooked by the scalar regression methods (only ranking 3,291 and 5,094 in the linear / B-spline Tobit results). Although Figure 1B depicts a clear temporal effect on mean Nestin expression, scalar regression does not prioritize this gene because these methods fail to properly consider the full spectrum of changes affecting different segments of the cell population in the multitude of other genes with similar mean-effects as Nestin.

Although the closely-related Linear TRENDS model appears to do nearly as well as TRENDS in our Figure 6 pseudo-sensitivity analysis, we find the linearity assumption overly restrictive, preventing the Linear TRENDS model from identifying important genes like TSPYL5, a nuclear transcription factor which suppresses levels of well-known myogenesis regulator p53 (Epping et al. 2011, Porrello et al. 2000). Linear TRENDS model only assigns this gene a p-value of 0.2 whereas TRENDS identifies it as significant ($p = 0.05$), since TSPYL5 expression follows a monotonic trend fairly closely ($R^2 = 0.95$) but is not as well approximated by a linear trend ($R^2 = 0.68$).

12. Discussion

While established methods exist to quantify change over a sequence of probability distributions, TRENDS addresses the scientific question of how much of the observed change can be attributed to sequential progression rather than nuisance variation. Although the TF algorithm resembles quantile-modeling techniques, our ideas are grounded under the unifying lens of the Wasserstein distance, which we use to measure effects (6), goodness-of-fit (5), and a distribution-based least-squares fit (4). Like linear regression, an immensely popular scientific method despite rarely reflecting true underlying relationships, our TRENDS model is not intended to accurately model/predict the data, which are likely subject to many more effects than our simple *trend* definition encompasses. Rather, TRENDS quantifies effects of interest, which remain highly interpretable (via our Wasserstein-perspective) despite being considered across fully nonparametric populations.

We recommend our model for data in which the underlying population is heteroge-

neous (possibly subject to diverse effects), each batch contains many samples ($n_i \geqslant 50$), and the sequence of levels $L \geqslant 3$ is short enough that effects of interest should follow persistent trends. When considering TRENDS analysis, it is important to ensure that the primary effects of interest are a priori expected to follow our trend definition. For the developmental scRNA-seq data considered in this work, this is a reasonable assumption because the experiments typically focus on a limited window of the underlying process. Furthermore, the severe prevalence of nuisance variation makes it preferable to identify a high-confidence developmentally-relevant subset of genes (e.g. because they display consistent effects over time), rather than attempting to characterize the complete set of genes displaying interesting effects.

While our trend definition produces good empirical results in these scRNA-seq analyses (and encompasses various conceptually interesting effects discussed in §S1), we emphasize that adopting this assumption narrowly restricts the sort of effects measured by our approach. Our limited definition is unlikely to characterize more complex effects of interest in general settings (particularly for longer sequences), and future work should explore extensions such as allowing change-points in the model. Note that our proposed Wasserstein-least-squares fit objective and Wasserstein-R^2 measure remain applicable for more general classes of regression functions on distributions. Furthermore, Lemma 2 provides an alternative definition of a trend which also applies to multidimensional distributions, and thus may be useful for applications such as spatiotemporal modeling. Nevertheless, the basic TRENDS methodology presented in this work can produce valuable insights. As simultaneously-profiled cell numbers grow to the many-thousands thanks to technological advances (Macosko et al. 2015), significant discoveries may be made by studying the evolution of population-wide expression distributions, and TRENDS provides a principled framework for this analysis.

References

Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D. K. & Jaakkola, T. S. (2003), 'Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes', *Proceedings of the National Academy of Sciences* **100**(18), 10146–51.

Bijleveld, C., van der Kamp, L. J. T., Van Der Kamp, P., Mooijaart, A., Van Der Van Der Kloot, W. A., Van Der Leeden, R. & Van Der Burg, E. (1998), *Longitudinal Data Analysis: Designs, Models and Methods*, Sage Publications.

Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. (2003), 'A comparison of normalization methods for high density oligonucleotide array data based on variance and bias', *Bioinformatics* **19**(2), 185–193.

Bondell, H. D., Reich, B. J. & Wang, H. (2010), 'Non-crossing quantile regression curve estimation', *Biometrika* **97**(4), 825–838.

Boyle, J. & Dykstra, R. (1986), 'A Method for Finding Projections onto the Intersection of Convex Sets in Hilbert Spaces', *Lecture Notes in Statistics* **37**, 28–47.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C. & Stegle, O. (2015), 'Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells', *Nat Biotechnol* **33**(2), 155–60.

de Leeuw, J. (1977), 'Correctness of Kruskal's algorithms for monotone regression with ties', *Psychometrika* **42**(1), 141–144.

Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. (2014), 'Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells', *Science* **343**(6167), 193–196.

Epping, M. T., Meijer, L. A. T., Krijgsman, O., Bos, J. L., Pandolfi, P. P. & Bernards, R. (2011), 'TSPYL5 suppresses p53 levels and function by physical interaction with USP7', *Nat Cell Biol* **13**(1), 102–108.

Fan, J., Yao, Q. & Tong, H. (1996), 'Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems', *Biometrika* **83**(1), 189–206.

Ge, Y., Dudoit, S. & Speed., T. P. (2003), 'Resampling-based multiple testing for microarray data analysis', *Test* **12**(1), 1–77.

Geiler-Samerotte, K. A., Bauer, C. R., Li, S., Ziv, N., Gresham, D. & Siegal, M. L. (2013), 'The details in the distributions: why and how to study phenotypic variability.', *Current opinion in biotechnology* **24**(4), 752–9.

Good, P. (1994), *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, Spring-Verlag.

Hall, P., Wolff, R. C. L. & Yao, Q. (1999), 'Methods for Estimating a Conditional Distribution Function', *Journal of the American Statistical Association* **94**(445), 154–163.

Heidelberger, P. & Lewis, P. A. W. (1984), 'Quantile Estimation in Dependent Sequences', *Operations Research* **32**(1), 185–209.

Honorio, J. & Jaakkola, T. (2014), 'Tight Bounds for the Expected Risk of Linear Classifiers and PAC-Bayes Finite-Sample Guarantees', *Fourteenth International Conference on Artificial Intelligence and Statistics* .

Hyndman, R. J. & Fan, Y. (1996), 'Sample Quantiles in Statistical Packages', *The American Statistician* **50**(4), 361–365.

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H. & Herwig, R. (2011), 'ConsensusPathDB: toward a more complete picture of cell biology.', *Nucleic acids research* **39**, D712–7.

Keen, K. J. (2010), *Graphics for Statistics and Data Analysis with R*, Taylor & Francis.

Kharchenko, P. V., Silberstein, L. & Scadden, D. T. (2014), 'Bayesian approach to single-cell differential expression analysis', *Nat. Meth* **11**(7), 740–742.

Krishnaswamy, S., Spitzer, M. H., Mingueneau, M., Bendall, S. C., Litvin, O., Stone, E., Pe'er, D. & Nolan, G. P. (2014), 'Conditional density-based analysis of T cell signaling in single-cell data', *Science* **346**(6213).

Levina, E. & Bickel, P. (2001), 'The Earth Mover's distance is the Mallows distance: some insights from statistics', *Proceedings. Eighth IEEE International Conference on Computer Vision* **2**, 251–256.

Macosko, E., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A., Kamitaki, N., Martersteck, E., Trombetta, J., Weitz, D., Sanes, J., Shalek, A., Regev, A. & McCarroll, S. (2015), 'Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets', *Cell* **161**(5), 1202–1214.

Mueller, J. & Jaakkola, T. (2015), 'Principal Differences Analysis: Interpretable Characterization of Differences between Distributions', *Advances in Neural Information Processing Systems* pp. 1702–1710.

Pallari, H.-M., Lindqvist, J., Torvaldson, E., Ferraris, S. E., He, T., Sahlgren, C. & Eriksson, J. E. (2011), 'Nestin as a regulator of Cdk5 in differentiating myoblasts', *Molecular Biology of the Cell* **22**(9), 1539–1549.

Porrello, A., Cerone, M. A., Coen, S., Gurtner, A., Fontemaggi, G., Cimino, L., Piaggio, G., Sacchi, A. & Soddu, S. (2000), 'p53 regulates myogenesis by triggering the differentiation activity of pRb.', *The Journal of cell biology* **151**(6), 1295–1304.

Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. (2014), 'Normalization of RNA-seq data using factor analysis of control genes or samples', *Nature Biotechnology* **32**(9), 896–902.

Silverman, B. W. & Young, G. A. (1987), 'The bootstrap: To smooth or not to smooth?', *Biometrika* **74**(3), 469–79.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. & Rinn, J. L. (2014), 'The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells', *Nat. Biotechnol* **32**(4), 381–386.

Zeisel, A., Munoz-Manchado, A. B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J. & Linnarsson, S. (2015), 'Brain structure. Cell types in the mouse

cortex and hippocampus revealed by single-cell RNA-seq.', *Science* **347**(6226), 1138–42.

Zielinski, R. (2006), 'Small-Sample Quantile Estimators in a Large Nonparametric Model',
Communications in Statistics - Theory and Methods **35**(7), 1223–1241.

SUPPLEMENTARY MATERIAL FOR

Modeling Persistent Trends in Distributions

Contents

S1. Conceptual examples of trends

Example 1. Any sequence of *stochastically ordered* distributions follows a trend. 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 $F_1(x) \geqslant F_2(x)\ \forall x$ (equivalently characterized as $\Pr(X_1 > x) \leqslant \Pr(X_2 > x)\ \forall x$) (Shaked & Shanthikumar G. 1994, Wolfstetter 1993). 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. Our trend definition also encompasses sequences where the distributions at intermediate values of ℓ are *monotonic quantile mixtures* of P_1 and P_L, i.e.

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

$$\text{s.t. } \{\omega_\ell \in [0,1] : \ell = 1, \ldots, L\} \text{ form a monotonic sequence} \tag{14}$$

Quantile mixtures are typically more appropriate than mixture distributions when there is no evident switching mechanism between distributions in the data-generating process (Gilchrist 2000). Condition (14) thus naturally characterizes the situation in which the underlying forces of interest gradually evolve distribution P_1 into P_L over $\ell = 1, \ldots, L$.

Example 3. In many applications, each P_ℓ is a mixture of the *same* K underlying subpopulation-specific distributions, where we let G_k denote the CDF of the kth subpopulation-specific distribution (mixing component) with ℓ-dependent mixing propor-

tion $\pi_\ell^{(k)}$. Each observed distribution can thus be expressed as:

$$\forall \ell \in \{1, \ldots, L\} : \ F_\ell = \sum_{k=1}^{K} \pi_\ell^{(k)} G_k \quad \text{where } \forall \ k, \ell : \ \pi_\ell^{(k)} \in [0, 1] \ , \ \pi_\ell^{(K)} = 1 - \sum_{k=1}^{K-1} \pi_\ell^{(k)} \ (15)$$

Here, the effects of interest alter the mixing proportions, so that a fraction of the individuals of one subpopulation transition to become part of another as ℓ 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 ensure the generality of this example, we avoid imposing a specific parameterization for 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 that imposing a stochastic ordering is much weaker than requiring G_k to have disjoint support).

To formalize the types of migration between subpopulations which meet our trend criterion, 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 $(\ell - 1) \to \ell$ (i.e. $\pi_\ell^{(i)} = \pi_{\ell-1}^{(i)} - \Delta$ and $\pi_\ell^{(j)} = \pi_{\ell-1}^{(j)} + \Delta$), then directed edges $i \to (i+1), (i+1) \to (i+2), \ldots, (j-1) \to j$ 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 $(\ell - 1) \to \ell$ is handled more delicately: First, we identify the sequence $\mathcal{S}$ of operations which produces the optimal transformation from mixing proportions vector $[\pi_{\ell-1}^{(1)}, \ldots, \pi_{\ell-1}^{(K)}] \to [\pi_\ell^{(1)}, \ldots, \pi_\ell^{(K)}]$, 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 $\mathcal{S}$, we introduce an edge into $\mathcal{G}$ between the corresponding nodes k and $k+1$ whose direction is specified by the sign of Δ (edge $k \to (k+1)$ if $\Delta > 0$, the reverse edge otherwise).

$\mathcal{G}$ 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 changes between $(\ell - 1) \rightarrow \ell$ as described above. Then, the sequence of distributions $P_1, \ldots, P_L$ follows a trend if $\mathcal{G}$ 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 ℓ increases, do not also cause migration in the reverse direction between these subpopulations at different values of ℓ. Figure 2D depicts an example of an evolving 3-component mixture model which follows a trend.

S2. Permutation testing with small batch numbers

Unfortunately, in many settings of interest such as most currently existing scRNA-seq 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 (e.g. $N < 7$). 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 entirely valid as it discards the (often substantial) correlations between genes and has been found to produce suboptimal results (Phipson & Smyth 2010).

To circumvent these issues, we propose a variant of our label-permutation-based procedure to obtain finer-grained but only approximate p-values (in the small N, L setting, rough approximations are all one can hope for since asymptotics-derived p-values are also error-prone). The underlying goal of our heuristic is to produce a richer picture of the

3

null distribution of R^2 (at the cost of resorting to approximation), which is accomplished as follows:

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

2. For data in which each distribution $\widehat{P}_i$ 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 $\widehat{P}_i^{(k)}$ using $\{Z_{i,s}^{(k)}\}_{s=1}^{n_i}$, and assemble the kth null-distributed dataset (under the current labels-permutation) by pairing the bootstrapped empirical distributions with the permuted labels ℓ_i^{perm}.

3. Apply TRENDS to each null-distributed dataset $\{(\ell_i^{\text{perm}}, \widehat{P}_i^{(k)})\}_{i=1}^N$ and compute a $R^2_{\text{perm},k}$ value via (5) which is distributed according to the desired null (where $K = 1$ and $\widehat{P}_i^{(k)} = \widehat{P}_i$ if bootstrapping is not performed).

4. Form a smooth approximation of the null distribution by fitting a kernel CDF estimate $\widehat{F}$ 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 (Altman & Leger 1995)). Finally, the approximate p-value is computed as $\widehat{p} := 1 - \widehat{F}(R^2)$, 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 ℓ_i corresponding to the actual ordering or its reverse are equally likely a priori as any other

4

permutation of the ℓ_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 (Tsai & Chen 2007), and point out that as the number of distributions per level N_ℓ grows, the approximation factor of our procedure shrinks, as is the case for p-values based on asymptotics which are themselves only approximations.

S2.1. Evaluating TRENDS p-values

Under the simulation setup of §10, we investigate the performance of our permutation technique to obtain TRENDS p-values. We draw samples from each of the underlying models S_1, S_2, S_3 with $n = 100, N_\ell = 1$, and $\sigma = 0.1$. To each simulated dataset (in total, 100 datasets are drawn from each model), we apply the TRENDS model and then determine the significance of the TRENDS R^2 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 these 1000 null samples as detailed in the technique described above for obtaining 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 estimated as follows: a minimum of $J = 10,000$ new datasets (i.e. batch sequences) from the same underlying model are drawn in which ℓ 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 corresponding 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 begin with $J = 10,000$ 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 S1 demonstrates that our approximation produces much better p-values than the basic permutation method.

Model	Average p	$\mathbb{E}[\widehat{p} - p]$	SD($\widehat{p}$)	MSE($\widehat{p}$)	$\mathbb{E}[p_{\text{perm}} - p]$	SD(p_{perm})	MSE(p_{perm})
S_1	0.13	-0.012	0.036	**1.2e-3**	-0.015	0.036	1.3e-3
S_2	0.19	0.039	0.068	**5.2e-3**	0.085	0.117	1.8e-2
S_3	0.51	0.056	0.084	**8.8e-3**	0.092	0.157	2.8e-2

Table S1: Comparing our approximate p-values ($\widehat{p}$) against the standard permutation test (p_{perm}). Column 2 lists the average true p-value (over 100 datasets) for each model S_1-S_3.

S3. Description of alternative methods

Here, we describe different methods that TRENDS is compared against. Note that the methods which model full distributions may be ordered based on increasing generality of the underlying assumption as follows: Linear TRENDS $\rightarrow$ TRENDS $\rightarrow$ KS / MI. By selecting a model later in this ordering, one can capture a wider diversity of underlying effects but only with decreased statistical power (and robustness against batch-effects).

S3.1. Kolmogorov-Smirnov method (KS)

This approach performs an omnibus test of the hypothesis that there exist ℓ_1 and ℓ_2 such that $\Pr(X \mid \ell_1) \neq \Pr(X \mid \ell_2)$. 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 $\ell_1 < \ell_2 \in \{1, \ldots, L\}$. Statistical significance is assessed via permutation testing, since the usual asymptotics are no longer valid after maximization.

S3.2. Mutual information method (MI)

Here, we estimate the size of the effect using the mutual information between ℓ and X. Because we operate in the fixed-design setting, ℓ is technically not a random variable, so we instead employ a conditional variant of the mutual information in which the marginal distribution of ℓ is disregarded, following the DREMI method of Krishnaswamy et al. (2014). First, we simply reweigh our batches to ensure the marginal distribution of ℓ is uniform over $\{1, \ldots, L\}$ in the given labels $\{\ell_i\}_{i=1}^N$. Subsequently, kernel density estimates of the reweighed joint (X, ℓ) distribution as well as each conditional $\Pr(X \mid \ell_1)$ are used to calculate the (conditional) mutual information, which is used to produce a ranking of genes' inferred developmental importance according to this method. A p-value is obtained via permutation testing, using the mutual information as the test statistic.

S3.3. Linear TRENDS (LT) model

This method is very similar to TRENDS, except it uses a more restrictive class of regression functions where each quantile evolves linearly (rather than the assumption of monotonicity used in our trend criterion). We thus operate on real-valued rather than ordinal covariates (e.g. the actual values of the time points t_ℓ when available in the scRNA-seq context, or the integer ℓ-values when there are no definitive numerical batch-labels, as in our simulation study). Linear TRENDS also relies on our notion of Wasserstein least-squares

fit, the Δ effect-size measure (used to rank genes), and the same permutation-procedure for testing significance in TRENDS (the sole difference between these models is that the former accounts for covariate scaling assuming that effects manifest linearly on this scale).

A similar linear multiple-quantile regression framework has been previously proposed in numerous contexts, although it is designed only for simultaneously estimating a few specific quantiles of the conditional distribution (Takeuchi et al. 2006, Bondell et al. 2010). Takeuchi et al. and Bondell et al. both fit this model jointly over the quantiles of interest via a quadratic program with constraints to ensure non-crossing quantiles. Linear quantile regression (with non crossing) could nonetheless be employed to model the full distribution by simply selecting a grid of quantiles spanning $(0, 1)$ as is done in TRENDS, but note that simple scalar measures such as our Δ and R^2 values do not exist in standard quantile regression which lacks the unifying Wasserstein perspective presented in this work.

In our setting, the empirical quantiles of each conditional distribution are available, so one can directly employ the usual squared error loss on the empirical quantiles themselves (as done in our TF algorithm) rather than relying on the quantile regression loss function used by Takeuchi et al. and Bondell et al. 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_ℓ, the time at which the batch is sampled. By replacing the PAVA step (over ℓ) of the TF algorithm with standard linear regression (where t_ℓ is the sole covariate) and also omitting the δ-search for the split between increasing and decreasing quantiles, our alternating projections method is trivially adapted to fit the set of non-crossing quantile linear regressions under the squared-loss. In the case where we estimate around 100 quantiles to represent the entire distributions, we find that this linearized TF algorithm is orders of magnitude faster than the quadratic program, which has difficulty dealing with the large number of constraints required in this setting (these methods are not intended to estimate full distributions).

We therefore fit the Linear TRENDS model using this linearized TF algorithm in our applications (computational efficiency is crucial when the model is fit thousands of times as in our gene-expression analyses), and find that besides the marked runtime improvement, Linear TRENDS produces nearly identical estimates as the linear multiple-quantile regression model of Bondell et al.

S3.4. Tobit model (censored regression)

Trapnell et al. (2014) introduce a scalar regression model specifically tailored for the analysis of single-cell gene expression over time (which only considers conditional expectations rather than the complete expression distribution across the cell population). Their approach ranks genes based on the significance of the regression coefficients in a Tobit-family generalized additive model fit to log-FPKM values vs. time. It is thus assumed that measured expression follows 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 highly expressed (this missing data issue plagues scRNA-seq measurements due to the small amount of RNA that can be isolated from one cell). We try both directly regressing X against t_ℓ (referring to this generalized linear model as the linear Tobit), as well as initially using a B-spline basis expansion of the t_ℓ values so the subsequent Tobit regression can capture diverse nonlinear effects (Trapnell et al. 2014).

S4. Simulation study details

Our negative binomial distribution parameters r_ℓ and p_ℓ correspond to the arguments `size` and `prob` used by the `NegBinomial` functions in the R `stats` package (here, a negative binomial random variable represents the number of failures occurring in a series of Bernoulli trials before r_ℓ successes take place). To ensure we are sampling from valid distributions after the introduction of noise, we subsequently enforce the following additional constraints: $\widetilde{r}_\ell \geqslant 1$, $0.05 \leqslant \widetilde{p}_\ell \leqslant 0.95$ before drawing our observations.

S5. Single cell RNA-seq analysis details

Trapnell et al. (2014) 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. After inducing differentiation in a culture of primary human myoblast cells, cells were sampled (and sequenced) in batches every 24 hours. While the microfluidic system in this experiment can capture 96 cells (one batch is sampled per time point), some of the captures contain visible debris and cannot be confirmed to come from a whole single cell. In addition to discarding these, Trapnell et al. stringently omit cells whose libraries were not sequenced deeply (≥ 1 million reads), since their analysis uses high-dimensional manifold methods which are not robust to noise. Because TRENDS is designed to distinguish biological effects from noise, we retain these cells embracing the additional (albeit noisy) insight on underlying expression. Omitting only the debris-cells, the data[2] we analyze consists of 17,341 genes profiled in the following number of cells at each time point: 0h: 93 cells, 24h: 93 cells, 48h: 93 cells, 72h: 76 cells.

In a scRNA-seq experiment of much larger scale, Zeisel et al. (2015) isolated 1,691 cells from the somatosensory cortex (the brain's sensory system) of juvenile CD1 mice aged P22-P32. We treat age (in postnatal days) as our covariate, whose ordinal representation takes one of $L = 10$ possible levels. Numerous batches of cells were captured from some identically-aged mice, implying $N_\ell > 1$ for many ℓ, and a total of 14,575 genes have nonzero expression measurements[3] in the sampled cells.

In both analyses, gene expression is represented in ($\log_{10}(x+1)$ transformed) Fragments Per Kilobase of transcript per Million mapped reads (FPKM) (Trapnell et al. 2014). Although TRENDS is nonparametric and can be applied to any expression representation, we find log-FPKM values favorable due to their interpretability and direct comparability

---

[2]Myoblast FPKM values are available in the Gene Expression Omnibus under accession GSE52529.

[3]We compute FPKM values from the somatosensory cortex sequencing read counts available in the Gene Expression Omnibus under accession GSE60361.

between different genes. The methods we compare TRENDS against (§S3) are all suited for log-FPKM values and do not hinge on the specific distributional assumptions often required for other expression-measures such as read counts Risso et al. (2014) or negative-binomial rates Kharchenko et al. (2014).

The word clouds (Figure 5) summarizing enriched biological process terms in the GO annotations for genes with significantly trending expression were made using the ConsensusPathDB[4] tool (Kamburov et al. 2011). Table S2 provides additional detail listing the most highly enriched terms identified in the significantly trending gene set from each dataset. Table S3 contains previously characterized developmental genes found among those with the ten largest TRENDS Δ values (i.e. the genes with the largest inferred effect-size). Table S4 lists the highly enriched GO terms (again found via Consensus-PathDB) in the 100 genes with largest Δ values in each dataset.

<table>
<tr><td colspan="3" align="center">(A) Myoblast</td><td colspan="3" align="center">(B) Somatosensory Cortex</td></tr>
<tr><td>Term</td><td>p-value</td><td>q-value</td><td>Term</td><td>p-value</td><td>q-value</td></tr>
<tr><td>liver development</td><td>1e-4</td><td>6e-3</td><td>transmission of nerve impulse</td><td>6e-8</td><td>2e-5</td></tr>
<tr><td>hepaticobiliary system development</td><td>1e-4</td><td>6e-3</td><td>multicellular organismal</td><td></td><td></td></tr>
<tr><td>anatomical structure development</td><td>3e-4</td><td>8e-3</td><td> signaling</td><td>1e-7</td><td>3e-5</td></tr>
<tr><td>gland development</td><td>3e-4</td><td>0.03</td><td>cell communication</td><td>6e-7</td><td>7e-5</td></tr>
<tr><td>system development</td><td>2e-3</td><td>0.08</td><td>neuron differentiation</td><td>1e-6</td><td>2e-4</td></tr>
<tr><td>regulation of cyclin-dependent protein</td><td></td><td></td><td>cell development</td><td>3e-6</td><td>2e-4</td></tr>
<tr><td> serine/threonine kinase activity</td><td>2e-3</td><td>0.08</td><td>ensheathment of neurons</td><td>3e-6</td><td>2e-4</td></tr>
<tr><td>single-multicellular organism process</td><td>3e-3</td><td>0.04</td><td>axon ensheathment</td><td>3e-6</td><td>3e-4</td></tr>
<tr><td>single-organism</td><td></td><td></td><td>single organism signaling</td><td>4-e6</td><td>3e-4</td></tr>
<tr><td> developmental process</td><td>4e-3</td><td>0.04</td><td>neurogenesis</td><td>1e-5</td><td>1e-3</td></tr>
<tr><td>central nervous system development</td><td>5e-3</td><td>0.07</td><td>regulation of biological quality</td><td>1e-5</td><td>4e-4</td></tr>
<tr><td>cardiovascular system development</td><td>5e-3</td><td>0.07</td><td>system development</td><td>1e-5</td><td>5e-4</td></tr>
<tr><td>circulatory system development</td><td>5e-3</td><td>0.07</td><td>neuron projection development</td><td>1e-5</td><td>1e-3</td></tr>
<tr><td>multicellular organismal</td><td></td><td></td><td>cell projection organization</td><td>1e-5</td><td>5e-4</td></tr>
<tr><td> development</td><td>5e-3</td><td>0.08</td><td>single-organism cellular process</td><td>2e-5</td><td>4e-4</td></tr>
<tr><td>cellular nitrogen compound</td><td></td><td></td><td>neuron development</td><td>2e-5</td><td>1e-3</td></tr>
<tr><td>catabolic process</td><td>5e-3</td><td>0.07</td><td>anatomical structure development</td><td>3e-5</td><td>5e-4</td></tr>
<tr><td>response to hormone</td><td>5e-3</td><td>0.08</td><td>nervous system development</td><td>3e-5</td><td>2e-3</td></tr>
<tr><td>nervous system development</td><td>e-3</td><td>0.07</td><td>cellular developmental process</td><td>5e-5</td><td>6e-4</td></tr>
<tr><td>heart development</td><td>5e-3</td><td>0.08</td><td>cell differentiation</td><td>6e-5</td><td>2e-3</td></tr>
<tr><td>regulation of cell cycle</td><td>6e-3</td><td>0.07</td><td>single-organism</td><td></td><td></td></tr>
<tr><td>organ development</td><td>6e-3</td><td>0.08</td><td> developmental process</td><td>7e-5</td><td>7e-4</td></tr>
</table>

Table S2: Most highly enriched terms in the biological process annotations of significantly trending genes. The *p*-values correspond to the statistical significance of each term's enrichment in the set of genes (false-discovery-rate correction produces *q*-values).

---

[4]ConsensusPathDB Link: `http://cpdb.molgen.mpg.de`

(A) Myoblast

Gene	Δ	R^2	p-value	Developmental Evidence
MT2A	0.46	0.98	0.11	Apostolova et al. (1999)
ACTA2	0.44	0.99	0.08	Petschnik et al. (2010)
MT1L	0.43	0.99	0.09	Apostolova et al. (1999)
TNNT1	0.42	0.95	0.13	Sebastian et al. (2013)
MYLPF	0.41	0.99	0.03	Sebastian et al. (2013)
MYH3	0.39	0.99	0.04	Trapnell et al. (2014)
MT1E	0.39	0.99	0.11	Apostolova et al. (1999)
AC004702.2	0.37	0.99	0.23	Unknown
FABP3	0.35	0.98	0.18	Myers et al. (2013)
DKK1	0.34	0.99	0.12	Han et al. (2011)

(B) Somatosensory Cortex

Gene	Δ	R^2	p-value	Developmental Evidence
Sst	0.23	0.22	0.05	Zeisel et al. (2015)
Xist	0.14	0.09	0.35	Unknown
Ptgds	0.13	0.24	0.02	Trimarco et al. (2014)
Plp1	0.13	0.16	0.14	Zeisel et al. (2015)
Mog	0.13	0.13	0.16	Zeisel et al. (2015)
Npy	0.12	0.11	0.23	Zeisel et al. (2015)
Rps26	0.11	0.12	0.20	Unknown
Tsix	0.11	0.12	0.23	Unknown
Apod	0.11	0.16	0.11	Sanchez et al. (2002)
Ermn	0.10	0.11	0.20	Zeisel et al. (2015)

Table S3: The top ten inferred developmental genes (with the largest Δ value) from each experiment. Shown are the TRENDS Δ, R^2, and p-value (after multiple-testing correction) for each gene, as well as existing literature (if known) which previously characterized the gene as playing an important role in developmental processes.
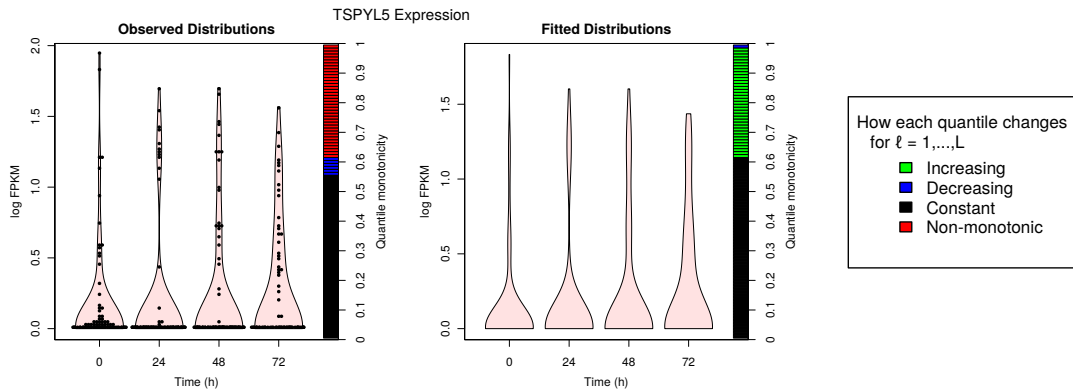

Figure S1: Violin plots depicting the empirical distribution of TSPYL5 expression measured in myoblast cells (on left), and the corresponding TRENDS fitted distributions (on right). Each point shows a sampled cell.

(A) Myoblast

Term	p-value	q-value
actin-mediated cell contraction	4e-9	9e-7
muscle structure development	6e-9	1e-6
striated muscle tissue development	8e-9	9e-7
muscle tissue development	1e-8	2e-6
muscle organ development	1e-8	2e-6
response to zinc ion	2e-8	2e-6
actin filament-based movement	3e-8	2e-6
organ development	1e-7	1e-5
muscle system process	1e-7	7e-6
response to inorganic substance	2e-7	1e-5
muscle contraction	2e-7	2e-5
negative regulation of growth	2e-7	1e-5
response to metal ion	2e-7	1e-5
mitotic cell cycle	3e-7	1e-5
response to transition metal nanoparticle	5e-7	2e-5
cellular response to metal ion	5e-7	3e-5
cellular response to inorganic substance	1e-6	4e-5
muscle cell development	2e-6	6e-5
cell cycle	5e-6	2e-4
muscle tissue morphogenesis	6e-6	2e-4
muscle organ morphogenesis	9e-6	2e-4
heart development	1e-5	4e-4
regulation of mitotic cell cycle	1e-5	6e-4
striated muscle cell development	2e-5	6e-4

(B) Somatosensory Cortex

Term	p-value	q-value
ensheathment of neurons	2e-10	3e-8
axon ensheathment	2e-10	5e-8
cellular homeostasis	3e-8	2e-6
cellular chemical homeostasis	4e-8	4e-6
transmission of nerve impulse	7e-8	5e-6
multicellular organismal signaling	1e-7	6e-6
glial cell differentiation	3e-7	2e-5
regulation of biological quality	4e-7	2e-5
glial cell development	7e-7	3e-5
chemical homeostasis	2e-6	6e-5
response to inorganic substance	4e-6	1e-4
homeostatic process	8e-6	3e-4
nervous system development	1e-5	5e-4
response to metal ion	2e-5	6e-4
response to oxygen-containing compound	4e-5	1e-3
system development	6e-5	1e-3
central nervous system development	6e-5	2e-3
detoxification of copper ion	7e-5	2e-3
response to steroid hormone stimulus	1e-4	2e-3
response to lipid	1e-4	2e-3
response to reactive oxygen species	2e-4	3e-3
response to toxic substance	2e-4	3e-3
anatomical structure development	2e-4	6e-3
neurogenesis	3e-4	5e-3

Table S4: Most highly enriched terms in the biological process annotations of the top 100 genes with largest Δ values in each experiment. The p-values correspond to the statistical significance of each term's enrichment in the set of genes (false-discovery-rate correction produces q-values).

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

Table S5: A list of all GO annotation terms containing both the words "muscle" and "development", used to produce the pseudo-sensitivity plots in Figure 6A.

	Gene Ontology ID	Annotation Term
1	GO:0007420	brain development
2	GO:0007399	nervous system development
3	GO:0014003	oligodendrocyte development
4	GO:0021860	pyramidal neuron development
5	GO:0022008	neurogenesis

Table S6: A list of the GO annotation terms relevant to the somatosensory cortex development, used to produce the pseudo-sensitivity plots in Figure 6B. This brain region is primarily composed of oligodendrocyte and pyramidal neuron cells (Zeisel et al. 2015).

S6. Model checking

In this section, we perform another simulation to demonstrate our proposed procedure for checking whether the TRENDS model is appropriate in analyses lacking prior domain knowledge about the effects of interest. Samples are generated from one of the following choices of the underlying trending distribution sequence $Q_1, \ldots, Q_L$ (with $L = 7$):

(R_1) $Q_\ell \sim N(0,1)$ for $\ell = 1, \ldots, 7$.

(R_2) $Q_\ell \sim N(\mu_\ell, 1)$ with $\mu_\ell = 0, 0.1, 0.1, 0.2, 0.5, 0.9, 1$ for $\ell = 1, \ldots, 7$.

(R_3) $Q_\ell \sim N(\mu_\ell, 1)$ with $\mu_\ell = 0, 0.1, 0.3, 0.5, 0.4, 0.2, 0$ for $\ell = 1, \ldots, 7$.

Note that the underlying sequence of distributions for R_3 severely violates our trend condition. Under each of these models, observed values for the ith batch is generated according to $x_{i,s} = \widetilde{x}_{i,s} + z_i$ where $\widetilde{x}_{i,s} \overset{iid}{\sim} Q_{\ell_i}$, and we independently draw a single noise-variable (i.e. batch-effect) $z_i \sim N(0, \sigma^2)$ for the entire batch.

For each quantile $p \in (0,1)$ used in our TRENDS-fit, we compute the value of the empirical residual function $\widehat{\mathcal{E}}_i(p) = \widehat{F}_i^{-1}(p) - \widehat{G}_{\ell_i}^{-1}(p)$, where $\widehat{F}_i^{-1}$ denotes the empirical quantiles of the distribution for the ith batch (estimated from $\{x_{i,s}\}_{s=1}^{n_i} \sim P_i$) and $\widehat{G}_{\ell_i}^{-1}$ denote the fitted quantiles produced by the TF algorithm applied the data (corresponding to inferred trending distributions Q_{ℓ_i}). Figure S2 depicts a diagnostic plot showing the distribution of $\widehat{\mathcal{E}}_i(p)$ vs. ℓ when TRENDS is fit to data from each of these models. Based on the clear pattern displayed by the residuals in the R_3 plot, one can easily correctly conclude that the TRENDS model is not very appropriate for this dataset. In contrast, the residual functions appear random for data from the other two underlying settings (which meet our TRENDS assumptions).

Under this simulation, we can evaluate the performance of our TRENDS estimates of misspecified effects. Motivated by our Δ statistic and Lemma 2, we employ the L_1 Wasserstein distance to define the true overall sequential-progression effect in this simulation as

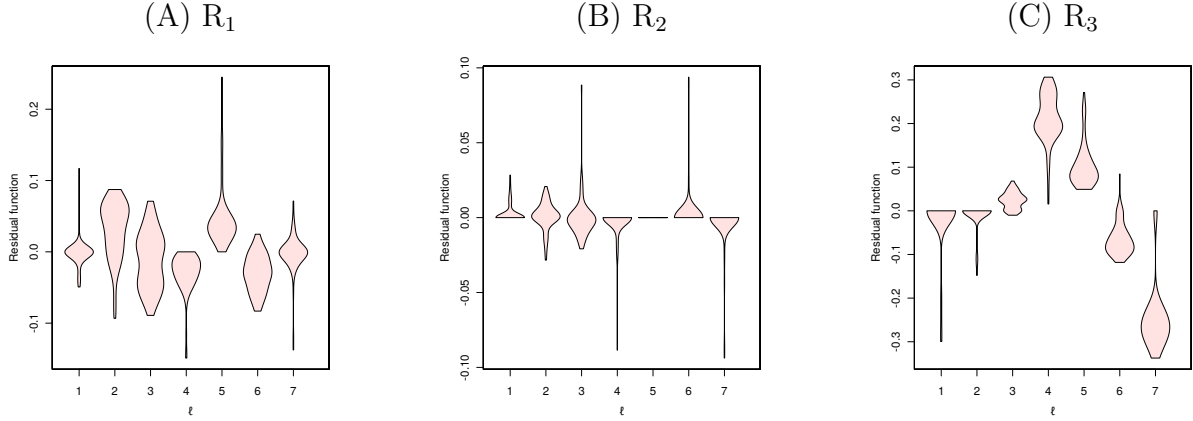|(A) R_1|(B) R_2|(C) R_3|

Figure S2: Diagnostic plot of the residual functions $\widehat{\mathcal{E}}_i(p)$ when TRENDS is fit to data from each underlying setting R_1, R_2, R_3 ($N_\ell = 1, n_i = 1000, \sigma = 0.1$). For each batch i, the plot depicts a kernel density estimate of the values taken by $\widehat{\mathcal{E}}_i(p)$ over $p = 0.01, 0.02, \ldots, 0.99$.

$\Delta_{\mathrm{true}} = \sum_{\ell=2}^{L} d_{L_1}(Q_{\ell-1}, Q_\ell)$, which is simply 1 for setting R_3. When all $N_\ell = 1$ (one batch per level), we can simply incorporate the Wasserstein distances between adjacent observed empirical distributions $\Delta_{\mathrm{emp}} = \sum_{\ell=2}^{L} d_{L_1}(P_{\ell-1}, P_\ell)$ as a basic estimate of Δ_{true}. Note that the batch-effects cause Δ_{emp} to have inflated variance beyond random-sampling deviations in the empirical quantile-estimates. In contrast, the Δ_{TRENDS} estimate produced by our TRENDS model is downwardly biased when applied to data from R_3, because of our restriction to monotone quantiles. Even in this misspecified setting, Figure S3 shows that under non-trivial amounts of noise, Δ_{TRENDS} remains a far superior estimator of Δ_{true} than Δ_{emp}, which is highly susceptible to variation arising from these batch-effects.

Finally, we investigate the residual functions when TRENDS is fit to the scRNA-seq data from genes known to play a major role in regulating developmental processes. Figure S4 does not indicate any systematic pattern in the residuals that would suggest our model is inappropriate for these data.
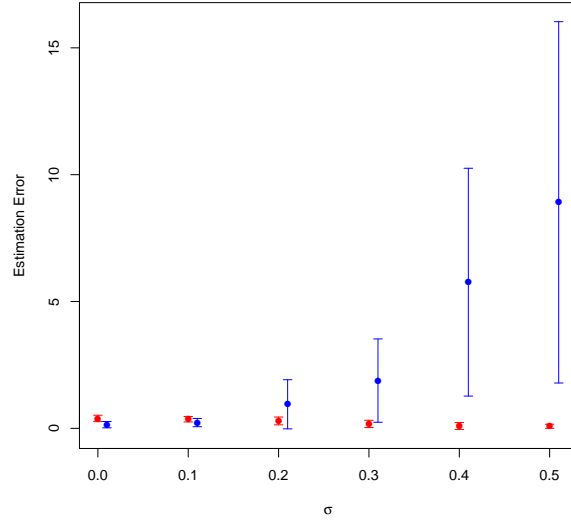
Figure S3: The mean/standard-deviation of the squared error of Δ_{emp} estimates (blue) and Δ_{TRENDS} estimates (red) over 100 datasets drawn from R_3 (under each value of σ, with $n_i = 100$ for each batch).
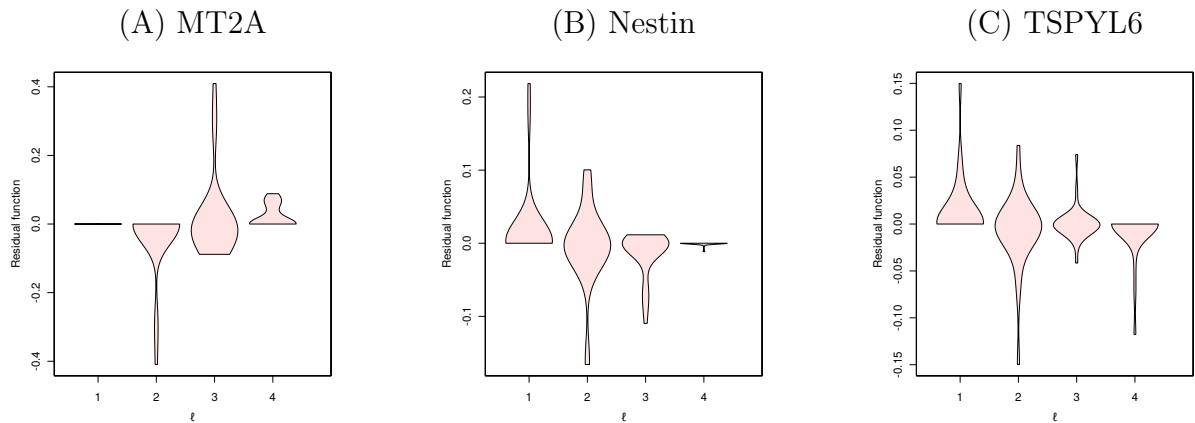

Figure S4: Diagnostic plot of the residual functions $\widehat{\mathcal{E}}_i(p)$ for TRENDS fit to scRNA-seq data from known regulatory genes of myoblast development. For each batch i, the plot depicts a kernel density estimate of the values taken by $\widehat{\mathcal{E}}_i(p)$ over $p = 0.01, 0.02, \ldots, 0.99$.

S7.   ACS income distribution analysis

To demonstrate the broader utility of TRENDS beyond scRNA-seq analysis, we present a brief study of incomes in various industries during the years 2007-2013 following the economic recession. Our goal is to quantify and compare effects across different industries' incomes during this post-recession period. Rather than measuring ephemeral decline/rebound in this analysis, our interests lie in consistent effects which enduringly altered an industry's incomes through 2013. 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 Microdata Series (Ruggle et al. 2010). After filtering out individuals with missing or \$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 separately, treating the observations from each year as a single batch and year-index in this time series as the label ($\ell = 1, \ldots, 7$).

Industry	R^2	p-value	Δ
Other information services	0.97	0.02	5465
Software publishers	0.78	0.10	2991
Electronic auctions	0.86	0.04	2584
Oil and gas extraction	0.78	0.12	2454
Miscellaneous petroleum and coal products	0.52	0.38	2415
Other telecommunication services	0.80	0.07	2414
Pharmaceutical and medicine manufacturing	0.98	0.04	2220
Management of companies and enterprises	0.66	0.12	2194
Metal ore mining	0.89	0.02	2074
Support activities for mining	0.88	0.03	1915
Electric and gas, and other combinations	0.82	0.03	1910
Non-depository credit and related activities	0.92	0.06	1860
Sound recording industries	0.51	0.38	1731
Electronic component and product manufacturing	0.99	0.02	1719
Securities, commodities, funds, trusts, and other financial investments	0.57	0.23	1665
Agricultural chemical manufacturing	0.77	0.09	1635
Communications, and audio and video equipment manufacturing	0.72	0.09	1628
Pipeline transportation	0.70	0.14	1620
Coal mining	0.90	0.04	1573
Natural gas distribution	0.69	0.11	1546

Table S7: The 20 industries with annual incomes most affected by temporal progression from 2007-2013 (as inferred by TRENDS). Broader sectors are: manufacturing (red), business/finance (green), energy (blue), technology (magenta).
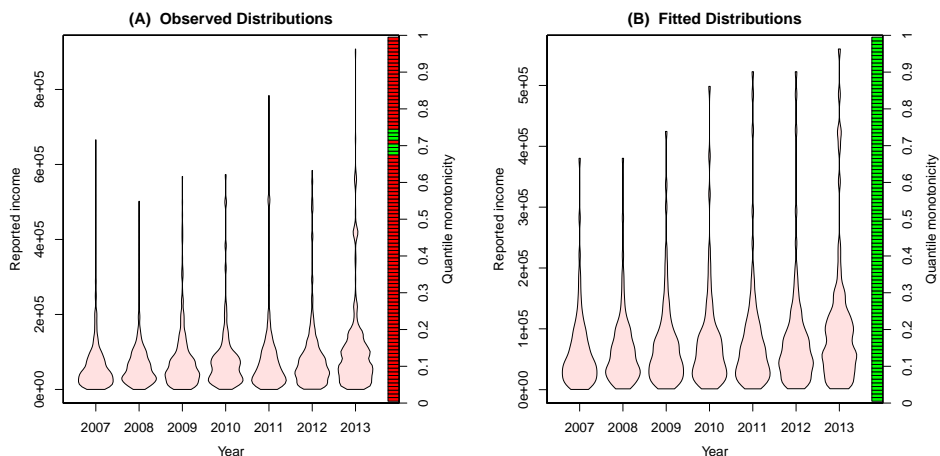
Figure S5: Distributions of reported income of individuals in the "other information services" industry. (A) kernel density estimates applied to the ACS survey results from each year (B) corresponding TRENDS fitted distributions.

Table S7 lists the industries which according to TRENDS are subject to the largest trending temporal effects in income distribution over this 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[5]. The other industries in which income distributions were subject to the largest temporal progression effects are predominantly technology-related, representing the continued growth in incomes in this sector, which has been unaffected by the recession.

Of particular note is the "other information services" industry (includes web search, internet publishing/broadcasting), where we observe the emergence of a distinct subgroup with reported incomes in the hundreds of thousands. While a few of the extreme reported incomes fell from 07-08, TRENDS conservatively estimates the underlying effects as consistently increasing all quantiles rather than including this change in Δ (such extrema are highly-variable, even at our large sample size). For reference, the average reported incomes of this industry in 2007-13 were: \$65.8k, \$66.6k, \$77.9k, \$78.7k, \$82.1k, \$84k.

---

[5]Reflecting 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.

S8.   Proofs and auxiliary lemmas

S8.1.   Proof of Lemma 1

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

$$\sum_{j=1}^{N} \int_0^1 \left(F_j^{-1}(p) - G^{-1}(p) \right)^2 \mathrm{d}p$$

$$= \int_0^1 \sum_{j=1}^{N} \left(F_j^{-1}(p) - H(p) - \frac{1}{N} \sum_{i=1}^{N} F_i^{-1}(p) \right)^2 \mathrm{d}p$$

$$\geq \int_0^1 \sum_{j=1}^{N} \left(F_j^{-1}(p) - \frac{1}{N} \sum_{i=1}^{N} F_i^{-1}(p) \right)^2 \mathrm{d}p$$

regardless of the value taken by $H(p)$ for each $p \in [0,1]$

$\square$

S8.2.   Proof of Lemma 2

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

$$d_{L_1}(P_i, P_j) = \int_0^1 \left| F_i^{-1}(p) - F_j^{-1}(p) \right| dp = \int_0^1 \sum_{\ell=i+1}^{j} \left| F_\ell^{-1}(p) - F_{\ell-1}^{-1}(p) \right| dp = \sum_{\ell=i+1}^{j} d_{L_1}(P_{\ell-1}, P_\ell)$$

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.

$\square$

S8.3. Proof of Theorem 1

Proof. We have:

$$\operatorname*{argmin}_{G_1^{-1},\ldots,G_L^{-1}} \left\{ \sum_{\ell=1}^{L} \sum_{i \in I_\ell} w_i \sum_{k=1}^{P-1} \left(\widehat{F}_i^{-1}(p_k) - G_\ell^{-1}(p_k) \right)^2 \left[\frac{p_{k+1} - p_{k-1}}{2} \right] \right\}$$

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

$$\equiv \operatorname*{argmin}_{v^{(1)},\ldots,v^{(L)}} \left\{ \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2} \right) \sum_{\ell=1}^{L} \sum_{i \in I_\ell} w_i \left(\widehat{F}_\ell^{-1}(p_k) - v_k^{(\ell)} \right)^2 \right\}$$

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

$$\text{s.t. } \forall \; k < k' \in \{1, \ldots, P-1\} : \begin{cases} \forall \ell : \; v_k^{(\ell)} < v_{k'}^{(\ell)} & \text{since } G_\ell^{-1} \text{ must be a valid quantile function} \\ v_k^{(1)}, \ldots, v_k^{(L)} \text{ is a monotone sequence whose direction } = \delta[k] \end{cases}$$

for one of the δ constructed in Step 6 or 8 of the procedure.

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

$$= \operatorname*{argmin}_{v^{(1)},\ldots,v^{(L)}} \left\{ \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2} \right) \sum_{\ell=1}^{L} w_\ell^* \left(\overline{\widehat{F}}_\ell^{-1}(p_k) - v_k^{(\ell)} \right)^2 \right\} \tag{16}$$

$$\text{s.t. } \forall \; k < k' \in \{1, \ldots, P-1\} : \begin{cases} \forall \ell : \; v_k^{(\ell)} < v_{k'}^{(\ell)} & \text{since } G_\ell^{-1} \text{ must be a valid quantile function} \\ v_k^{(1)}, \ldots, v_k^{(L)} \text{ is a monotone sequence whose direction } = \delta[k] \end{cases}$$

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

We will now show that for any δ constructed in Step 6 or 8, the corresponding y_ℓ produced by the AlternatingProjections algorithm are the optimal valid quantile-functions if we impose the additional constraint that for any k, the p_kth 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 δ, 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

$$d_W\left([v^{(1)},\ldots,v^{(L)}],[w^{(1)},\ldots,w^{(L)}]\right) = \sum_{k=1}^{P-1}\left(\frac{p_{k+1}-p_{k-1}}{2}\right)\sum_{\ell=1}^{L}w_\ell^*\left(v_k^{(\ell)}-w_k^{(\ell)}\right)^2 \qquad (17)$$

Lemmas 4 and 5 show that our AlternatingProjections algorithm is equivalent to Dykstra's method of alternating projections (Boyle & Dykstra 1986) 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) $\left[x^{(1)},\ldots,x^{(L)}\right]$ 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 Combettes & Pesquet (2011), Boyle & Dykstra (1986) 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 (8) under the additional constraint imposed by δ. $\qquad \square$

Lemma 3 (de Leeuw (1977)). *Given weights $w_1,\ldots,w_N \geqslant 0$ 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 $\widehat{y}_1,\ldots,\widehat{y}_L$ produced by tertiary-variant of PAVA are guaranteed to be the best-fitting nondecreasing sequence in the least-squares sense, i.e.*

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

Lemma 4. *Recall the definitions from the TF algorithm and the proof of Theorem 1. Given any $\left[x^{(1)},\ldots,x^{(L)}\right] \in \mathcal{X}$, 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*

$$y_k^{(1)},\ldots,y_k^{(L)} = \mathbf{PAVA}\left((x_k^{(1)},w_{*1}),\ldots,(x_k^{(L)},w_L^*);\delta[k]\right)$$

Proof of Lemma 4. Choose any $[z^{(1)}, \ldots, z^{(L)}] \in \mathcal{Y}$. By consequence of Lemma 3

$$\mathbf{PAVA}\left((x_k^{(1)}, w_1^*), \ldots, (x_k^{(L)}, w_L^*); \delta[k]\right)$$

$$= \underset{\text{monotone } \lambda_1, \ldots, \lambda_L}{\text{argmin}} \left\{ \sum_{\ell=1}^{L} w_\ell^* \left(x_k^{(\ell)} - \lambda_\ell\right)^2 \right\} \quad \text{where the } \lambda_\ell \text{ are only increasing if specified by } \delta[k]$$

$$\implies \sum_{\ell=1}^{L} w_\ell^* \left(y_k^{(\ell)} - x_k^{(\ell)}\right)^2 \leqslant \sum_{\ell=1}^{L} w_\ell^* \left(z_k^{(\ell)} - x_k^{(\ell)}\right)^2 \quad \forall k$$

$$\text{since } z_k^{(1)}, \ldots, z_k^{(L)} \text{ have monotonicity specified by } \delta$$

$$\implies \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \sum_{\ell=1}^{L} w_\ell^* \left(y_k^{(\ell)} - x_k^{(\ell)}\right)^2 \leqslant \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \sum_{\ell=1}^{L} w_\ell^* \left(z_k^{(\ell)} - x_k^{(\ell)}\right)^2$$

$$\square$$

Lemma 5. *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*

$$x_1^{(\ell)}, \ldots, x_{P-1}^{(\ell)} = \mathbf{PAVA}\left(\left(y_1^{(\ell)}, \frac{p_2 - p_0}{2}\right), \ldots, \left(y_{P-1}^{(\ell)}, \frac{p_P - p_{P-2}}{2}\right); \text{``nondecreasing''}\right)$$

Proof of Lemma 5. Choose any $[z^{(1)}, \ldots, z^{(L)}] \in \mathcal{X}$. By Lemma 3

$$\mathbf{PAVA}\left(\left(y_1^{(\ell)}, \frac{p_2 - p_0}{2}\right), \ldots, \left(y_{P-1}^{(\ell)}, \frac{p_P - p_{P-2}}{2}\right); \text{``nondecreasing''}\right)$$

$$= \underset{\lambda_1 \leqslant \cdots \leqslant \lambda_{P-1}}{\text{argmin}} \left\{ \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \left(y_k^{(\ell)} - \lambda_k\right)^2 \right\} \quad \text{for each } \ell$$

$$\implies \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \left(x_k^{(\ell)} - y_k^{(\ell)}\right)^2 \leqslant \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \left(z_k^{(\ell)} - y_k^{(\ell)}\right)^2 \quad \forall \ell$$

$$\text{since } [z^{(1)}, \ldots, z^{(L)}] \in \mathcal{X} \implies \forall \ell : z_1^{(\ell)} \leqslant \cdots \leqslant z_{P-1}^{(\ell)}$$

$$\implies \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \sum_{\ell=1}^{L} w_\ell^* \left(x_k^{(\ell)} - y_k^{(\ell)}\right)^2 \leqslant \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \sum_{\ell=1}^{L} w_\ell^* \left(x_k^{(\ell)} - z_k^{(\ell)}\right)^2$$

$$\square$$

S8.4. Proof of Theorem 2

Proof. Recalling that $G^{-1}(p)$ denotes the pth quantile of $Q_\ell \equiv f(\ell)$, we also define:

$$\bar{F}_\ell^{-1}(p) := \frac{1}{N_\ell} \sum_{i \in I_\ell} F_i^{-1}(p) \tag{18}$$

By a standard application of the Chernoff bound (Vershynin 2012, Boucheron et al. 2013):

$$\Pr\left(\left|\bar{F}^{-1}(p) - G_\ell^{-1}(p)\right| > \eta\right) = \Pr\left(\left|\frac{1}{N_\ell}\sum_{i\in I_\ell}\mathcal{E}_i(p)\right| > \eta\right) \leqslant 2\exp\left(-\frac{\eta^2 N_\ell}{2\sigma^2}\right) \quad \forall \eta > 0$$

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

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

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

$$\text{by a union-bound}$$

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

$$\leqslant 2PL\exp\left(-\frac{\eta N_\ell}{2\sigma^2 L}\right) \tag{19}$$

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

$$\sum_{\ell=1}^{L}\sum_{i\in I_\ell}d_W\left(F_i^{-1}, \widehat{G}_\ell^{-1}\right)^2 \leqslant \sum_{\ell=1}^{L}\sum_{i\in I_\ell}d_W\left(F_i^{-1}, G_\ell^{-1}\right)^2$$

$$\Rightarrow \sum_{\ell=1}^{L}d_W\left(\bar{F}_\ell^{-1}, \widehat{G}_\ell^{-1}\right)^2 \leqslant \sum_{\ell=1}^{L}d_W\left(\bar{F}_\ell^{-1}, G_\ell^{-1}\right)^2 \quad \text{by Lemma 1}$$

$$\Rightarrow \forall \ell:\ d_W\left(\bar{F}_\ell^{-1}, \widehat{G}_\ell^{-1}\right)^2 \leqslant \sum_{\ell=1}^{L}d_W\left(\bar{F}_\ell^{-1}, G_\ell^{-1}\right)^2$$

Thus, by the triangle-inequality:

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

which implies $\forall \epsilon > 0$ we can combine this result with (19) setting $\eta := \epsilon^2/4$ to get:

$$\Pr\left(\exists \ell: d_W(\widehat{G}_\ell^{-1}, G_\ell^{-1}) > \epsilon\right) \leqslant \Pr\left(\sum_{\ell=1}^{L}d_W\left(\bar{F}_\ell^{-1}, G_\ell^{-1}\right)^2 > \frac{\epsilon^2}{4}\right) \leqslant 2PL\exp\left(-\frac{\epsilon^2 N_\ell}{8\sigma^2 L}\right)$$

$\square$

S8.5. Proof of Theorem 3

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

$$\overline{\widehat{F}}_\ell^{-1}(p) := \frac{1}{N_\ell} \sum_{i \in I_\ell} \widehat{F}_i^{-1}(p) \tag{20}$$

by Theorem 8 and Lemma 1, we have:

$$\sum_{\ell=1}^L d_W \left(\widehat{G}_\ell^{-1}, \overline{\widehat{F}}_\ell^{-1} \right)^2 \leqslant \sum_{\ell=1}^L d_W \left(G_\ell^{-1}, \overline{\widehat{F}}_\ell^{-1} \right)^2$$

$$\Rightarrow \ d_W \left(\widehat{G}_\ell^{-1}, \overline{\widehat{F}}_\ell^{-1} \right)^2 \leqslant \sum_{\ell=1}^L d_W \left(G_\ell^{-1}, \overline{\widehat{F}}_\ell^{-1} \right)^2 \quad \forall \ell$$

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

$$\forall \ell : \ d_W \left(\widehat{G}_\ell^{-1}, G_\ell^{-1} \right) \leqslant d_W \left(\widehat{G}_\ell^{-1}, \overline{\widehat{F}}_\ell^{-1} \right) + d_W \left(\overline{\widehat{F}}_\ell^{-1}, G_\ell^{-1} \right) \quad \text{by the triangle-inequality}$$

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

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

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

Therefore $\forall \epsilon > 0$:

$$\Pr \left(\exists \ell : d_W \left(\widehat{G}_\ell^{-1}, G_\ell^{-1} \right) > \epsilon \right) \leqslant \Pr \left(\sum_{\ell=1}^L d_W (\bar{F}_\ell^{-1}, G_\ell^{-1})^2 + \sum_{\ell=1}^L d_W \left(\overline{\widehat{F}}_\ell^{-1}, \bar{F}_\ell^{-1} \right)^2 > \frac{\epsilon^2}{8} \right)$$

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

and we can use (19) to bound the first summand, resulting in the following bound

$$\Pr \left(\exists \ell : d_W \left(\widehat{G}_\ell^{-1}, G_\ell^{-1} \right) > \epsilon \right) \leqslant 2PL \exp \left(\frac{-\epsilon^2 N_\ell}{32\sigma^2 L} \right) + \Pr \left(\sum_{\ell=1}^L d_W \left(\overline{\widehat{F}}_\ell^{-1}, \bar{F}_\ell^{-1} \right)^2 > \frac{\epsilon^2}{16} \right) \tag{21}$$

Finally, Lemma 7 implies:

$$\Pr\left(\sum_{\ell=1}^{L} d_W\left(\widehat{\bar{F}}_{\ell}^{-1}, \bar{F}_{\ell}^{-1}\right)^2 > \frac{\epsilon^2}{16}\right) \leqslant 2N_\ell PL \exp\left(-2nR\left(\frac{\epsilon}{4\sqrt{L}}\right)^2\right)$$

which produces the desired bound when combined with (21). □

S8.6. Proof of Theorem 4

Proof. By Lemma 8, (A.11) $\Rightarrow$ (A.12), so we only need to show the result assuming (A.12) holds. Lemma 9 then implies:

$$\Pr\left(\sum_{\ell=1}^{L} d_W\left(\widehat{\bar{F}}_{\ell}^{-1}, \bar{F}_{\ell}^{-1}\right)^2 > \frac{\epsilon^2}{16}\right) \leqslant 2P \exp\left(-\frac{c^2}{8}N_\ell n\epsilon^2\right)$$

Note that the bound in (21) only requires the assumptions from Theorem 2, so we can combine it with the above expression to obtain the desired bound. □

S8.7. Proof of Theorem 5

Proof.

$$\begin{aligned}
\text{Consider} \quad &\Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) \\
&= \Pr\left(\widehat{F}_i\left(F_i^{-1}(k/P) + \epsilon\right) \leqslant \frac{k}{P}\right) \\
&= \Pr\left(\sum_{j=1}^{n} \mathbb{1}\left[X_{i,j} \leqslant F_i^{-1}(k/P) + \epsilon\right] \leqslant \frac{nk}{P}\right) \quad (22)
\end{aligned}$$

This is the CDF evaluated at $\widetilde{x} := \frac{nk}{P}$ of a binomial random variable with success probability $\widetilde{p} := F_i\left(F_i^{-1}(k/P) + \epsilon\right)$ in n trials.
Now assume $\epsilon + F_i^{-1}(k/P) \geqslant B > 0$, which implies $n\widetilde{p} \geqslant \widetilde{x}$.
Letting $D(\alpha \,||\, \beta)$ denote the relative entropy between the Bernoulli(α) and Bernoulli(β) distributions, we can thus apply a tail-inequality for the binomial CDF which Arratia &

Gordon (1989) derived from the Chernoff bound to upper-bound (22) by

$$\leqslant \exp\left(-nD\left(\frac{\tilde{x}}{n} \parallel \tilde{p}\right)\right)$$

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

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

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

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

$$\quad \text{where } C(k) := \frac{k}{P}\log\left(\frac{k}{P}\right) + \left(1 - \frac{k}{P}\right)\log\left(1 - \frac{k}{P}\right) \geqslant -1$$

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

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

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

$$\quad \text{because } 1 - k/P > 0 \text{ and by (A.13): } F_i\left(F_i^{-1}(k/P)+\epsilon\right) \geqslant z$$

$$\quad \text{since we've assumed } F_i^{-1}(k/P) + \epsilon \geqslant B$$

$$= e^{-n} \cdot \exp\left(-2an\left(1 - \frac{k}{P}\right)\left(F_i^{-1}(k/P) + \epsilon - B + b\right)^2\right)$$

$$\leqslant e^{-n} \cdot \exp\left(-2an\left(1 - \frac{k}{P}\right)\frac{\min\left\{b^2, \left(B - F_i^{-1}(k/P)\right)^2\right\}}{(B - F_i^{-1}(k/P))^2}\epsilon^2\right)$$

$$\quad \text{because } \epsilon \geqslant B - F_i^{-1}(k/P) \text{ implies}$$

$$\frac{\min\left\{b^2, \left(B - F_i^{-1}(k/P)\right)^2\right\}\epsilon^2}{(B - F_i^{-1}(k/P))^2} \leqslant \left(F_i^{-1}(k/P) + \epsilon - B + b\right)^2$$

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

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

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

$$\leqslant \exp\left(-n\left(\frac{2a\left(1-\frac{k}{P}\right)b^2-1}{4B^2}\right)\epsilon^2\right) \qquad \text{because by (A.13) and (A.15):}$$

$$- F_i^{-1}(k/P) \leqslant B \text{ and } 0 < b \leqslant B$$

And finally, we can use the fact that $k \leqslant P-1$ to obtain the following bound

$$\Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) \leqslant \exp\left(-n\left(\frac{2ab^2-1}{4PB^2}\right)\epsilon^2\right) \tag{23}$$

Following the proof of Lemma 8, one can show that (A.13) implies

$$\Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) \leqslant \exp(-2nc^2\epsilon^2) \quad \text{if } 0 < \epsilon < B - F_i^{-1}(k/P) \tag{24}$$

Combining (24) with (23), we thus have

$$\Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) \leqslant \exp\left(-nr\epsilon^2\right) \quad \forall \epsilon > 0$$

where $r := \min\left\{2c^2, \frac{2ab^2-1}{4PB^2}\right\} > 0$ by (A.14).

One can show by an identical argument that

$$\Pr\left(F_i^{-1}(k/P) - \widehat{F}_i^{-1}(k/P) > \epsilon\right) \leqslant \exp\left(-nr\epsilon^2\right) \quad \forall \epsilon > 0$$

and therefore

$$\Pr\left(\left|\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P)\right| > \epsilon\right) \leqslant 2\exp\left(-nr\epsilon^2\right) \quad \forall \epsilon > 0 \tag{25}$$

$\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P)$ is thus sub-Gaussian with parameter $\frac{1}{2nr}$ and independent of $\widehat{F}_j^{-1}(k/P) - F_j^{-1}(k/P) \ \forall j \neq i$ because we assumed the simple quantile-estimator defined in (A.10) is used. Following the proof of Lemma 9, $\forall \gamma > 0$:

$$\Pr\left(\sum_{\ell=1}^{L} d_W\left(\widehat{\bar{F}}_\ell^{-1}, \bar{F}_\ell^{-1}\right)^2 > \frac{\epsilon^2}{16}\right) \leqslant 2P\exp\left(-\frac{r}{16}N_\ell n\epsilon^2\right) \tag{26}$$

Note that the bound in (21) only requires the assumptions from Theorem 2, so we can combine it with the above inequality to obtain the desired bound. $\qquad\square$

Lemma 6 (Serfling (1980): Theorem 2.3.2). *For $p \in (0,1)$: if $\exists$ unique x s.t. $F(x) = p$ and $\widehat{F}^{-1}(p)$ is estimated using n i.i.d. samples from CDF F_i, then $\forall \gamma > 0$:*

$$\Pr\left(\left| \widehat{F}_i^{-1}(p) - F_i^{-1}(p) \right| > \gamma \right) \leqslant 2\exp\left(-2nR(\gamma, i, p)^2 \right)$$

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 7. *Under the assumptions of Theorem 3 and definitions (11), (18), (20)*

$$\forall \gamma > 0: \quad \Pr\left(\sum_{\ell=1}^{L} d_W\left(\overline{\widehat{F}}_\ell^{-1}, \bar{F}_\ell^{-1} \right)^2 > \gamma \right) \leqslant 2N_\ell PL \exp\left(-2nR\left(\sqrt{\gamma/L} \right)^2 \right)$$

Proof of Lemma 7.

$$\Pr\left(\sum_{\ell=1}^{L} d_W\left(\overline{\widehat{F}}_\ell^{-1}, \bar{F}_\ell^{-1} \right)^2 > \gamma \right)$$

$$= \Pr\left(\sum_{\ell=1}^{L} \frac{1}{N_\ell} \sum_{i \in I_\ell} \sum_{k=1}^{P-1} \frac{1}{P} \left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) \right)^2 > \gamma \right)$$

$$\leqslant N_\ell L \sum_{k=1}^{P-1} \Pr\left(\left| \widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) \right| > \sqrt{\frac{\gamma}{L}} \right) \quad \text{by the union-bound}$$

$$\leqslant 2N_\ell L \sum_{k=1}^{P-1} \exp\left(-2nR\left(\sqrt{\gamma/L}, i, k/P \right)^2 \right) \quad \text{by (A.9) and Lemma 6}$$

$$\leqslant 2N_\ell LP \exp\left(-2nR\left(\sqrt{\gamma/L} \right)^2 \right) \quad \text{by definition (11)}$$

$\qquad\square$

Lemma 8. *If we assume (A.8) and (A.9), then condition (A.11) implies condition (A.12).*

Proof of Lemma 8. Assume WLOG that $F_i^{-1}(k/P) \geqslant 0$ and note that $F_i^{-1}(k/P) \leqslant B$ by (A.11).
Then, by a bound established in the proof of Lemma 6 given in (Serfling 1980), $\forall \epsilon > 0$:

$$\Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon \right) \leqslant \exp\left(-2n\,R(\epsilon, i, k/P)^2 \right) \tag{27}$$

and
$$\Pr\left(F_i^{-1}(k/P) - \widehat{F}_i^{-1}(k/P) > \epsilon\right) \leqslant \exp\left(-2n\,R(\epsilon, i, k/P)^2\right) \tag{28}$$

By (A.11): $f_i(x) = \frac{d}{dx}F_i(x) \geqslant c \;\forall x \in (-B, B)$ which implies

$$R(\gamma, i, p) \geqslant c\gamma > 0 \;\; \text{if } F_i^{-1}(p) \pm \gamma \in (-B, B) \tag{29}$$

because recall that we defined $R(\gamma, i, p) := \min\left\{F_i\left(F_i^{-1}(p) + \gamma\right) - p\;,\;\; p - F_i\left(F_i^{-1}(p) - \gamma\right)\right\}$. Together with (29), (27) and (28) imply

$$\Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) \leqslant \exp(-2nc^2\epsilon^2) \;\; \text{if } F_i^{-1}(k/P) + \epsilon < B \tag{30}$$

and

$$\Pr\left(F_i^{-1}(k/P) - \widehat{F}_i^{-1}(k/P) > \epsilon\right) \leqslant \exp(-2nc^2\epsilon^2) \;\; \text{if } F_i^{-1}(k/P) - \epsilon > -B \tag{31}$$

Note that because $f_i(x) = 0 \;\forall x \geqslant B$, we have

$$\Pr\left(\widehat{F}_i^{-1}(k/P) > F_i^{-1}(k/P) + \epsilon\right) = 0 \;\; \text{if } \epsilon \geqslant B - F_i^{-1}(k/P)$$
$$\implies \Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) = 0 \;\; \text{if } \epsilon \geqslant B - F_i^{-1}(k/P) \tag{32}$$

as well as

$$\Pr\left(\widehat{F}_i^{-1}(k/P) < F_i^{-1}(k/P) - \epsilon\right) = 0 \;\; \text{if } \epsilon \geqslant B + F_i^{-1}(k/P)$$
$$\implies \Pr\left(F_i^{-1}(k/P) - \widehat{F}_i^{-1}(k/P) > \epsilon\right) = 0 \;\; \text{if } \epsilon \geqslant B + F_i^{-1}(k/P) \tag{33}$$

Putting together (30), (31), (32), and (33), we thus have

$$\Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) \leqslant \exp(-2nc^2\epsilon^2) \;\;\;\; \forall \epsilon > 0$$

and
$$\Pr\left(F_i^{-1}(k/P) - \widehat{F}_i^{-1}(k/P) > \epsilon\right) \leqslant \exp(-2nc^2\epsilon^2) \;\;\;\; \forall \epsilon > 0$$

which implies

$$\Pr\left(\left|F_i^{-1}(k/P) - \widehat{F}_i^{-1}(k/P)\right| > \epsilon\right) \leqslant 2\exp(-2nc^2\epsilon^2) \;\;\;\; \forall \epsilon > 0$$

$\square$

Lemma 9. *Under condition (A.12) and definitions (11), (18), (20)*

$$\forall \gamma > 0: \quad \Pr\left(\sum_{\ell=1}^{L} d_W\left(\widehat{\bar{F}}_\ell^{-1}, \bar{F}_\ell^{-1}\right)^2 > \gamma\right) \leqslant 2P\exp\left(-2nc^2 N_\ell \gamma\right)$$

Proof of Lemma 9.

$$\Pr\left(\sum_{\ell=1}^{L} d_W\left(\overline{\widehat{F}}_\ell^{-1}, \bar{F}_\ell^{-1}\right)^2 > \gamma\right)$$

$$= \Pr\left(\frac{1}{LN_\ell}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\sum_{k=1}^{P-1}\frac{1}{P}\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P)\right)^2 > \frac{\gamma}{L}\right)$$

$$\leqslant \sum_{k=1}^{P-1}\Pr\left(\left|\frac{1}{LN_\ell}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P)\right| > \sqrt{\frac{\gamma}{L}}\right) \quad \text{by the union-bound}$$

$$\leqslant 2\sum_{k=1}^{P-1}\exp\left(-2nc^2 LN_\ell\sqrt{\frac{\gamma}{L}}^2\right) = 2P\exp\left(-2nc^2 N_\ell\gamma\right)$$

where in the last inequality, we have used the fact that (A.12) implies the $\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P)$ are independent sub-Gaussian random variables with parameter $\frac{1}{4nc^2}$, so the inequality follows from a standard application of the Chernoff bound (Vershynin 2012, Boucheron et al. 2013). $\qquad\square$

Additional References for the Supplementary Material

Altman, N. & Leger, C. (1995), 'Bandwidth selection for kernel distribution function estimation', *J. Stat. Plan. Inference* **46**, 195–214.

Apostolova, M. D., Ivanova, I. A. & Cherian, M. (1999), 'Metallothionein and Apoptosis during Differentiation of Myoblasts to Myotubes: Protection against Free Radical Toxicity', *Toxicology and Applied Pharmacology* **159**(3), 175–184.

Arratia, R. & Gordon, L. (1989), 'Tutorial on large deviations for the binomial distribution', *Bulletin of Mathematical Biology* **51**(1), 125–131.

Bondell, H. D., Reich, B. J. & Wang, H. (2010), 'Non-crossing quantile regression curve estimation', *Biometrika* **97**(4), 825–838.

Boucheron, S., Lugosi, G. & Massart, P. (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press.

Boyle, J. & Dykstra, R. (1986), 'A Method for Finding Projections onto the Intersection of Convex Sets in Hilbert Spaces', *Lecture Notes in Statistics* **37**, 28–47.

Combettes, P. L. & Pesquet, J. C. (2011), Proximal splitting methods in signal processing, *in* H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke & H. Wolkowicz, eds, 'Fixed-Point Algorithms for Inverse Problems in Science and Engineering', Springer, New York, pp. 185–212.

de Leeuw, J. (1977), 'Correctness of Kruskal's algorithms for monotone regression with ties', *Psychometrika* **42**(1), 141–144.

Gilchrist, W. (2000), *Statistical Modelling with Quantile Functions*, Taylor & Francis.

Han, X. H., Jin, Y.-R., Seto, M. & Yoon, J. K. (2011), 'A WNT/β-Catenin Signaling Activator, R-spondin, Plays Positive Regulatory Roles during Skeletal Myogenesis', *Journal of Biological Chemistry* **286**(12), 10649–10659.

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H. & Herwig, R. (2011), 'ConsensusPathDB: toward a more complete picture of cell biology.', *Nucleic acids*

research **39**, D712–7.

Kharchenko, P. V., Silberstein, L. & Scadden, D. T. (2014), 'Bayesian approach to single-cell differential expression analysis', *Nat. Meth* **11**(7), 740–742.

Krishnaswamy, S., Spitzer, M. H., Mingueneau, M., Bendall, S. C., Litvin, O., Stone, E., Pe'er, D. & Nolan, G. P. (2014), 'Conditional density-based analysis of T cell signaling in single-cell data', *Science* **346**(6213).

Myers, S. A., Nield, A., Chew, G.-S. & Myers, M. A. (2013), 'The Zinc Transporter, Slc39a7 (Zip7) Is Implicated in Glycaemic Control in Skeletal Muscle Cells', *PLoS ONE* **8**(11), e79316.

Petschnik, A. E., Fell, B., Kruse, C. & Danner, S. (2010), 'The role of alpha-smooth muscle actin in myogenic differentiation of human glandular stem cells and their potential for smooth muscle cell replacement therapies.', *Expert opinion on biological therapy* **10**(6), 853–861.

Phipson, B. & Smyth, G. K. (2010), 'Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn.', *Statistical applications in genetics and molecular biology* **9**, Article39.

Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. (2014), 'Normalization of RNA-seq data using factor analysis of control genes or samples', *Nature Biotechnology* **32**(9), 896–902.

Ruggle, S., Alexander, J. T., Genadek, K., Goeken, R. & Matthew B. Schroeder, M. S. (2010), 'Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]'.

Sanchez, D., Ganfornina, M. D. & Martinez, S. (2002), 'Expression pattern of the lipocalin apolipoprotein D during mouse embryogenesis.', *Mechanisms of development* **110**(1-2), 225–229.

Sebastian, S., Faralli, H., Yao, Z., Rakopoulos, P., Palii, C., Cao, Y., Singh, K., Liu, Q.-C., Chu, A., Aziz, A., Brand, M., Tapscott, S. J. & Dilworth, F. J. (2013), 'Tissue-specific splicing of a ubiquitously expressed transcription factor is essential for muscle

differentiation.', *Genes & development* **27**(11), 1247–1259.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley and Sons, New York.

Shaked, M. & Shanthikumar G., J. (1994), *Stochastic orders and their applications*, Academic Press, Boston.

Takeuchi, I., Le, Q. V., Sears, T. D. & Alexander J. Smola (2006), 'Nonparametric Quantile Estimation', *Journal of Machine Learning Research* **7**, 1231–1264.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. & Rinn, J. L. (2014), 'The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells', *Nat. Biotechnol* **32**(4), 381–386.

Trimarco, A., Forese, M. G., Alfieri, V., Lucente, A., Brambilla, P., Dina, G., Pieragostino, D., Sacchetta, P., Urade, Y., Boizet-Bonhoure, B., Boneschi, F. M., Quattrini, A. & Taveggia, C. (2014), 'Prostaglandin D2 synthase/GPR44: a signaling axis in PNS myelination', *Nature Neuroscience* **17**(12), 1682–1692.

Tsai, C.-A. & Chen, J. J. (2007), 'Kernel estimation for adjusted -values in multiple testing', *Computational Statistics & Data Analysis* **51**(8), 3885–3897.

Vershynin, R. (2012), Introduction to the non-asymptotic analysis of random matrices, *in* 'Compressed Sensing', Cambridge University Press, Cambridge, pp. 210–268.

Wolfstetter, E. (1993), *Stochastic dominance: theory and applications*, Wirtschaftswiss. Fak., Humboldt-Universitat.

Zeisel, A., Munoz-Manchado, A. B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J. & Linnarsson, S. (2015), 'Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.', *Science* **347**(6226), 1138–42.