Flexible models for understanding and optimizing complex populations

by

Jonas W. Mueller

Submitted to the

Department of Electrical Engineering and Computer Science

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Signature redacted

Author .....

Department of Electrical Engineering and Computer Science

May 23, 2018

Signature redacted

Certified by .....

Tommi S. Jaakkola

Professor of Electrical Engineering and Computer Science

Thesis Supervisor

Signature redacted

Certified by .....

David K. Gifford

Professor of Electrical Engineering and Computer Science

Thesis Supervisor

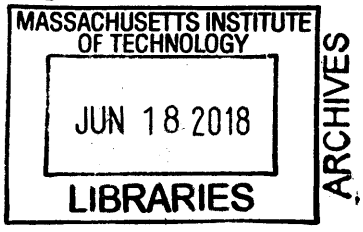
Signature redacted

Accepted by .....

Leslie A. Kolodziejski

Professor of Electrical Engineering and Computer Science

Chair, Department Committee on Graduate Students



Flexible models for understanding and optimizing complex populations

by

Jonas W. Mueller

Submitted to the Department of Electrical Engineering and Computer Science
on May 23, 2018, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

Data analysis is often driven by the goals of understanding or optimizing some population of interest. The first of these two objectives aims to produce insights regarding characteristics of the underlying population, often to facilitate scientific understanding. Crucially, this requires models which produce results that are highly interpretable to the analyst. On the other hand, notions of interpretability are not necessarily as central for determining how to optimize populations, where the aim is to build data-driven systems which learn how to act upon individuals in a manner that maximally improves certain outcomes of interest across the population. In this thesis, we develop interpretable yet flexible modeling frameworks for addressing the former goal, as well as black-box nonparametric methods for addressing the latter. Throughout, we demonstrate various empirical applications of our algorithms, primarily in the biological context of modeling gene expression in large cell populations.

For better understanding populations, we introduce two nonparametric models that can accurately reflect interesting characteristics of complex distributions without reliance on restrictive assumptions, while simultaneously remaining highly interpretable through their use of the Wasserstein (optimal transport) metric to summarize changes over an entire population. One approach is *principal differences analysis*, a projection-based technique that interpretably characterizes differences between two arbitrary high-dimensional probability distributions. Another approach is the *TRENDS* model, which quantifies the underlying effects of temporal progression in an evolving sequence of distributions that also vary due to confounding noise. While the aforementioned techniques fall under the frequentist regime, we subsequently present a Bayesian framework for the task of optimizing populations. Drawing upon the Gaussian process toolkit, our method learns how to best conservatively intervene upon heterogeneous populations in settings with limited data and substantial uncertainty about the underlying relationship between actions and outcomes.

Thesis Supervisor: Tommi Jaakkola
Professor of EECS

Thesis Supervisor: David Gifford
Professor of EECS

Acknowledgments

This thesis and my PhD work are a culmination of the intellectual contributions, mentorship, and inspiration I received from many people close to me.

I am extremely fortunate to have had the guidance of two incredible advisors: Tommi Jaakkola and David Gifford, who both immensely shaped my research and overall scientific outlook. Encouraging high levels of rigor in all aspects of my work, Tommi repeatedly awed me with his deep insights on the numerous technical obstacles I encountered. Dave has always motivated me to pursue projects with massive real-world impact, pushing my research efforts to a grander scale and emboldening me with an extremely ambitious vision of how machine learning should drive biological advancement. Together, these two have guided me through a wide range of scientific problems and avidly fostered my development as a researcher.

I am also deeply grateful my third thesis committee member, Edoardo Airoldi, who contributed a great deal to my statistics education, and provided valuable career advice and feedback on my thesis and other research projects. Most of my work would not have been possible without a number of other people who I thank for playing instrumental roles throughout my PhD: Vasilis Syrgkanis & Matt Taddy for their insightful mentorship across multiple fields spanning online optimization and economics, Nadav Sharon & Douglas Melton for introducing me to the innovative world of single-cell genomics, Regina Barzilay & Suvrit Sra for teaching me how to teach, Tamara Broderick & Stephanie Jegelka for serving on my RQE committee, Tejas Kulkarni & Andrea Tacchetti for our joint organization of the MIT Machine Learning Tea seminars, Jeanne Darling & Linda Lynch & Teresa Cataldo for their vigilance in ensuring my needs as a graduate student, and Richard Sherwood who always readily offered his wealth of biological knowledge.

In addition to having two incredible advisors, I have been fortunate to be a part of their two wonderful lab groups, among which I found many great friends and scientific collaborators. From hunting for free food, Mondays at the Muddy, engaging in Coup subterfuge, ski trips, and heated neural net vs. kernels debates, hanging around the lab has always been a joy, and occasionally even productive as well. I am particularly thankful for my inspiring discussions with Charlie O'Donnell, Tatsu Hashimoto, Tahin Syed, Matt Edwards, Haoyang Zeng, Nisha Rajagopal, Yuchun Guo, Daniel Kang, Logan Engstrom, Saber Liu, Grace Yeo, Max Shen, Konstantin Krismer, Jen Hammelman, Brandon Carter, Sid Jain, Jean Honorio, Yu Xin, Andreea Gane, Paresh Malalur, David Reshef, David Alvarez-Mellis, Vikas Garg, Tim Plump, and George Du.

Finally, I am eternally thankful to my parents and Laura for the endless love and support as well as the countless times they have gracefully beared with my odd hours and neuroses. You are my greatest sources of happiness and inspiration.

Contents

List of Figures

List of Tables

Nomenclature

For ease of reference, we present a list of particular notation and technical definitions used in each chapter. Hat notation $\widehat{z}$ is used throughout the thesis to represent an empirical realization of underlying property z which is observed/estimated from data. C is used to represent universal constants, whose value may change from line to line.

Chapter 1

P_X denotes the probability distribution of random variable X.

The L_p norm of a vector $x \in \mathbb{R}^d$ is given by: $||x||_p = (|x_1|^p + |x_2|^p + \ldots |x_d|^p)^{1/p}$.

The L_p norm of a measurable function $f : \mathcal{D} \to \mathbb{R}$ is given by: $||f||_p = \left(\int_{\mathcal{D}} |f(x)|^p \, \mathrm{d}x \right)^{1/p}$.

For matrices A, B of equal dimensions: $\langle A, B \rangle$ is defined as the inner product of the vectors formed by flattening the matrices, and is equivalent to the trace of their matrix product: $\mathrm{tr}(AB)$. This is a linear function of both A and B.

$\mathcal{M}$ is the set of all $n \times m$ nonnegative matching matrices with fixed row sums $= 1/n$ and column sums $= 1/m$.

Chapter 2

A^C denotes the complement of set A.

$|| \cdot ||_0$ denotes the cardinality function (the number of nonzero entries) of a vector/matrix.

$|| \cdot ||_1$ denotes the ℓ_1 norm of a vector/matrix (the sum of the magnitudes of the

entries).

$||\cdot||_2$ and $||\cdot||_F$ denote the Euclidean and Frobenius norm of a vector and matrix, respectively. Throughout chapter 2, we interchangeably apply operators to both vectors and matrices assuming the context is clear.

$\text{tr}(A)$ denotes the *trace* of matrix A, i.e. the sum of the diagonal entries.

The overloaded $\text{diag}(\cdot)$ operator returns either the vector formed by the diagonal elements of the input matrix, or alternatively the diagonal matrix whose only nonzero elements are the entries of the input vector.

$D(\cdot,\cdot)$ is taken to be the squared L_2 Wasserstein distance between distributions from §2.3 onward.

$\mathcal{B} := \{\beta \in \mathbb{R}^d : ||\beta||_2 \leq 1, \beta_1 \geq 0\}$ is the feasible set of unit-length projection vectors considered in PDA.

Chapter 3

F_X denotes the cumulative distribution function of univariate probability measure P_X, and F_X^{-1} denotes the corresponding quantile function.

We slightly abuse notation using $d_{L_q}(\cdot,\cdot)$ to denote both L_q-Wasserstein distances between distributions or the corresponding quantile functions' L_q-distance (both $q = 1, 2$ are used in this chapter).

Random variable $X \in \mathbb{R}$ is said to follow a sub-Gaussian(σ) distribution if $\mathbb{E}[X] = 0$ and $\Pr(|X| > t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$ for any $t > 0$.

The Wasserstein R^2 measures how much of the variation in the observed distributions is captured by the TRENDS model's fitted distributions.

Δ measures the magnitude of the inferred trend-effect (i.e. the effect size).

Chapter 4

All points $x \in \mathbb{R}^d$ lie in convex and compact domain $\mathcal{C} \subset \mathbb{R}^d$.

All occurrences of f are implicitly referring to $f \mid \mathcal{D}_n$.

$\mu_n(\cdot)$, $\sigma_n^2(\cdot)$, and $\sigma_n(\cdot,\cdot)$ respectively denote the mean, variance, and covariance function of our posterior for $f \mid \mathcal{D}_n$ under the $\mathrm{GP}\big(0, k(x,x')\big)$ prior.

$F_Z^{-1}(\alpha)$ denotes the α^{th} quantile of random variable Z.

$\Phi^{-1}(\cdot)$ denotes the $N(0,1)$ quantile function.

$||\cdot||_k$ denotes the norm of reproducing kernel Hilbert space $\mathcal{H}_k$.

$\mathcal{B}_\delta(x) \subset \mathbb{R}^d$ denotes the ball of radius δ centered at $x \in \mathcal{C}$.

$\mathcal{I} \subseteq \{1,\ldots,d\}$ represents the set of variables which are intervened upon in sparse settings.

$\mathrm{pa}(Y)$ denotes the set of variables which are parents of Y in a causal *directed acyclic graph* (DAG) (Pearl 2000)

$\mathrm{desc}(\mathcal{I})$ is the set of variables which are descendants of at least one variable in $\mathcal{I}$ according to the causal DAG.

The *squared exponential* kernel (with length-scale parameter $l > 0$) is defined:

$$k(x,x') = \exp\Big(-\frac{1}{2l^2}||x-x'||^2\Big)$$

The *Matérn* kernel (with another parameter $\nu > 0$ controlling smoothness of sample paths) is defined:

$$k(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)} r^\nu B_\nu(r) \quad \text{where } r = \frac{\sqrt{2\nu}}{l}||x-x'||, B_\nu \text{ is a modified Bessel function}$$

Random variables $\varepsilon^{(1)},\ldots,\varepsilon^{(n)}$ form a *martingale difference sequence* which is *uniformly bounded* by σ if $\mathbb{E}[\varepsilon^{(i)} \mid \varepsilon^{(i-1)},\ldots,\varepsilon^{(1)}] = 0$ and $\varepsilon^{(i)} \leq \sigma \;\; \forall i \in \mathbb{N}$.

A function f is *Lipshitz continuous* with constant L if: $|f(x) - f(x')| \leq L|x-x'|$ for every $x, x' \in \mathcal{C}$.

Suppose $\rho > 0$ is expressed as $\rho = m + \eta$ for nonnegative integer m and $0 < \eta \leq 1$. The *Hölder space* $C^\rho[0,1]^d$ is the space of functions with existing partial derivatives of orders $(k_1,\ldots,k_d)$ for all integers $k_1,\ldots,k_d \geq 0$ satisfying $k_1 + \cdots + k_d \leq m$. Additionally, each function's highest order partial derivative must form a function h that satisfies: $|h(x) - h(y)| \leq C|x-y|^\eta$ for any x, y.

Chapter 1

Introduction

Across nearly all sciences and industries, the collection and analysis of data has become the primary instrument by which we learn about our world and how to enact desired changes. Technological advancements in computing and measurement techniques (e.g. higher-resolution sensors, internet/mobile applications, high-throughput experimentation) have lead to an explosion in the amount of available data. To fully realize the immense potential impact buried within rapidly accumulating data, we require proper analytic methods designed to best utilize newly emerging data types.

One fundamental goal of scientific data analysis is to better understand the underlying characteristics of a particular population. Classical models for interpretable data analysis often rely on overly simple parametric distributions which are especially inaccurate for high-dimensional measurements, or produce results that must be understood with respect to the "average" individual in the population. However, this hypothetical average-individual may not actually exist in a heterogeneous population and inferred average-effects can be misleading[1]. In contrast, black-box methods from machine learning and nonparametrics can approximate arbitrary relationships, but do not provide much insight about the underlying population. To address these shortcomings, this thesis will introduce nonparametric models that can accurately reflect interesting characteristics of complex distributions, while simultaneously remaining highly interpretable to data analysts. Our key idea is to measure changes across an entire population via the Wasserstein (optimal transport) metric, instead of relying on traditional coarser summaries like expectation/covariance differences.

In other application domains, such as business and medicine, one is more concerned with acting to optimize the underlying population rather than simply understanding its characteristics. Here, data is used to infer an action policy that will lead to desired outcomes. While advances in reinforcement learning (RL) and ban-

[1]Consider for example the famous case of Simpson's paradox, in which a particular average-case relationship appears in the aggregate data from different subpopulations but disappears or reverses when these subpopulations are combined into a single larger dataset.

dit/Bayesian optimization have shown great promise (Shahriari et al. 2016), these sequential methods typically require a vast number of experimental rounds before they begin to reliably identify good actions. Such reliance on a multitude of sequential experiments has limited the applicability of these methods primarily to digital environments where one can rapidly iterate between modeling and experimentation.

It is often more logistically feasible in real-world applications to obtain a single fixed dataset with many observations (e.g. from an external agency or scientific collaborator), upon which all subsequent decision-making will be based. Although more widely applicable, learning from a fixed (and often observational) dataset will inherently involve substantial uncertainty due to the limited number of samples, and it is undesirable to prescribe actions whose outcomes are unclear. In particular, to prove the value of sophisticated statistical learning methods to an external party (for example to encourage the creation of a sequential experimentation infrastructure), it is critical that the analyst provides a data-driven policy that does not produce harmful outcomes at the outset. This unfortunately cannot be ensured under typical exploration strategies employed in RL and bandit/Bayesian optimization. Even in settings where sequential experimentation is feasible, many applications require learning to halt at the time of deployment (consider an autonomous vehicle which is deployed with a fixed policy to ensure reliability). Because complete exploration of complex real-world environments is likely not achieved by the time of deployment, defining the optimal policy will still require dealing with substantial uncertainty, which can be quantified by studying the agent's entire past experience as a fixed dataset.

The final portion of this thesis considers such fixed-data settings from a Bayesian perspective and formalizes the of role of uncertainty in data-driven actions. Importantly, we consider the case where the data do not even contain examples of beneficial actions, and thus these must be identified without explicit supervision (unlike say imitation learning). Adopting a Gaussian process framework, we introduce a conservative definition of the optimal intervention which can be either tailored on an individual basis or globally enacted over a population. Gradient methods are employed to identify the best intervention and a key theme of the approach is carefully constraining this optimization to avoid regions of high uncertainty. Various applications of this methodology are presented including gene expression manipulation and improving the popularity of news articles.

1.1 Single cell RNA-sequencing

Much of the methodology presented in thesis is motivated by scientific applications in single-cell genomics, where the goal is to model a heterogeneous cell population in which key biological processes of interest take place within individual members. In particular, we analyze data containing gene-expression measurements obtained via single-cell RNA-sequencing (scRNA-seq). The recent introduction of RNA-sequencing

techniques to obtain transcriptome-wide gene expression profiles from within individual cells has drawn massive interest across the field of biology, as described by 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. For a detailed survey of the technical steps involved in extracting and quantifying RNA molecules from within single cells, we refer the reader to the recent publication of Haque et al. (2017).

To highlight the importance of making expression measurements within single cells, Geiler-Samerotte et al. (2013) articulate the following analogy: "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". A key limitation for discovery is that existing statistical methods are primarily designed to operate on crude summary statistics such as expectations and covariances. However, such coarse analysis fails to leverage the finer grained insight into biological processes that scRNA-seq measurements can provide, since one could study these same quantities via (aggregate) tissue-level RNA-seq data. Futhermore, statistical results which are interpreted in terms of some hypothetical "average" cell may be severely misleading. 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). Thus, in order to fully characterize biological processes, it is crucial to study the full distribution of gene expression in the underlying cell population. As the number of cells per experiment is increased by advances in single-cell profiling technology such as Drop-seq (Macosko et al. 2015) or inDrops (Zilionis et al. 2017), the study of population-wide cellular gene expression distributions will enable many future discoveries, and the tools introduced in this thesis provide principled and effective statistical frameworks for this analysis.

1.2 The Wasserstein distance

Many of our ideas for modeling populations leverage the Wasserstein distance, a metric between probability distributions that themselves are defined over a common metric space (Dobrushin 1970, Villani 2008). While other popular measures of difference between distributions such as f-divergences (e.g. Kullback-Leibler), total variation, or Bhattacharyya distances only take into account changes in probability measure, the Wasserstein distance (and the broader class of *integral probability metrics* to which it belongs) additionally considers the magnitude of changes, measuring both the amount

of probability mass moved as well as the distance this mass is moved. Since it was first introduced a statistical context by Mallows (1972), many special cases of this statistical divergence have appeared under a variety of names including the Kantorovich, Mallows, Dudley, optimal transport, or earth-mover distance (Levina & Bickel 2001).

For random variables X, Y defined on a common metric space with metric $d(\cdot, \cdot)$, the Wasserstein distance between their corresponding distributions P_X, P_Y is formally defined as the solution of the following optimization over all joint distributions P_{XY} with marginals that match the given distributions:

$$D(X,Y) = \min_{P_{XY}} \mathbb{E}_{P_{XY}} \left[d(X,Y) \right] \quad \text{s.t.} \quad (X,Y) \sim P_{XY}, \; X \sim P_X, \; Y \sim P_Y \tag{1.1}$$

Intuitively interpreted as the minimal amount of "work" that must be done to transform one distribution into the other, the Wasserstein distance has enjoyed enormous success across a number of application domains where differences between distributions should be measured along a meaningful scale (Levina & Bickel 2001). For vector-valued variables, the underlying metric $d(\cdot, \cdot)$ is commonly taken to be an L_p norm (typically either the Euclidean – L_2 – or Manhattan – L_1 – distance). This metric offers a natural dissimilarity measure between populations because it accounts for the proportion of individuals that are different as well as *how* different these individuals are (i.e. it integrates the amount of probability mass moved times the distance moved). Lemma 1 formalizes this idea in a single dimension, showing that the Wasserstein distance can in this case be expressed as a distance between quantile functions.

Lemma 1 (Levina & Bickel (2001)). *Suppose $X, Y \in \mathbb{R}$ are continuous univariate variables in a metric space where distances are measured via a L_p norm. Then, Wasserstein distance between them is given by:*

$$D(X,Y) = \left(\int_0^1 \left| F_X^{-1}(q) - F_Y^{-1}(q) \right|^p \, \mathrm{d}q \right)^{1/p}$$

where F_X, F_Y are the CDFs of distributions P_X, P_Y and F_X^{-1}, F_Y^{-1} are the corresponding quantile functions.

An empirical estimate of the Wasserstein distance from data $x^{(1)}, \ldots, x^{(n)} \overset{IID}{\sim} P_X$ and $y^{(1)}, \ldots, y^{(m)} \overset{IID}{\sim} P_Y$ (assuming $m \leq n$ without loss of generality) is typically obtained by solving the following optimal transport problem:

$$\widehat{D}(X,Y) = \min_{M \in \mathcal{M}} \langle M, K \rangle \tag{1.2}$$

where $\mathcal{M}$ is the polytope of all $n \times m$ nonnegative *matching* matrices with fixed row sums $= 1/n$ and column sums $= 1/m$ (see Villani (2008) for details), and K is a $n \times m$ *transportation cost* matrix with entries $K_{ij} = d(x^{(i)}, y^{(j)})$.

The linear program in (1.2) is often solved using network simplex or interior point methods (Pele & Werman 2009), but it remains quite computationally intensive, requiring at least $O(n^3 \log n)$ runtime. While far more efficient approximation algorithms have been introduced by Cuturi (2013), we note that the computational complexity of this optimal transport problem remains prohibitive for many practical applications. In contrast, estimation of the Wasserstein distance from univariate data can be done far more efficiently by way of the equivalent formulation in Lemma 1, where quantile function estimates are easily obtained by sorting the data (merely requiring $O(n \log n)$ runtime). This property is heavily exploited by our methods in Chapters 2 and 3. We finally emphasize that when using an L_p-Wasserstein distance in data analysis, it is critical to first ensure that the measurements of different variables have been represented on a directly comparable scale.

1.3 Contrasting multivariate distributions

Characterizing differences between populations is one of the most fundamental tasks encountered across the sciences. Such analysis seeks to answer whether or not the populations differ and, if so, which variables or relationships contribute most to this difference. In many applications, information is lacking about the nature of possible differences and the distribution of measurements in the underlying populations are high-dimensional quite complex. The first novel method introduced in this thesis is *principal differences analysis* (PDA) for analyzing differences between high-dimensional distributions. This approach not only produces a p-value for an empirically observed difference, but also interpretably quantifies how much each variable contributes to the overall difference. PDA operates by finding the projection that maximizes the Wasserstein divergence between the resulting univariate populations. Representing the first practical algorithm derived from the Cramer-Wold theory, our approach requires no assumptions about the form of the underlying distributions, nor the nature of their inter-class differences. A sparse variant of the method is introduced to identify features responsible for the differences. We provide algorithms for both the original minimax formulation as well as a convex semidefinite relaxation.

In addition to deriving some convergence results, we illustrate how the approach may be applied to identify differences between cell populations in the somatosensory cortex and hippocampus as manifested by single cell RNA-seq Zeisel et al. (2015). We find that PDA exhibits high power in two-sample testing (even in high-dimensions with sparse underlying differences), and is the only provably sparsistent variable-selection method that does not rely on strong assumptions like those required for the logistic lasso or sparse discriminants analysis. When applied to heavily normalized gene expression data from cells sampled in two different brain regions, our PDA method identified numerous interesting genes involved in regulatory interactions. These could not be found via de-facto differential expression analyses that consider each gene marginally. While Chapter 2 introduces PDA in the context of

the Wasserstein distance, our broader framework extends beyond this specific choice of statistical divergence.

1.4 Quantifying trends in evolving populations

Chapter 3 subsequently presents 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. 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 thus 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 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.

In many scientific and survey settings, real-valued observations are sampled in batches, where the observations in each batch share a common label. This numerical/ordinal value is the covariate. The primary interest in such analyses is to assess the affect of the covariate (measured across a batch) on other measurements (measured within individuals). When each batch consists of a large number of i.i.d. observations, the empirical distribution of individual observation-values in a batch may be a good approximation of the underlying population distribution conditioned on the value of the covariate. In order to quantify the covariate's overall effect on these conditional distributions, we can consider changes across all segments of the population. In the case of high-dimensional observations, one can measure this effect separately for each profiled 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. Referred to as batch effects in the scientific literature (Risso et al. 2014), this type of variation can cause standard methods to overestimate the underlying effects of interest.

Our TRENDS (Temporally Regulated Effects on Distribution Sequences) regression model is designed to infer the magnitude of 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 restriction that fitted-values are smooth/linear in the covariate is replaced by the restriction that fitted distributions follow a *trend*. Here, we formally define the concept of a trend as a sequence of distri-

butions that evolve linearly under the Wasserstein metric. TRENDS extends scalar-valued regression to full distributions while retaining the ability to distinguish effects of interest from extraneous noise. Inspired by applications in single-cell genomics, where longitudinal measurements of individual cells are unobtainable, TRENDS redefines the objectives and measures of fit/effect-size employed in classical regression (such as least-squares and R^2) by using the language of optimal transport to interpretably adapt regression measures for distributions in place of scalar quantities. The corresponding estimation algorithm we propose combines Dykstra's alternating projections with the pool-adjacent-violators technique (de Leeuw 1977, Boyle & Dykstra 1986), and is practically efficient and guaranteed to find a global optimum.

One exciting avenue of experimentation which has become possible with the advent of single-cell RNA-sequencing technology involves profiling groups of cells sampled at various times from tissues / cell-cultures undergoing development. The hope is that such data could reveal which developmental genes regulate/mark the emergence of new cell types over the course of development. However, current scRNA-seq cost/labor constraints prevent dense sampling of cells continuously across the entire time-continuum. Researchers must instead focus on a few time-points, simultaneously isolating batches of cells at each time and subsequently generating RNA-seq transcriptome profiles for each individual cell that has been sampled. 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. Rather than treating the cells from a single time point identically, it is desirable to retain batch information and 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. When applied to scRNA-seq time course datasets from differentiating myoblast cells (Trapnell et al. 2014) as well as the developing somatosensory cortex of juvenile mice (Zeisel et al. 2015), TRENDS is able to accurately uncover the genes that regulate development by disentangling the temporal effects on expression variation from batch effects.

1.5 Bayesian inference of optimal interventions

In Chapter 4, we turn our attention from simply understanding populations to actively optimizing them through external intervention. We introduce a nonparametric Bayesian framework for determining how to best intervene upon heterogeneous populations in order to maximally improve individual outcomes. This work aims to discover narrowly focused interventions (impacting few covariates) that may be individually tailored or globally enacted over the entire population, in order to shape the population as desired. The conservative definition of the optimal intervention that we propose is particularly designed for applications where proposing a harmful action

is drastically worse than proposing no change at all.

Our methodology adapts principles from Bayesian optimization (and bandits/RL) for settings where learning is restricted to a single dataset rather than sequential experimentation. Under limited data, vast regions of the feature-space are inevitably associated with large outcome-uncertainty, and a primary goal of this research is to identify interventions that are beneficial not only in expectation, but also with high confidence. The approach we introduce never suggests wild transformations to covariate-values never before encountered in the training data, and will conservatively suggest no change at all for an individual whose measurements look very different from all previously observed data-points. While our optimal-intervention goals appear similar to Bayesian optimization (Shahriari et al. 2016), additional data is not acquired in our setting. Acquisition functions tailored for exploration of uncertain areas are therefore not appropriate. Furthermore, we seek interventions that lead to the greatest improvement in each individual's outcome rather than finding a single answer (the maximum across the population). For example, in a writing improvement application, we wish to inform a particular author of simple modifications likely to improve their existing article rather than proposing a single optimal article. Similarly, Bayesian optimization does not address optimization of a given population for which the underlying distribution of covariate-values is unknown.

In many data-driven applications, including medicine, the primary interest in causality has to do with identifying actions that are likely to produce a desired change in some outcome of interest. Typically, this is done by analyzing data using models which facilitate understanding of the relations between variables (eg. assuming linearity/additivity). Based on conclusions drawn from this analysis, the analyst decides how to intervene in a manner they confidently believe will improve outcomes. Formalizing such beliefs via Bayesian inference, our framework instead automatically identifies beneficial interventions directly from the data.

In a general setting, optimal intervention requires understanding both the statistical relationship between covariates and outcomes as well as the underlying causal structure. While existing methods for causal inference aim to learn both of these, they remain limited to large sample sizes and few dimensions. By restricting ourselves to applications that meet a set stringent causal conditions, we explore an alternative paradigm to improve outcomes that dispenses with causal modeling, instead treating the underlying mechanisms as a black-box function to be optimized. If the underlying relationship obeys an invariance condition, our approach can identify beneficial interventions directly from observational data. We provide theoretical guarantees for gains obtained via our approach when Gaussian process regression modeling is used to provide posterior estimates of the underlying relationship between covariates and outcomes (Rasmussen 2006). Although our methods assume covariates can be precisely adjusted, they remain capable of improving outcomes in misspecified settings where interventions incur unintentional downstream effects (meaning they affect additional covariates beyond those intended to be altered).

We demonstrate two applications of this methodology. One is a writing improvement task where the data consists of documents labeled with associated outcomes (eg. grades, impact, popularity) and the goal is to suggest beneficial changes to the author. Our second example is a gene perturbation task where the expression of certain regulatory genes can be up/down-regulated in a population (eg. cells or yeast) with the goal of inducing a particular phenotype or downstream gene expression pattern. These examples depict settings where features cause outcomes (not vice-versa) and the assumptions of our approach may hold approximately, depending on the types of external intervention used to actually alter the features.

Although our methods for learning beneficial interventions from observational datasets rely on stringent causal and statistical assumptions, they empirically perform well in both intentionally-misspecified and complex real-world settings. As supervised learning algorithms grow ever more popular, we expect intervention-decisions in many domains will increasingly rely on predictive models. Our conservative definition of the optimal intervention provides a principled approach to handle the inherent uncertainty in these settings as a consequence of limited data. Because we are able to employ any Bayesian regressor (including Gaussian processes and Bayesian neural networks), our ideas are widely applicable, considering practical types of interventions that can either be individually personalized or enacted uniformly over a population.

1.6 Previously published work

This thesis contains material from various previous publications with collaborators who were instrumental in developing many of the results presented here. The material of Chapters 2, 3, and 4, has respectively appeared in the following following publications:

J. Mueller and T. Jaakkola. Principal Differences Analysis: Interpretable Characterization of Differences between Distributions. *Advances in Neural Information Processing Systems*, 2015.

J. Mueller, T. Jaakkola, and D. Gifford. Modeling Persistent Trends in Distributions. *Journal of the American Statistical Association*, 2017.

J. Mueller, D. Reshef, G. Du, and T. Jaakkola. Learning Optimal Interventions. *Artificial Intelligence and Statistics*, 2017.

Chapter 2

Principal differences analysis

Understanding differences between populations is a common task across disciplines, from biomedical data analysis to demographic or textual analysis. For example, in biomedical analysis, a set of variables (features) such as genes may be profiled under different conditions (e.g. cell types, disease variants), resulting in two or more populations to compare. The hope of this analysis is to answer whether or not the populations differ and, if so, which variables or relationships contribute most to this difference. In many cases of interest, the comparison may be challenging primarily for three reasons:

1. The number of variables profiled may be large.

2. Populations are represented by finite, unpaired, high-dimensional sets of samples from potentially complex underlying distributions with strong interactions between variables.

3. Information may be lacking about the nature of possible differences (exploratory data analysis).

We will focus on the comparison of two high dimensional populations. Therefore, given two unpaired i.i.d. sets of samples $\mathbf{X}^{(n)} = x^{(1)}, \ldots, x^{(n)} \sim P_X$ and $\mathbf{Y}^{(m)} = y^{(1)}, \ldots, y^{(m)} \sim P_Y$, the goal is to answer the following two questions about the underlying multivariate random variables $X, Y \in \mathbb{R}^d$:

(Q1) Is $P_X = P_Y$?

(Q2) If not, what is the minimal subset of features $S \subseteq \{1, \ldots, d\}$ such that the marginal distributions differ $P_{X_S} \neq P_{Y_S}$ while $P_{X_{S^C}} = P_{Y_{S^C}}$ for the complement?

A finer version of (Q2) may additionally be posed which asks how much each feature contributes to the overall difference between the two probability distributions (with respect to the given scale on which the variables are measured).

Many two-sample analyses have focused on characterizing limited differences such as mean shifts (Lopes et al. 2011, Clemmensen et al. 2011). More general differences beyond the mean of each feature remain of interest, however, including variance/covariance of demographic statistics such as income. It is also undesirable to restrict the analysis to specific parametric differences, especially in exploratory analysis where the nature of the underlying distributions may be unknown. In the univariate case, a number of nonparametric tests of distribution equality based on statistical divergence are available with accompanying concentration results (van der Vaart & Wellner 1996). Popular examples of such *divergences* (also referred to as probability metrics) include: f-divergences (Kullback-Leibler, Hellinger, total-variation, etc.), the Kolmogorov distance, or the Wasserstein metric (Gibbs & Su 2002). Unfortunately, this univariate simplicity vanishes as the dimensionality d of the data grows, and a variety of complex statistical divergences have been designed to address some of the difficulties that appear in high-dimensional settings (Wei et al. 2015, Rosenbaum 2005, Szekely & Rizzo 2004, Gretton et al. 2012).

In this chapter, we propose the *principal differences analysis* (PDA) framework which circumvents the curse of dimensionality through explicit reduction back to the univariate case. Given a pre-specified statistical divergence D which measures the difference between univariate probability distributions, PDA seeks to find a projection β which maximizes $D(\beta^T X, \beta^T Y)$ subject to the constraints $||\beta||_2 \leq 1, \beta_1 \geq 0$ (the nonnegativity constraint on the first entry of β is merely included to avoid underspecification). This reduction is justified by the Cramer-Wold device, which ensures that $P_X \neq P_Y$ *if and only if* there exists a direction along which the univariate linearly projected distributions differ (Cramer & Wold 1936, Cuesta-Albertos et al. 2007, Jirak 2011). Assuming D is a *positive definite* divergence (meaning it is nonzero between any two distinct univariate distributions), the projection vector produced by PDA can thus capture arbitrary types of differences between high-dimensional P_X and P_Y. Furthermore, the approach can be straightforwardly modified to address (Q2) by introducing a sparsity penalty on β and examining the features with nonzero weight in the resulting optimal projection. The resulting comparison pertains to marginal distributions up to the sparsity level. We refer to this approach as *sparse differences analysis* or SPARDA.

2.1 Related work

Due to its fundamental scientific value, the problem of characterizing differences between populations, including feature selection, has received a great deal of study (Clemmensen et al. 2011, Tibshirani 1996, Bradley & Mangasarian 1998, Wei et al. 2015, Lopes et al. 2011). We limit our discussion to projection-based methods which, as a family of methods, are closest to our approach. For multivariate two-class data, the most widely adopted methods include sparse linear discriminant analysis (LDA) proposed by Clemmensen et al. (2011) and the logistic lasso of Tibshirani (1996).

While interpretable, these methods seek specific differences (e.g., covariance-rescaled average differences) or operate under stringent assumptions (e.g., log-linear model). In contrast, SPARDA (with a positive-definite divergence) aims to find features that characterize a priori unspecified differences between general multivariate distributions.

Perhaps most similar to our general approach is Direction-Projection-Permutation (DiProPerm) procedure of Wei et al. (2015), in which the data is first projected along the normal to the separating hyperplane (found using linear SVM, distance weighted discrimination, or the centroid method) followed by a univariate two-sample test on the projected data. The projections could also be chosen at random (Lopes et al. 2011). In contrast to our approach, the choice of the projection in such methods is not optimized for the test statistics. We note that by restricting the divergence measure in our technique, methods such as the linear support vector machine of Bradley & Mangasarian (1998) could be viewed as special cases. The divergence in this case would measure the margin between projected univariate distributions. While suitable for finding well-separated populations, it may fail to uncover more general differences between possibly multi-modal populations whose distributions heavily overlap in the feature space.

2.2 Using projections to characterize differences in distributions

For a given divergence measure D between two univariate random variables, the general idea of principal differences analysis is to find the projection $\widehat{\beta}$ that solves

$$\max_{\beta \in \mathcal{B}, ||\beta||_0 \leq k} \left\{ D(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(m)}) \right\} \tag{2.1}$$

where $\mathcal{B} := \{\beta \in \mathbb{R}^d : ||\beta||_2 \leq 1, \beta_1 \geq 0\}$ is the feasible set, $||\beta||_0 \leq k$ is the sparsity constraint, and $\beta^T \widehat{X}^{(n)}$ denotes the observed random variable that follows the empirical distribution of n samples of $\beta^T X$. Instead of imposing a hard cardinality constraint $||\beta||_0 \leq k$, we may instead penalize by adding a penalty term[1] $-\lambda||\beta||_0$ or its natural relaxation, the ℓ_1 shrinkage used in Lasso (Tibshirani 1996), sparse LDA (Clemmensen et al. 2011), and sparse PCA (D'Aspremont et al. 2007, Amini & Wainwright 2009). Sparsity in our setting explicitly restricts the comparison to the marginal distributions over features with non-zero coefficients. We can evaluate the null hypothesis $P_X = P_Y$ (or its sparse variant over marginals) using permutation testing with test statistic $D(\widehat{\beta}^T \widehat{X}^{(n)}, \widehat{\beta}^T \widehat{Y}^{(m)})$. In the basic permutation test, one simply randomizes the assignments of the observed data $x^{(1)}, \ldots, x^{(n)}, y^{(1)}, \ldots, y^{(m)}$ to create two new groups whose underlying distributions must be equal, as specified by the null hypothesis (Wei et al. 2015, Good 1994).

[1] In practice, shrinkage parameter λ (or explicit cardinality constraint k) may be chosen via cross-validation by maximizing the divergence between held-out samples.

The divergence D plays a key role in our analysis. If D is defined in terms of density functions as in f-divergence, one can use univariate kernel density estimation to approximate projected pdfs with additional tuning of the bandwidth hyperparameter. For a suitably chosen kernel (e.g. Gaussian), the unregularized PDA objective (without shrinkage) is a smooth function of β, and thus amenable to the projected gradient method (or its accelerated variants (Duchi et al. 2011, Wright 2010)). In contrast, when D is defined over the cumulative density functions along the projected direction – e.g. the Kolmogorov or Wasserstein distance that we focus on in this chapter – the objective is nondifferentiable due to the discrete jumps in the empirical CDF. We specifically address the combinatorial problem implied by the Wasserstein distance. Moreover, since the divergence assesses general differences between distributions, Equation (2.1) is typically a non-concave optimization. To this end, we develop a semi-definite relaxation of the problem into a concave formulation for use with the Wasserstein distance.

2.3 PDA using the Wasserstein distance

In the remainder of this chapter, we focus on the squared L_2 Wasserstein distance, defined as

$$D(X,Y) = \min_{P_{XY}} \mathbb{E}_{P_{XY}} ||X - Y||^2 \quad \text{s.t.} \quad (X,Y) \sim P_{XY},\ X \sim P_X,\ Y \sim P_Y \tag{2.2}$$

where the minimization is over all joint distributions over (X,Y) with given marginals P_X and P_Y. Intuitively interpreted as the amount of *work* required to transform one distribution into the other, D provides a natural dissimilarity measure between populations that integrates both the fraction of individuals which are different and the magnitude of these differences. While component analysis based on the Wasserstein distance has been limited to the work of Sandler & Lindenbaum (2011), this divergence has been successfully used in many other applications (Levina & Bickel 2001). In the univariate case, (2.2) may be analytically expressed as the L_2 distance between quantile functions. We can thus efficiently compute empirical projected Wasserstein distances by sorting X and Y samples along the projection direction to obtain quantile estimates. See §1.2 for more detailed background information on Wasserstein distance.

Under the squared L_2 Wasserstein distance, the empirical objective in Equation (2.1) between unpaired sampled populations $\{x^{(1)}, \ldots, x^{(n)}\}$ and $\{y^{(1)}, \ldots, y^{(m)}\}$ can be shown to be

$$\max_{\substack{\beta \in \mathcal{B} \\ ||\beta||_0 \leq k}} \left\{ \min_{M \in \mathcal{M},} \sum_{i=1}^{n} \sum_{j=1}^{m} (\beta^T x^{(i)} - \beta^T y^{(j)})^2 M_{ij} \right\} = \max_{\substack{\beta \in \mathcal{B} \\ ||\beta||_0 \leq k}} \left\{ \min_{M \in \mathcal{M}} \beta^T W_M \beta \right\} \tag{2.3}$$

where $\mathcal{M}$ is the set of all $n \times m$ nonnegative matching matrices defined in §1.2 with

fixed row sums $= 1/n$ and column sums $= 1/m$, $W_M := \sum_{i,j}[Z_{ij} \otimes Z_{ij}]M_{ij}$, and $Z_{ij} := x^{(i)} - y^{(j)}$. Note that if we omitted (or fixed) the inner minimization over the matching matrices and set $\lambda = 0$, the solution of (2.3) would be simply the largest eigenvector of W_M. Similarly, for the sparse variant without minizing over M, the problem would be solvable as sparse PCA (D'Aspremont et al. 2007, Amini & Wainwright 2009, Wang et al. 2014). The actual max-min problem in (2.3) is more complex and non-concave with respect to β. We propose a two-step procedure similar to "tighten after relax" framework used to attain minimax-optimal rates in sparse PCA by Wang et al. (2014). First, we first solve a convex relaxation of the problem and subsequently run a steepest ascent method (initialized at the global optimum of the relaxation) to greedily improve the current solution with respect to the original nonconvex problem whenever the relaxation is not tight.

Finally, we emphasize that PDA (and SPARDA) not only computationally resembles (sparse) PCA, but PCA is actually a special case of PDA in the Gaussian, paired-sample-differences setting. This connection is made explicit by considering the two-class problem with *paired* samples $(x^{(i)}, y^{(i)})$ where X, Y follow two multivariate Gaussian distributions. Here, the largest principal component of the (uncentered) differences $x^{(i)} - y^{(i)}$ is in fact equivalent to the direction which maximizes the projected Wasserstein difference between the distribution of $X - Y$ and a delta distribution at 0. Thus, PDA may be viewed as an extension of PCA to model the variation between unpaired data samples from two populations, where an optimal pairing is first inferred via the matching defined in the Wasserstein distance.

2.3.1 Semidefinite relaxation

The SPARDA problem may be expressed in terms of $d \times d$ symmetric matrices B as

$$\begin{aligned} &\max_{B} \min_{M \in \mathcal{M}} \operatorname{tr}(W_M B) \\ &\text{subject to } \operatorname{tr}(B) = 1,\ B \succeq 0,\ ||B||_0 \leq k^2,\ \operatorname{rank}(B) = 1 \end{aligned} \tag{2.4}$$

where the correspondence between (2.3) and (2.4) comes from writing $B = \beta \otimes \beta$ (note that any solution of (2.3) will have unit norm). When $k = d$, i.e., we impose no sparsity constraint as in PDA, we can relax by simply dropping the rank-constraint. The objective is then a supremum of linear functions of B and the resulting semidefinite problem is concave over a convex set and may be written as:

$$\max_{B \in \mathcal{B}_r} \min_{M \in \mathcal{M}} \operatorname{tr}(W_M B) \tag{2.5}$$

where $\mathcal{B}_r$ is the convex set of positive semidefinite $d \times d$ matrices with trace $= 1$. If $B^* \in \mathbb{R}^{d \times d}$ denotes the global optimum of this relaxation and $\operatorname{rank}(B^*) = 1$, then the best projection for PDA is simply the dominant eigenvector of B^* and the relaxation is tight. Otherwise, we can truncate B^* as in D'Aspremont et al. (2007), treating the

dominant eigenvector as an approximate solution to the original problem (2.3).

To obtain a relaxation for the sparse version where $k < d$ (SPARDA), we follow D'Aspremont et al. (2007) closely. Because $B = \beta \otimes \beta$ implies $||B||_0 \leq k^2$, we obtain an equivalent cardinality constrained problem by incorporating this nonconvex constraint into (2.4). Since $\mathrm{tr}(B) = 1$ and $||B||_F = ||\beta||_2^2 = 1$, a convex relaxation of the squared ℓ_0 constraint is given by $||B||_1 \leq k$. By selecting λ as the optimal Lagrange multiplier for this ℓ_1 constraint, we can obtain an equivalent penalized reformulation parameterized by λ rather than k. The sparse semidefinite relaxation is thus the following concave problem

$$\max_{B \in \mathcal{B}_r} \Big\{ \min_{M \in \mathcal{M}} \mathrm{tr}\,(W_M B) - \lambda ||B||_1 \Big\} \tag{2.6}$$

While our relaxation bears strong resemblance to the sparse PCA relaxation proposed by D'Aspremont et al. (2007), the inner minimization over matchings within our problem prevents direct application of general semidefinite programming solvers. Let $M(B)$ denote the matching that minimizes $\mathrm{tr}\,(W_M B)$ for a given B. Standard projected subgradient ascent could be applied to solve (2.6), where at the t^{th} iterate the (matrix-valued) subgradient is $W_{M(B^{(t)})}$. However, this approach requires solving optimal transport problems with large $n \times m$ matrices at each iteration. Instead, we turn to a dual form of (2.6), assuming $n \geq m$ (cf. Bertsekas (1998), Bertsekas & Eckstein (1988))

$$\max_{B \in \mathcal{B}_r, u \in \mathbb{R}^n, v \in \mathbb{R}^m} \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m} \min\{0,\ \mathrm{tr}([Z_{ij} \otimes Z_{ij}]\, B) - u_i - v_j\} + \frac{1}{n} \sum_{i=1}^{n} u_i + \frac{1}{m} \sum_{j=1}^{m} v_j - \lambda ||B||_1 \tag{2.7}$$

(2.7) is simply a maximization over $B \in \mathcal{B}_r$, $u \in \mathbb{R}^n$, and $v \in \mathbb{R}^m$ which no longer requires matching matrices nor their cumbersome row/column constraints. While dual variables u and v can be solved in closed form for each fixed B (via sorting), we describe a simple sub-gradient approach that works better in practice.

RELAX Algorithm: Solves the dualized semidefinite relaxation of SPARDA (2.7). Returns the largest eigenvector of the solution to (2.6) as the desired projection direction for SPARDA.

Input: d-dimensional data $x^{(1)}, \ldots, x^{(n)}$ and $y^{(1)}, \ldots, y^{(m)}$ (with $n \geq m$)
Parameters: $\lambda \geq 0$ controls the amount of regularization, $\gamma > 0$ is the step-size used for B updates, $\eta > 0$ is the step-size used for updates of dual variables u and v, T is the maximum number of iterations without improvement in cost after which algorithm terminates.

Initialize $\beta^{(0)} \leftarrow \left[\frac{\sqrt{d}}{d}, \ldots, \frac{\sqrt{d}}{d}\right]$, $B^{(0)} \leftarrow \beta^{(0)} \otimes \beta^{(0)} \in \mathcal{B}_r$, $u^{(0)} \leftarrow \mathbf{0}_{n\times 1}$, $v^{(0)} \leftarrow \mathbf{0}_{m\times 1}$

While the number of iterations since last improvement in objective function is less than T:

  $\partial u \leftarrow [1/n, \ldots, 1/n] \in \mathbb{R}^n$, $\partial v \leftarrow [1/m, \ldots, 1/m] \in \mathbb{R}^m$, $\partial B \leftarrow \mathbf{0}_{d\times d}$

  For $i, j \in \{1, \ldots, n\} \times \{1, \ldots, m\}$:

    $Z_{ij} \leftarrow x^{(i)} - y^{(j)}$

    If $\mathrm{tr}([Z_{ij} \otimes Z_{ij}]B^{(t)}) - u_i^{(t)} - v_j^{(t)} < 0$:

      $\partial u_i \leftarrow \partial u_i - 1/m$, $\partial v_j \leftarrow \partial v_j - 1/m$, $\partial B \leftarrow \partial B + Z_{ij} \otimes Z_{ij}/m$

  End For

  $u^{(t+1)} \leftarrow u^{(t)} + \eta \cdot \partial u$ and $v^{(t+1)} \leftarrow v^{(t)} + \eta \cdot \partial v$

  $B^{(t+1)} \leftarrow \textbf{Projection}\left(B^{(t)} + \frac{\gamma}{||\partial B||_F} \cdot \partial B \; ; \; \lambda, \; \gamma/||\partial B||_F\right)$

Output: $\widehat{\beta}_{\text{relax}} \in \mathbb{R}^d$ defined as the largest eigenvector (based on corresponding eigenvalue's magnitude) of the matrix $B^{(t^*)}$ which attained the best objective value over all iterations.

Projection Algorithm: Projects matrix onto positive semidefinite cone of unit-trace matrices $\mathcal{B}_r$ (the feasible set in our relaxation). Step 4 applies soft-thresholding proximal operator for sparsity.

Input: $B \in \mathbb{R}^{d \times d}$
Parameters: $\lambda \geq 0$ controls the amount of regularization, $\delta = \gamma / ||\partial B||_F \geq 0$ is the actual step-size used in the B-update.
1: $Q \Lambda Q^T \leftarrow$ eigendecomposition of B
2: $w^* \leftarrow \arg\min \left\{ ||w - \text{diag}(\Lambda)||_2^2 : w \in [0,1]^d, ||w||_1 = 1 \right\}$ (Quadratic program)
3: $\widetilde{B} \leftarrow Q \cdot \text{diag}\{w_1^*, \ldots, w_d^*\} \cdot Q^T$
4: **If** $\lambda > 0$: **For** $r, s \in \{1, \ldots, d\}^2$: $\widetilde{B}_{r,s} \leftarrow \text{sign}(\widetilde{B}_{r,s}) \cdot \max\{0, |\widetilde{B}_{r,s}| - \delta\lambda\}$
Output: $\widetilde{B} \in \mathcal{B}_r$

The RELAX algorithm is a projected subgradient method with supergradients computed in Steps 3 - 8. For scaling to large samples, one may alternatively employ *incremental* supergradient directions (Bertsekas 2011) where Step 4 would be replaced by drawing random (i, j) pairs. After each subgradient step, projection back into the feasible set $\mathcal{B}_r$ is done via a quadratic program involving the current solution's eigenvalues. In SPARDA, sparsity is encouraged via the soft-thresholding proximal map corresponding to the ℓ_1 penalty. The overall form of our iterations matches subgradient-proximal updates (4.14)-(4.15) in Bertsekas (2011). By the convergence analysis in §4.2 of Bertsekas (2011), the RELAX algorithm (as well as its incremental variant) is guaranteed to approach the optimal solution of the dual which also solves (2.6), provided we employ sufficiently large T and small step-sizes. In practice, fast and accurate convergence is attained by: (a) renormalizing the B-subgradient (Step 10) to ensure balanced updates of the unit-norm constrained B, (b) using diminishing learning rates which are initially set larger for the unconstrained dual variables (or even taking multiple subgradient steps in the dual variables per each update of B).

2.3.2 Tightening after relaxation

It is unreasonable to expect that our semidefinite relaxation is always tight. Therefore, we can sometimes further refine the projection $\widehat{\beta}_{\text{relax}}$ obtained by the RELAX algorithm by using it as a starting point in the original non-convex optimization. We introduce a sparsity constrained *tightening* procedure for applying projected gradient ascent for the original nonconvex objective $J(\beta) = \min_{M \in \mathcal{M}} \beta^T W_M \beta$ where β is now forced to lie in $\mathcal{B} \cap \mathcal{S}_k$ and $\mathcal{S}_k := \{\beta \in \mathbb{R}^d : ||\beta||_0 \leq k\}$. The sparsity level k is fixed based on the relaxed solution ($k = ||\widehat{\beta}_{\text{relax}}||_0$). After initializing $\beta^{(0)} = \widehat{\beta}_{\text{relax}} \in \mathbb{R}^d$, the tightening procedure iterates steps in the gradient direction of J followed by straight-

forward projections into the unit half-ball $\mathcal{B}$ and the set $\mathcal{S}_k$ (accomplished by greedily truncating all entries of β to zero besides the largest k in magnitude).

Let $M(\beta)$ again denote the matching matrix chosen in response to β. J fails to be differentiable at the $\widetilde{\beta}$ where $M(\widetilde{\beta})$ is not unique. This occurs, e.g., if two samples have identical projections under $\widetilde{\beta}$. While this situation becomes increasingly likely as $n, m \to \infty$, J interestingly becomes smoother overall (assuming the distributions admit density functions). For all other β: $M(\beta') = M(\beta)$ where β' lies in a small neighborhood around β and J admits a well-defined gradient $2W_{M(\beta)}\beta$. In practice, we find that the tightening always approaches a local optimum of J with a diminishing step-size. We note that, for a given projection, we can efficiently calculate gradients without recourse to matrices $M(\beta)$ or $W_{M(\beta)}$ by sorting ${\beta^{(t)}}^T x^{(1)}, \ldots, {\beta^{(t)}}^T x^{(n)}$ and ${\beta^{(t)}}^T y^{(1)}, \ldots, {\beta^{(t)}}^T y^{(m)}$. The gradient is directly derivable from expression (2.3) where the nonzero $M_{ij}(\beta)$ entries are easily determined by appropriately matching empirical quantiles (represented by sorted indices) of the data. This extremely efficient computation of $M(\beta)$ is possible since the univariate Wasserstein distance is simply the L_2 distance between quantile functions (recall Lemma 1 in §1.2). Additional computation can be saved by employing insertion sort which runs in nearly linear time for almost sorted points (in iteration $t-1$, the points have been sorted along the $\beta^{(t-1)}$ direction and their sorting in direction $\beta^{(t)}$ is likely similar under small step-size). Thus the tightening procedure is much more efficient than the RELAX algorithm (respective runtimes are $O(dn \log n)$ vs. $O(d^3 n^2)$ per iteration).

We require the combined steps for good performance. The projection found by the tightening algorithm heavily depends on the starting point $\beta^{(0)}$, finding only the closest local optimum (as in Figure 2-1). It is thus important that $\beta^{(0)}$ is already a good solution, as can be produced by our RELAX algorithm. Additionally, we note that as first-order methods, both the RELAX and tightening algorithms are amendable to a number of (sub)gradient-acceleration schemes (e.g. momentum techniques, adaptive learning rates, or FISTA and other variants of Nesterov's method (Wright 2010, Duchi et al. 2011, Beck & Teboulle 2009)).

2.3.3 Properties of semidefinite relaxation

We conclude the algorithmic discussion by informally highlighting a few settings in which our PDA relaxation is tight. Assuming $n, m \to \infty$, each of (i)-(iii) implies that the B^* which maximizes (2.5) is nearly rank one, or equivalently $B^* \approx \widetilde{\beta} \otimes \widetilde{\beta}$. Thus, the tightening procedure initialized at $\widetilde{\beta}$ will produce a global maximum of the PDA objective in each of these cases.

(i) There exists direction in which the *projected* Wasserstein distance between X and Y is nearly as large as the overall Wasserstein distance in $\mathbb{R}^d$. This occurs for example if $||\mathbb{E}[X] - \mathbb{E}[Y]||_2$ is large while both $||\mathrm{Cov}(X)||_F$ and $||\mathrm{Cov}(Y)||_F$

are small (the distributions need not be Gaussian).

(ii) $X \sim N(\mu_X, \Sigma_X)$ and $Y \sim N(\mu_Y, \Sigma_Y)$ with $\mu_X \neq \mu_Y$ and $\Sigma_X \approx \Sigma_Y$.

(iii) $X \sim N(\mu_X, \Sigma_X)$ and $Y \sim N(\mu_Y, \Sigma_Y)$ with $\mu_X = \mu_Y$ where the underlying covariance structure is such that $\arg\max_{B \in \mathcal{B}_r} ||(B^{1/2}\Sigma_X B^{1/2})^{1/2} - (B^{1/2}\Sigma_Y B^{1/2})^{1/2}||_F^2$ is nearly rank 1. For example, if the primary difference between covariances is a shift in the marginal variance of some features, i.e. $\Sigma_Y \approx V \cdot \Sigma_X$ where V is a diagonal matrix.

Condition (i) is derived by noting that (2.5) has rank one solution when the objective is approximately linear in B. Conditions (ii) and (iii) follow from the fact that (2.5) is actually the Wasserstein distance between random variables $B^{1/2}X$ and $B^{1/2}Y$. Furthermore, when X is Gaussian, AX follows a $N(A\mu_X, A\Sigma_X A^T)$ distribution, and the Wasserstein distance between (multivariate) Gaussian distributions can be analytically expressed as

$$W(X, Y) = ||\mu_X - \mu_Y||_2^2 + ||\Sigma_X^{1/2} - \Sigma_Y^{1/2}||_F^2$$

2.4 Theoretical results

In this section, we characterize various statistical properties of an empirical divergence-maximizing projection $\widehat{\beta} := \arg\max_{\beta \in \mathcal{B}} D(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(n)})$, although we note that the algorithms may not succeed in finding such a global maximum for severely non-convex problems. Throughout, D denotes the squared L_2 Wasserstein distance between univariate distributions, the C values (with various subscripts) represent universal constants that change from line to line, and we employ hat notation to represent empirical versions of all distributional quantities. F is defined the cumulative density function of a random variable, and the corresponding quantile function is $F^{-1}(p) := \inf\{x : F(x) \geq p\}$.

To simplify our analysis, we make the following assumptions throughout:

(A1) The sample number of samples is drawn from each distribution ($n = m$).

(A2) Random variables X and Y admit continuous density functions.

(A3) X and Y are compactly supported with nonzero density in the Euclidean ball of radius R.

Note that these assumptions ensure the quantile function equals the unique inverse of any projected CDF. Our theory can be generalized beyond these conditions to obtain similar (but far more complex) statements through careful treatment of the distributions' tails and zero-density regions where CDFs are flat.

Theorem 1 provides basic concentration results for the projections used in empirical applications our method. To relate distributional differences between X, Y in the ambient d-dimensional space with their estimated divergence along the univariate linear representation chosen by PDA, we turn to Theorems 2 and 3. Finally, Theorem 4 provides sparsistency guarantees for SPARDA in the case where X, Y exhibit large differences over a certain feature subset (of known cardinality).

Theorem 1. *Suppose (A1)-(A3) hold and there exists direction $\beta^* \in \mathcal{B}$ such that $D(\beta^{*T}X, \beta^{*T}Y) \geq \Delta$. Then:*

$$D(\widehat{\beta}^T\widehat{X}^{(n)}, \widehat{\beta}^T\widehat{Y}^{(n)}) > \Delta - \epsilon \qquad \textit{with probability greater than } 1 - 4\exp\left(-\frac{n\epsilon^2}{16R^4}\right)$$

Proof. Since $\widehat{\beta}$ maximizes the empirical divergence, we have:

$$\begin{aligned}
&\Pr(D(\widehat{\beta}^T\widehat{X}^{(n)}, \widehat{\beta}^T\widehat{Y}^{(n)}) > \Delta - \epsilon) \\
\geq &\Pr(D(\beta^{*T}\widehat{X}^{(n)}, \beta^{*T}\widehat{Y}^{(n)}) > \Delta - \epsilon) \\
\geq &\Pr(D(\beta^{*T}\widehat{X}^{(n)}, \beta^{*T}X) + D(\beta^{*T}\widehat{Y}^{(n)}, \beta^{*T}Y) \leq \epsilon) \\
\geq &1 - 4\exp\left(-\frac{n\epsilon^2}{16R^4}\right) \text{ applying Lemma 2 and the union bound.}
\end{aligned}$$

□

Theorem 2. *Under the same assumptions as Theorem 1, if X and Y are identically distributed in $\mathbb{R}^d$, then: $D(\widehat{\beta}^T\widehat{X}^{(n)}, \widehat{\beta}^T\widehat{Y}^{(n)}) < \epsilon$ with probability greater than*

$$1 - C_1\left(1 + \frac{R^2}{\epsilon}\right)^d \exp\left(-\frac{C_2}{R^4}n\epsilon^2\right)$$

Proof. We first construct a fine grid of points $\{\alpha_1, \ldots, \alpha_S\}$ which form an (ϵ/R^2)-net covering the surface of the unit ball in $\mathbb{R}^d$. When $P_X = P_Y$, the Cramer-Wold device (Cramer & Wold 1936) implies that for any point in our grid: $D(\alpha_s^T X, \alpha_s^T Y) = 0$. A result analogous to Theorem 1 implies $D(\alpha_s^T\widehat{X}^{(n)}, \alpha_s^T\widehat{Y}^{(n)}) > \epsilon$ with probability $< C_1\exp\left(-\frac{C_2}{R^4}n\epsilon^2\right)$.

Subsequently, we apply the union bound over the finite set of all grid points. The total number of points under consideration is the covering number of the unit-sphere which is $\left(1 + \frac{2R^2}{\epsilon}\right)^d$. Thus, the probability that $D(\alpha_s^T\widehat{X}^{(n)}, \alpha_s^T\widehat{Y}^{(n)}) < \epsilon$ simultaneously for all points in the grid is at least

$$C_1\left(1 + \frac{2R^2}{\epsilon}\right)^d \exp\left(-\frac{C_2}{R^4}n\epsilon^2\right)$$

By construction, there must exist grid point α_0 such that $||\widehat{\beta} - \alpha_0||_2 < \epsilon/R^2$. By Lemma 3

$$D(\widehat{\beta}^T \widehat{X}^{(n)}, \widehat{\beta}^T \widehat{Y}^{(n)}) \leq D(\alpha_0{}^T \widehat{X}^{(n)}, \alpha_0{}^T \widehat{Y}^{(n)}) + C\epsilon$$

thus completing the proof. □

To measure the difference between the original (unprojected) random variables $X, Y \in \mathbb{R}^d$, we define the following metric between distributions on $\mathbb{R}^d$ which is parameterized by $a \geq 0$ (see also Jirak (2011)):

$$T_a(X, Y) := |\Pr(|X_1| \leq a, \ldots, |X_d| \leq a) - \Pr(|Y_1| \leq a, \ldots, |Y_d| \leq a)| \tag{2.8}$$

For our subsequent theory, we also adopt the following simplifying assumptions:

(A4) Y has sub-Gaussian tails, meaning CDF F_Y satisfies: $1 - F_Y(y) \leq \frac{C}{y} \exp(-y^2/2)$.

(A5) $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ (note that mean differences can trivially be captured by linear projections, so these are not the differences of interest in the following theory).

(A6) The data have been normalized such that $\text{Var}(X_\ell) = 1$ for $\ell = 1, \ldots, d$.

Theorem 3. *Suppose (A1)-(A6) hold and there exists* $a \geq 0$ *s.t.* $T_a(X, Y) > h\left(g(\Delta)\right)$ *where* $h\left(g(\Delta)\right) := \min\{\Delta_1, \Delta_2\}$ *with*

$$\Delta_1 := (a + d)^d (g(\Delta) + d) + \exp(-a^2/2) + \psi \exp\left(-1/(\sqrt{2}\psi)\right) \tag{2.9}$$

$$\Delta_2 := \left(g(\Delta) + \exp(-a^2/2)\right) \cdot d \tag{2.10}$$

$\psi := ||Cov(X)||_1$, $g(\Delta) := \Delta^4 \cdot (1 + \Phi)^{-4}$, *and* $\Phi := \sup_{\alpha \in \mathcal{B}} \left\{ \sup_y |f_{\alpha^T Y}(y)| \right\}$ *with* $f_{\alpha^T Y}(y)$ *defined as the density of the projection of* Y *in the* α *direction. Then:*

$$D(\widehat{\beta}^T \widehat{X}^{(n)}, \widehat{\beta}^T \widehat{Y}^{(n)}) > C\Delta - \epsilon \tag{2.11}$$

with probability greater than $1 - C_1 \exp\left(-\frac{C_2}{R^4} n\epsilon^2\right)$

Proof. Our proof relies primarily on a quantitative form of the Cramer-Wold result derived by Jirak (2011). This statement only requires that the distribution of one of X or Y has rapidly decaying tails. We adapt Theorem 3.1 from Jirak (2011) in its contrapositive form: If $\exists\ a \geq 0$ such that $T_a(X, Y) > h(g(\Delta))$, then $\exists \beta \in \mathcal{B}$ such that

$$\sup_{z \in \mathbb{R}} \left| \Pr\left(\beta^T X \leq z\right) - \Pr\left(\beta^T Y \leq z\right) \right| > g(C\Delta) \tag{2.12}$$

Subsequently we leverage a number of well-characterized relationships between different probability metrics (summarized in Gibbs & Su (2002)) to lower bound the projected (squared) Wasserstein distance between the underlying random variables.

Letting K_β denote the projected Kolmogorov distance in (2.12), we have that the β-projected Lévy-distance, L_β satisfies:

$$K_\beta \leq (1+\Phi)L_\beta \qquad \text{where } \Phi := \sup_{\alpha\in\mathcal{B}} \left\{ \sup_y |f_{\alpha^T Y}(y)| \right\} \tag{2.13}$$

and $f_{\alpha^T Y}(y)$ is the density of the projection of Y in the α direction.

In turn the projected Lévy distance L_β is bounded above by the Prokhorov metric which itself is bounded above by the square root of the β-projected Wasserstein distance. Following the chain of inequalities, we obtain: $D(\beta^T X, \beta^T Y) \geq C\Delta$, to which we can apply Theorem 1 to obtain the desired probabilistic bound on the empirical projected divergence. □

Theorem 4. *Define C as in (2.11) and assume (A1)-(A6). Suppose there exists certain feature subsets $S \subset \{1,\ldots,d\}$ such that $\max_a T_a(X_S, Y_S) \geq h\left(g\left(\epsilon(d+1)/C\right)\right)$, and remaining marginal distributions X_{S^C} , Y_{S^C} are identical. If we take S to be the smallest of all such feature subsets and S is unique with cardinality $|S| = k$, then:*

$$\widehat{\beta}^{(k)} := \arg\max_{\beta\in\mathcal{B}} \left\{ D(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(n)}) : ||\beta||_0 \leq k \right\}$$

satisfies $\widehat{\beta}_i^{(k)} \neq 0$ and $\widehat{\beta}_j^{(k)} = 0 \;\; \forall\, i \in S, j \in S^C$ with probability greater than

$$1 - C_1 \left(1 + \frac{R^2}{\epsilon}\right)^{d-k} \exp\left(-\frac{C_2}{R^4} n\epsilon^2\right)$$

Proof. Intuitively, the properties of the feature subset S imply that Theorem 4 holds for underlying distributions whose marginal distributions over S^C are identical, while the marginal distributions over S are highly different, but only if all variables in S are considered (removal of any feature $i \in S$ results in a substantially decreased difference in the remaining marginal distributions over $S\backslash i$). Theorem 2 implies that with high probability, any unit-vector $\beta_{S^C} \in \mathbb{R}^{d-k}$ must satisfy $D(\beta_{S^C}^T \widehat{X}_{S^C}^{(n)}, \beta_{S^C}^T \widehat{Y}_{S^C}^{(n)}) < \epsilon$, while Theorem 3 specifies the probability that there exists unit-vector $\beta_S \in \mathbb{R}^k$ such that $D(\widehat{\beta}_S^T \widehat{X}_S^{(n)}, \widehat{\beta}_S^T \widehat{Y}_S^{(n)}) > d \cdot \epsilon$.

A bound for the probability that both theorems' conclusions hold simultaneously may be obtained by the union bound. When this is the case, it is clear that the optimal k-sparse $\widehat{\beta} \in \mathbb{R}^d$ must obey the sparsity pattern specified in the statement of Theorem 4. To see this, consider any $\beta \in \mathcal{B}$ with $\beta_j \neq 0$ for some $j \in S^C$ and note that it is always possible to produce a strictly superior projection by setting $\beta_j = 0$ and distributing the additional weight $|\beta_j|$ among the features in S in an optimal manner. □

2.4.1 Auxiliary lemmas

Lemma 2. *For bounded univariate random variable $Z \in [-R, R]$ with nonzero continuous density in this region, we have*

$$D(\widehat{Z}^{(n)}, Z) > \epsilon \qquad \text{with probability at most } 2\exp\left(-\frac{n\epsilon^2}{8R^4}\right)$$

Proof. On the real line, the (squared) Wasserstein distance is given by:

$$\begin{aligned}
D(\widehat{Z}^{(n)}, Z) &= \int_0^1 \left(\widehat{F}_Z^{-1}(p) - F_Z^{-1}(p)\right)^2 \mathrm{d}p \\
&= 4R^2 \int_0^1 \left(\frac{\widehat{F}_Z^{-1}(p) - F_Z^{-1}(p)}{2R}\right)^2 \mathrm{d}p \quad \text{where } \left|\frac{\widehat{F}_Z^{-1}(p) - F_Z^{-1}(p)}{2R}\right| \leq 1 \text{ for each } p \in (0,1) \\
&\leq 4R^2 \int_0^1 \left|\frac{\widehat{F}_Z^{-1}(p) - F_Z^{-1}(p)}{2R}\right| \mathrm{d}p \\
&= 2R \int_0^1 \left|\widehat{F}_Z^{-1}(p) - F_Z^{-1}(p)\right| \mathrm{d}p \\
&= 2R \int_{-\infty}^{\infty} \left|\widehat{F}_Z(z) - F_Z(z)\right| \mathrm{d}z \\
&\quad \text{by the equivalence of the (empirical) quantile function and inverse (empirical) CDF} \\
&\leq 4R^2 \cdot \sup_z \left|\widehat{F}_Z(z) - F_Z(z)\right| \\
&\leq \epsilon \text{ with probability } \geq 1 - 2\exp\left(-\frac{n\epsilon^2}{8R^4}\right)
\end{aligned}$$

by the Dvoretzky-Kiefer-Wolfowitz inequality (van der Vaart & Wellner 1996). □

Lemma 3. *For $\alpha, \beta \in \mathcal{B}$ such that $||\alpha - \beta||_2 < \epsilon$, we have:*

$$|D(\alpha^T \widehat{X}^{(n)}, \alpha^T \widehat{Y}^{(n)}) - D(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(n)})| \leq C\epsilon R^2 \tag{2.14}$$

Proof. We assume that the α-projected divergence is larger than the β-projected divergence, and write:

$$D(\beta^T \widehat{X}^{(n)}, \beta^T \widehat{Y}^{(n)}) = \min_{M \in \mathcal{M},} \sum_{i=1}^{n} \sum_{j=1}^{m} (\beta^T x^{(i)} - \beta^T y^{(j)})^2 M_{ij}$$

recalling that $\mathcal{M}$ is the set of matching matrices defined previously. Let $M(\beta)$ denote the matrix which is used in the computation of the β-projected empirical Wasserstein distance (the minimizer of the righthand side of the above expression). Thus, we can

express (2.14) as:

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{m}(\alpha^T x^{(i)} - \alpha^T y^{(j)})^2 M(\alpha)_{ij} - \sum_{i=1}^{n}\sum_{j=1}^{m}(\beta^T x^{(i)} - \beta^T y^{(j)})^2 M(\beta)_{ij} \\
\leq &\sum_{i=1}^{n}\sum_{j=1}^{m}(\alpha^T x^{(i)} - \alpha^T y^{(j)})^2 M(\beta)_{ij} - \sum_{i=1}^{n}\sum_{j=1}^{m}(\beta^T x^{(i)} - \beta^T y^{(j)})^2 M(\beta)_{ij} \\
\leq &\sum_{i=1}^{n}\sum_{j=1}^{m}\left[(\alpha^T(x^{(i)} - y^{(j)}))^2 - (\beta^T(x^{(i)} - y^{(j)}))^2\right] M(\beta)_{ij} \\
= &\sum_{i=1}^{n}\sum_{j=1}^{m}\left[(\alpha-\beta)^T(x^{(i)} - y^{(j)}) \cdot (\alpha+\beta)^T(x^{(i)} - y^{(j)})\right] M(\beta)_{ij} \\
\leq &\sum_{i=1}^{n}\sum_{j=1}^{m} C\epsilon R^2 M(\beta)_{ij} = C\epsilon R^2
\end{aligned}$$

□

2.5 Empirical results

2.5.1 Nonconvexity of the PDA objective

Figure 2-1 illustrates the cost function of PDA pertaining to two 3-dimensional distributions. Note that only dimensions 2 and 3 of the projection-space are plotted in the figure since $\beta_1 = \sqrt{1 - \sum_{\ell=2}^{d} \beta_\ell^2}$ is fixed for the unit-norm projections of interest. Here, we apply PDA to $n = m = 1000$ points sampled from mean-zero Gaussian distributions with the following respective covariance matrices:

$$\Sigma_X = \begin{bmatrix} 1 & 0.2 & 0.4 \\ 0.2 & 1 & 0 \\ 0.4 & 0 & 1 \end{bmatrix} \qquad \Sigma_Y = \begin{bmatrix} 1 & -0.9 & 0 \\ -0.9 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Due to the large sample sizes, the empirical distributions accurately represent the underlying populations, and thus the projection produced by the tightening procedure (in green) is significantly inferior to the projection produced by the RELAX algorithm (in red) in terms of actual divergence captured. In order to ensure good results in practice, it is therefore important to use RELAX before tightening as we previously advised. This example also illustrates a setting where our convex relaxation is tight (the RELAX solution is already globally optimal without any further greedy optimization).

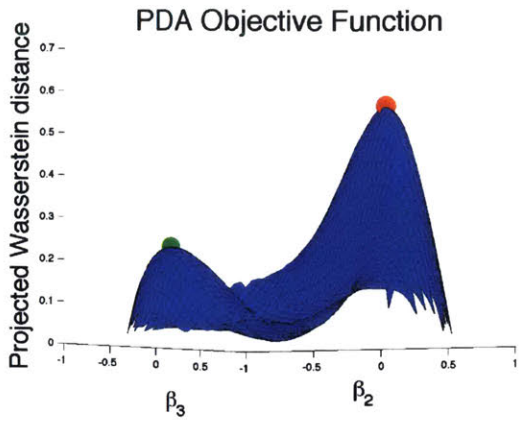


Figure 2-1: PDA objective function for data sampled from two 3-dimensional Gaussian distributions with mean zero. The point of convergence $\widehat{\beta}$ of the tightening method after random initialization (in green) is significantly inferior to the solution produced by the RELAX algorithm (in red).

2.5.2 Variable selection

The synthetic MADELON dataset used in the NIPS 2003 feature selection challenge consists of points ($n = m = 1000, d = 500$) which have 5 features scattered on the vertices of a five-dimensional hypercube (so that interactions between features must be considered in order to distinguish the two classes), 15 features that are noisy linear combinations of the original five, and 480 useless features (Guyon et al. 2006). While the focus of the challenge was on extracting features useful to classifiers, we direct our attention toward more interpretable models. Figure 2-2 demonstrates how well SPARDA, the top sparse principal component (Zou et al. 2005), sparse LDA (Clemmensen et al. 2011), and the logistic lasso (Tibshirani 1996) are able to identify the 20 relevant features over different settings of their respective regularization parameters (which control how many variables are selected by each method). The red asterisk indicates the SPARDA result with λ automatically selected via our cross-validation procedure (without information of the underlying features' importance), and the black asterisk indicates the best reported result in the challenge (Guyon et al. 2006).

The restrictive assumptions in logistic regression and linear discriminant analysis are not satisfied in this complex dataset resulting in poor performance. Despite being class-agnostic, PCA was successfully utilized by numerous challenge participants (Guyon et al. 2006), and we find that the sparse PCA performs on par with logistic regression and LDA. Although the lasso fairly efficiently picks out 5 relevant features, it struggles to identify the rest due to severe multi-colinearity. Similarly, the challenge-winning Bayesian SVM with Automatic Relevance Determination (Guyon et al. 2006) only selects 8 of the 20 relevant features. In many applications, the goal is to thoroughly characterize the set of differences rather than select a subset of features that maintains predictive accuracy. SPARDA is better suited for this alternative objective.

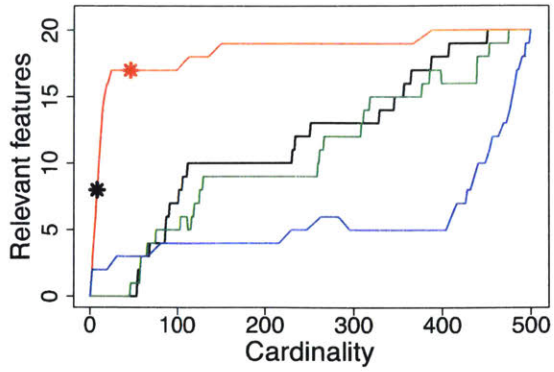


Figure 2-2: Number of relevant variables correctly selected by different methods applied to the MADELON data, over different settings of their respective regularization parameters (which determine the cardinality of each method's chosen variable set). The curves indicate the variable-selection precision of: SPARDA (red), the top sparse principal component (black), sparse LDA (green), and the logistic lasso (blue).

Many settings of λ return 14 of the relevant features with zero false positives. If λ is chosen automatically through cross-validation, the projection returned by SPARDA contains 46 nonzero elements of which 17 correspond to relevant features.

2.5.3 Two-sample testing in high dimensions

Figure 2-3 compares the statistical power of SPARDA, PDA, Maximum Mean Discrepancy (Gretton et al. 2012), DiProPerm (Wei et al. 2015), and the overall Wasserstein distance (over the ambient feature-space $\mathbb{R}^d$) in two-sample testing problems where $P_X \neq P_Y$ and the underlying differences have varying degrees of sparsity. Here, d indicates the overall number of features in the data being tested, of which only the first 3 exhibit any difference between the two populations.

We set the features of the underlying X and Y to mean-zero multivariate Gaussian distributions in blocks of 3, where within each block, (common) covariance parameters are sampled from the Wishart($\mathbf{I}_{3\times 3}$) distribution with 3 degrees of freedom. Only for the first block of 3 features do we sample a separate covariance matrix for X and a separate covariance matrix for Y, so all differences between the two distributions lie in the first 3 features. To generate a dataset with $d = 3 \times \ell$, we simply concatenate ℓ of our blocks together (always including the first block with the underlying difference) and draw $n = m = 100$ points from each class. We generate 20 datasets by increasing ℓ (so the largest $d = 60$), and repeat this entire experiment 10 times reporting the average p-values in Figure 2-3.

As we evaluate the significance of each method's statistic via permutation testing, all the tests are guaranteed to exactly control Type I error (Good 1994), and we thus only compare their respective power in determining $P_X \neq P_Y$ setting. Each p-value

is obtain by randomly permuting the class labels and recomputing the test statistic 100 times (where we use the same permutations between all datasets). In SPARDA, regularization parameter λ is re-selected using our cross-validation technique in each permutation. The overall Wasserstein distance in the ambient space is computed by solving the optimal transport problem via linear programming (Levina & Bickel 2001), and we note the similarity between this statistic and the cross-match test of Rosenbaum (2005). A popular kernel method for testing high-dimensional distribution equality, the mean map discrepancy, is computed using the Gaussian kernel with bandwidth parameter chosen by the "median trick" of Gretton et al. (2012) (which is very similar to the energy test of Szekely & Rizzo (2004)). Finally, we also compute the DiProPerm statistic, employing the the DWD-t variant recommended for testing general equality of distributions by Wei et al. (2015).

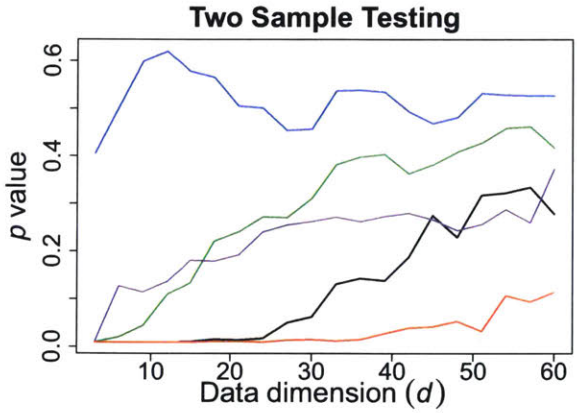


Figure 2-3: Average p-values (over 10 repetitions) for multivariate two-sample tests produced by SPARDA (red), PDA (purple), the overall Wasserstein distance in $\mathbb{R}^d$ (black), Maximum Mean Discrepancy (green), and DiProPerm (blue). While the dimensionality of the underlying distributions is varied, the underlying differences remain limited to 3 dimensions in all cases.

Figure 2-3 demonstrates clear superiority of SPARDA which leverages the underlying sparsity to maintain high power even with the increasing overall dimensionality. Even when all the features differ (when $d = 3$), SPARDA matches the power of methods that consider the full space despite only selecting a single direction (which cannot be based on mean-differences as there are none in this controlled data). This experiment also demonstrate that the unregularized PDA retains much greater power than DiProPerm, a similar projection-based method that performs poorly when the data are not linearly separable (as is the case here).

2.6 Cellular gene expression differences between the somatosensory cortex and hippocampus

Recent technological advances allow complete transcriptome profiling in thousands of individual cells with the goal of fine molecular characterization of cell populations. The (beyond the crude average-tissue-level expression measure that is currently standard) (Geiler-Samerotte et al. 2013). We apply SPARDA to expression measurements of 10,305 genes profiled in 1,691 single cells from the somatosensory cortex and 1,314 hippocampus cells sampled from the brains of juvenile mice by Zeisel et al. (2015). Playing critical roles in the brain, the somatosensory cortex (linked to the senses) and hippocampal region (linked to memory regulation and spatial coding) contain a diversity of cell types. It is thus of great interest to identify how cell populations in these regions diverge in developing brains, a question we address by applying SPARDA to scRNA-seq data from these regions.

Following Trapnell et al. (2014), we represent gene expression by log-transformed FPKM computed from the sequencing read counts[2], so values are directly comparable between genes. Because expression measurements from individual cells are poorer in quality than transcriptome profiles obtained in aggregate across tissue samples (due to a drastically reduced amount of available RNA), it is important to filter out poorly measured genes and we retain a set of 10,305 genes that are measured with sufficient accuracy for informative analysis.

The resulting $\widehat{\beta}$ produced by SPARDA identifies many previously characterized subtype-specific genes and is in many respects more informative than the results of standard differential expression methods. Table 2.1 and Figure 2-4 demonstrate that SPARDA discovers many interesting genes which are already known to play important functional roles in these regions of the brain. For comparison, we also run LIMMA, a standard method for differential expression analysis which tests for marginal mean-differences on a gene-by-gene basis (Ritchie et al. 2015). Ordering the significant genes under LIMMA by magnitude of their mean expression difference, we find that 3 of the top 10 genes identified by SPARDA also appear in this top 10 list (*Crym*, *Spink8*, *Neurod6*), demonstrating SPARDA's implicit attraction toward large first-order differences over more nuanced changes in practice. Because only few genes can feasibly be considered for subsequent experimentation in these studies, a good tool for differential expression analysis must rank the most relevant genes very highly in order for researchers to take note.

One particularly relevant gene in this data is *Snca*, a presynaptic signaling and membrane trafficking gene whose defects are implicated in both Parkinson and Alzheimer's disease (Lesage & Brice 2009, Linnertz et al. 2014). While *Snca* is ranked 11^{th} highest under SPARDA, it only ranks 349 according to LIMMA p-values and 95 based on absolute mean-shift. Figure 2-5 shows that the primary change in *Snca* expression

[2]available in NCBI's Gene Expression Omnibus (under accession GSE60361)

GENE	WEIGHT	DESCRIPTION
Cck	0.0593	Primary distinguishing gene between distinct interneuron classes identified in the cortex and hippocampus (Jasnow et al. 2009)
Neurod6 mutation	0.0583	General regulator of nervous system development whose induced displays different effects in neocortex vs. the hippocampal region (Bormuth et al. 2013)
Stmn3	0.0573	Up-expressed in hippocampus of patients with depressive disorders (Oh et al. 2010)
Plp1	0.0570	An oligodendrocyte- and myelin-related gene which exhibits cortical differential expression in schizophrenia (Wu et al. 2012)
Crym	0.0550	Plays a role in neuronal specification (Molyneaux et al. 2007)
Spink8	0.0536	Serine protease inhibitor specific to hippocampal pyramidal cells (Zeisel et al. 2015)
Gap43	0.0511	Encodes plasticity protein important for axonal regeneration and neural growth
Cryab	0.0500	Stress induction leads to reduced expression in the mouse hippocampus (Hagemann et al. 2012)
Mal	0.0494	Regulates dendritic morphology and is expressed at lower levels in cortex than in hippocampus (Shiota et al. 2006)
Tspan13	0.0488	Membrane protein which mediates signal transduction events in cell development, activation, growth and motility

Table 2.1: Genes with the greatest weight in the projection $\widehat{\beta}$ produced by SPARDA analysis of the mouse brain single cell RNA-seq data. Where not cited, the description of the genes are taken from the standard ontology annotations.

between the cell types is not a shift in the distributions, but rather the movement of a large fraction of low (1-2.5 log-FPKM) expression cells into the high-expression ($>$ 2.5 log-FPKM) regime. As this type of change does not match the restrictive assumptions of LIMMA's t-test, the method fails to highly-rank this gene while the Wasserstein distance employed by SPARDA is perfectly suited for measuring this sort of effect.

2.6.1 Identifying genes with differential interactions

Finally, we also apply SPARDA to normalized expression data with mean-zero & unit-variance marginal distributions. Since this removes nearly all of the difference between the two populations in terms of any single gene's expression, this explicitly restricts our search to genes whose relationship with other genes' expression is different between hippocampus and cortex cells. After restricting our analysis to only the top

gene ontology term		category, level	set size	candidates contained	p-value	q-value
GO:0019226	transmission of nerve impulse	BP 4	490	18 (3.7%)	5.46e-11	1.31e-08
GO:0007268	synaptic transmission	BP 4	391	13 (3.3%)	1.04e-07	1.02e-05
GO:0055082	cellular chemical homeostasis	BP 4	632	16 (2.5%)	1.27e-07	1.02e-05
GO:0032051	clathrin light chain binding	MF 4	3	3 (100.0%)	1.33e-07	4.66e-06
GO:0048666	neuron development	BP 4	646	16 (2.5%)	1.87e-07	1.09e-05
GO:0022008	neurogenesis	BP 4	1029	20 (1.9%)	2.28e-07	1.09e-05
GO:0032846	positive regulation of homeostatic process	BP 4	57	6 (10.5%)	4.72e-07	1.89e-05
GO:0048878	chemical homeostasis	BP 4	838	17 (2.0%)	1.12e-06	3.82e-05
GO:0007399	nervous system development	BP 4	1486	23 (1.6%)	1.31e-06	3.93e-05
GO:0030182	neuron differentiation	BP 4	854	17 (2.0%)	1.57e-06	4.18e-05
GO:0031175	neuron projection development	BP 4	529	13 (2.5%)	3.21e-06	7.7e-05
GO:0051969	regulation of transmission of nerve impulse	BP 4	194	8 (4.1%)	7.32e-06	0.00016
GO:0048858	cell projection morphogenesis	BP 4	516	12 (2.3%)	1.37e-05	0.000275
GO:0032990	cell part morphogenesis	BP 4	542	12 (2.2%)	2.19e-05	0.000405
GO:0007010	cytoskeleton organization	BP 4	763	14 (1.8%)	3.33e-05	0.000571
GO:0048168	regulation of neuronal synaptic plasticity	BP 4	38	4 (10.5%)	4.29e-05	0.000686
GO:0000902	cell morphogenesis	BP 4	814	14 (1.7%)	6.91e-05	0.00093
GO:0050877	neurological system process	BP 4	2024	24 (1.2%)	6.97e-05	0.00093
GO:0044057	regulation of system process	BP 4	427	10 (2.3%)	7.09e-05	0.00093
GO:0008366	axon ensheathment	BP 4	84	5 (6.0%)	7.36e-05	0.00093
GO:0008344	adult locomotory behavior	BP 4	86	5 (5.8%)	8.23e-05	0.000988
GO:0007611	learning or memory	BP 4	151	6 (4.0%)	0.000131	0.0015
GO:0006900	membrane budding	BP 4	21	3 (14.3%)	0.000165	0.0018
GO:0071822	protein complex subunit organization	BP 4	900	14 (1.6%)	0.000192	0.00201
GO:0001662	behavioral fear response	BP 4	27	3 (11.1%)	0.000356	0.00341
GO:0002209	behavioral defense response	BP 4	27	3 (11.1%)	0.000356	0.00341
GO:0030913	paranodal junction assembly	BP 4	6	2 (33.3%)	0.00039	0.0036
GO:0007626	locomotory behavior	BP 4	188	6 (3.2%)	0.000405	0.0036

Figure 2-4: Biological process terms most significantly enriched in the annotations of the top 100 genes identified by SPARDA.

500 genes with largest initial average expression (since genes playing important roles in interactions should be nontrivially expressed), we normalize each gene's expression values to have mean zero and unit variance within in the cells of each class.

Subsequent application of SPARDA reveals that most of the genes corresponding to the ten greatest values of the resulting $\widehat{\beta}$ are known to play important roles in in signaling and regulation (see Table 2.2). This analysis reveals many genes known to be heavily involved in signaling, regulating important processes, and other forms of functional interaction between genes. These types of important changes cannot

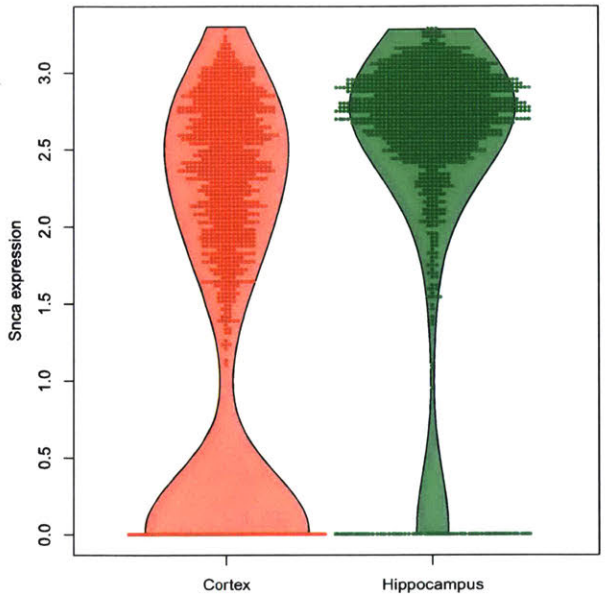


Figure 2-5: Distribution of *Snca* expression across cells of the somatosensory cortex and hippocampus.

be detected by standard differential expression analyses which consider each gene in isolation or require gene-sets to be explicitly identified as features (Geiler-Samerotte et al. 2013).

GENE	WEIGHT	DESCRIPTION
Thy1	0.1245	Plays a role in cell-cell & cell-ligand interactions during synaptogenesis and other processes in the brain
Vsnl1	0.1245	Modulates intracellular signaling pathways of nervous system
Stmn3	0.1222	Stathmins form important protein complex with tubulins
Stmn2	0.1188	Note: Tubulins Tubb3 and Tubb2 are ranked 20^{th} and 25^{th} by weight in $\widehat{\beta}$
Tmem59	0.1176	Fundamental regulator of neural cell differentiation. Knock out in the hippocampus results in drastic expression changes of many other genes (Zhang et al. 2011)
Basp1	0.1171	Transcriptional cofactor which can divert the differentiation of cells to a neuronal-like morphology (Goodfellow et al. 2011)
Snhg1	0.1166	Unclassified non-coding RNA gene
Mllt11	0.1145	Promoter of neurodifferentiation and axonal/dendritic maintenance (Lederer et al. 2007)
Uchl1	0.1137	Loss of function leads to profound degeneration of motor neurons (Jara et al. 2015).
Cck	0.1131	Targets pyramidal neurons and enables neocortical plasticity allowing for example the auditory cortex to detect light stimuli (Li et al. 2014, Gallopin et al. 2006)

Table 2.2: Genes with the greatest weight in the projection $\widehat{\beta}$ produced by SPARDA analysis of the marginally normalized expression data.

Chapter 3

Modeling persistent trends in distributions

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, a numerical/ordinal value whose effects on the observations is the item of interest. The batch label is also referred to as the *covariate*, which may be of random or fixed design. 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 chapter 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, 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 §3.3, a trend describes a sequence of distributions where the p^{th} 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 application: the analysis of single-cell RNA-sequencing time course data (see §3.12 for an alternative application to income data). One promising experimental design made feasible by the advent of scRNA-seq technology involves sampling groups of cells at various times from tissues / cell-cultures undergoing development and measuring transcriptome-wide gene expression within each individual 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 $N_\ell \geq 1$ 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 ($1 \leq s \leq n_i$), 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 differ between equidistant times. Thus, changes in the expression of genes regulating different 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 predicting 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 3-1. 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

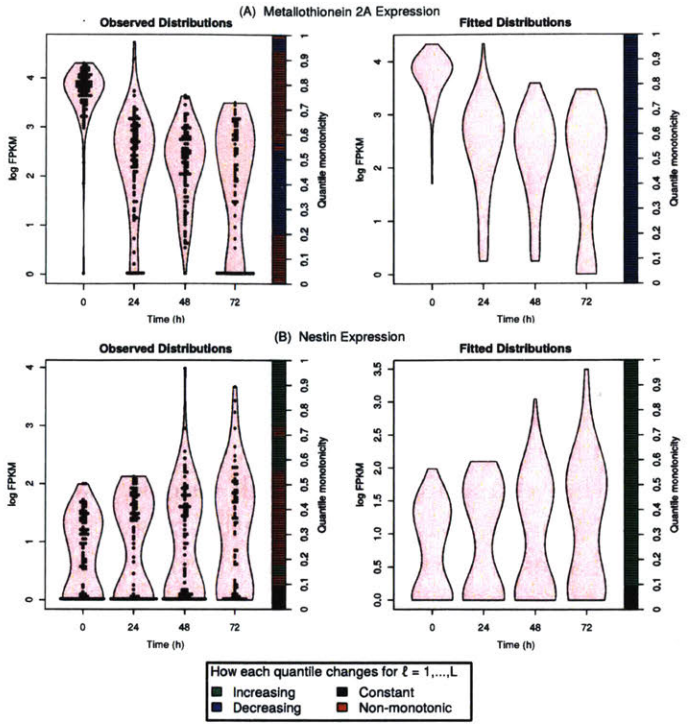


Figure 3-1: Violin plots (kernel density estimates) depicting the empirical distribution of known developmental genes' expression measured in myoblast cells (on left), and the corresponding TRENDS fitted distributions (on right). Each point shows a sampled cell.

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 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, as discussed in §1.1, reducing a cell population to a crude summary statistic may be highly misleading, because cell populations can exhibit enormous heterogeneity, particularly in developmental or in vivo settings, and transcript levels can vary immensely between seemingly equivalent cells (Geiler-Samerotte et al. 2013). 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 on their presumed developmental relevance or employ hypothesis testing to determine whether observed temporal variation in expression is biologically

relevant.

3.1 Related work

To better motivate the ideas subsequently presented in this chapter, 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, we wish to model the full (conditional) distribution of values, which is critical in the case of a highly 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 §3.10) 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 $L \leq 10$), 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.

3.2 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 $\ell_i \in \{1, \ldots, L\}$ is an ordinal-valued label associated with the ith batch, for which we have univariate empirical distribution $\widehat{P}_i$. 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 §3.3). 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. Since we do not believe exogenous effects unrelated to ℓ-progression are likely to follow a trend over ℓ, we can use this presumption to denoise any spurious variation. This is achieved by identifying the sequence of trending distributions which best models the variation in $\{\widehat{P}_{\ell_i}\}_{i=1}^N$ and subsequently concluding that changes in this trending sequence reflect the ℓ-progression-related forces affecting P_ℓ.

3.2.1 Use of the Wasserstein distance

TRENDS employs the Wasserstein distance to measure divergence between distributions. 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, Lemma 1 in §1.1 states that the L_q-Wasserstein distance is simply the L_q distance between quantile functions given by:

$$d_{L_q}(P,Q) = \left(\int_0^1 \left| F^{-1}(p) - G^{-1}(p) \right|^q \, \mathrm{d}p \right)^{1/q} \tag{3.1}$$

where F, G are the CDFs of P, Q and F^{-1}, G^{-1} are the corresponding *quantile* functions. Slightly abusing notation, we use $d_{L_q}(\cdot,\cdot)$ in this chapter to denote both Wasserstein distances between distributions or the corresponding quantile functions' L_q-distance (both $q = 1, 2$ are used in our 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 4. Calling this average the *Wasserstein mean*, we note its implicit use in the popular quantile normalization technique (Bolstad et al. 2003).

Lemma 4. *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:*

$$\overline{\mathbf{F}}^{-1} := \frac{1}{N} \sum_{i=1}^N F_i^{-1} = \underset{G^{-1} \in \mathcal{Q}}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \int_0^1 \left(F_i^{-1}(p) - G^{-1}(p) \right)^2 \mathrm{d}p \right\} \tag{3.2}$$

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:

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

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

3.3 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.

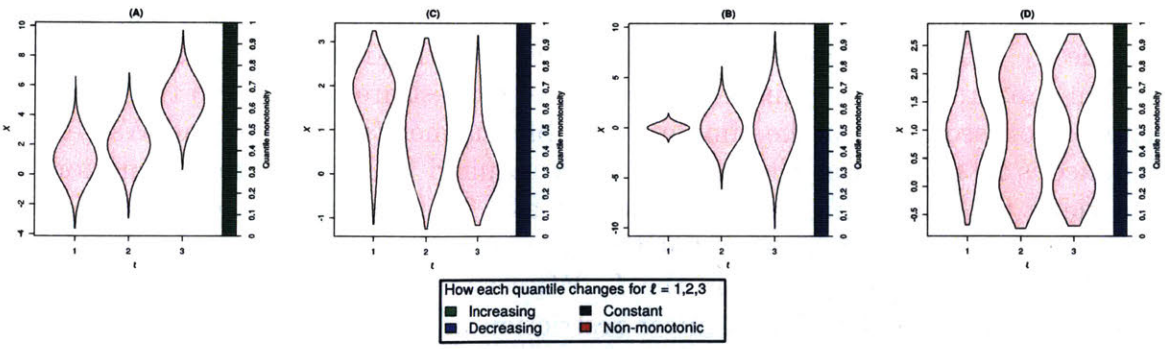


Figure 3-2: Violin plots depicting four different sequences of distributions which follow a trend. The pth rectangle in the color bar on the righthand side indicates the monotonicity of the pth quantile over the sequence of distributions (for $p = 0.01, 0.02, \ldots, 0.99$).

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 chapter is intended for relatively short sequences and does not include cyclic/seasonal patterns studied in time-series modeling.

Lemma 5. *If distributions $P_1, \ldots, P_L$ follow a trend, then*

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

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

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

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. □

Measuring how much the distributions are perturbed between each pair of levels via the L_1 Wasserstein metric, Lemma 5 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 5 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 3-3C). 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 3-3C). 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 §3.6) 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 3-2 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 §3.3.1 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.

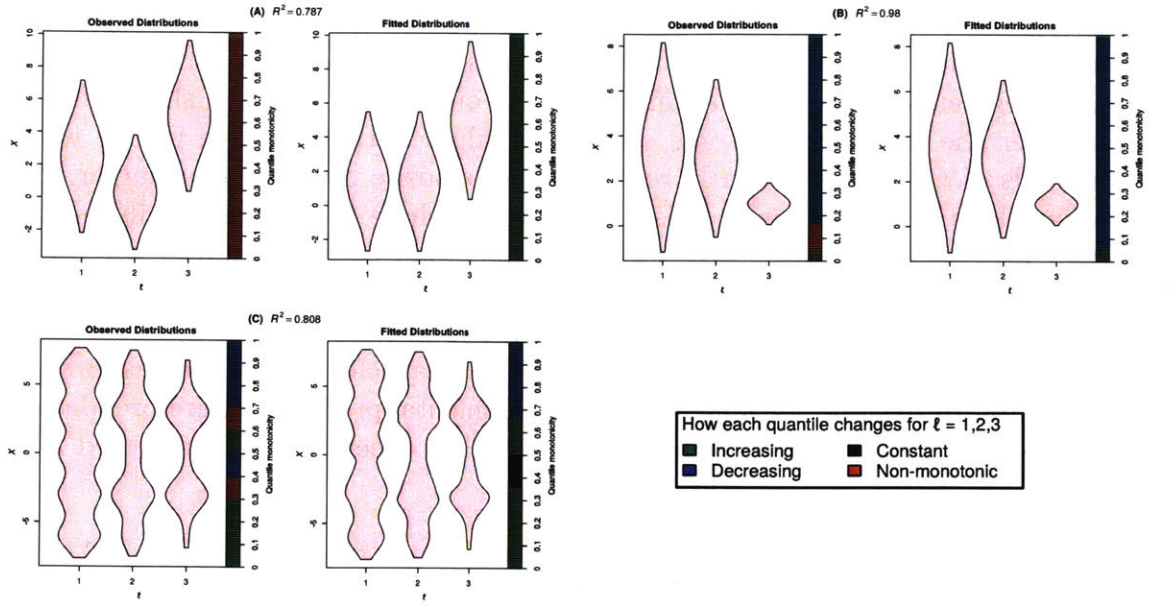


Figure 3-3: Violin plots depicting sequences of distributions which do *not* follow a trend (Observed Distributions in lefthand panels). Shown to the right of each example are the corresponding fitted distributions estimated by TRENDS (with the TRENDS R^2 value).

The lefthand panels of Figure 3-3 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 (3.6)), each distribution of which corresponds to our estimate of Q_ℓ (introduced in §4.4). 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 §3.5). 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 3-3A (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 3-3B 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 §3.3.1. The trend criterion fits our a priori knowledge well, while remaining flexible with respect to the precise nature of expression changes.

3.3.1 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 $F_1(x) \geq F_2(x)\ \forall x$ (equivalently characterized as $\Pr(X_1 > x) \leq \Pr(X_2 > x)\ \forall x$) (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.

$$\forall \ell:\ F_\ell^{-1} = \omega_\ell F_1^{-1} + (1-\omega_\ell)F_L^{-1} \quad \text{s.t. } \{\omega_\ell \in [0,1] : \ell = 1,\ldots,L\} \text{ form a monotonic sequence} \tag{3.3}$$

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

$$\forall \ell \in \{1,\ldots,L\}:\ F_\ell = \sum_{k=1}^{K} \pi_\ell^{(k)} G_k \quad \text{where } \forall\ k,\ell:\ \pi_\ell^{(k)} \in [0,1]\ ,\ \pi_\ell^{(K)} = 1 - \sum_{k=1}^{K-1} \pi_\ell^{(k)} \tag{3.4}$$

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 $(\ell - 1) \to \ell$ 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 3-2D depicts an example of an evolving 3-component mixture model which follows a trend.

3.4 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:

$$F_i^{-1} = G_{\ell_i}^{-1} + \mathcal{E}_i \text{ such that } G_1^{-1}, \ldots, G_L^{-1} \text{ follow a trend, and the following hold:} \tag{3.5}$$

(A7) $\mathcal{E}_i : (0, 1) \to \mathbb{R}$ is constrained so that $G_{\ell_i}^{-1}$ and F_i^{-1} are valid quantile functions.

(A8) 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 $\Pr(|\mathcal{E}_i(p)| > t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$ for any $t > 0$.

(A9) 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 (A7) 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 (A9) 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 (A9) 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 (A8) 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 (3.5) 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 (3.5), an analogous f maps each ordinal level {1,..., 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 (3.5). 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:

$$\widehat{Q}_1, \ldots, \widehat{Q}_L = \underset{Q_1,\ldots,Q_L}{\operatorname{argmin}} \left\{ \sum_{\ell=1}^{L} \sum_{i \in I_\ell} d_{L_2}(Q_\ell, \widehat{P}_i)^2 \right\} \text{ where } Q_1, \ldots, Q_L \text{ follow a trend} \tag{3.6}$$

where I_ℓ is the set of batch-indices i such that $\ell_i = \ell$, and we require $N_\ell := |I_\ell| \geq 1$ 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 3-3 shows some examples of fitted distributions produced by TRENDS regression. The objective in (3.6) 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. In §3.6, we introduce an efficient algorithm that is always guaranteed to produce the optimal Wasserstein-least-squares fit.

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.

3.5 Measuring goodness of 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 $\widehat{Q}_1, \ldots, \widehat{Q}_L$:

$$R^2 := 1 - \left(\frac{1}{N} \sum_{i=1}^{N} d_{L_2}(\widehat{Q}_{\ell_i}, \widehat{P}_i)^2 \right) \Bigg/ \left(\frac{1}{N} \sum_{i=1}^{N} d_{L_2}(\widehat{P}_i, \overline{\mathbf{F}}^{-1})^2 \right) \quad \in [0, 1] \tag{3.7}$$

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 4, the numerator and denominator in (3.7) 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 (3.5) in the $\widehat{P}_i$ over the full ordinal progression $\ell = 1, \ldots, L$ using:

$$\Delta := \frac{1}{L} \cdot d_{L_1}(\widehat{Q}_1, \widehat{Q}_L) \tag{3.8}$$

The L_1 Wasserstein distance is a natural choice, since by Lemma 5, 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 3-3B, 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.

Additionally, 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 (3.7). Due to the quantile-noise functions $\mathcal{E}_i(\cdot)$ assumed in our model (3.5), 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. Statistical accuracy and computational burden can be traded off by choosing an appropriate number of permutations. We note that within each permutation of the data, the Wasserstein-least-squares fit can be computed very efficiently in practice (as detailed in §3.6). Unfortunately, N and L may be small in some applications, which undesirably limits the number of possible label-permutations. In §3.5.1, 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). §3.9.2 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 §3.9.2 also empirically demonstrates that despite its restrictive assumptions, the TRENDS model can provide superior estimates of severely-misspecified effects than the initial empirical distributions.

3.5.1 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 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 $R_{\text{perm},k}^2$ value via (3.7) 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 $\{R_{\text{perm},k}^2\}$ 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 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.

3.6 Fitting the TRENDS model

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

$$\widehat{Q}_1, \ldots, \widehat{Q}_L = \arg\min_{Q_1,\ldots,Q_L} \left\{ \sum_{\ell=1}^{L} \sum_{i \in I_\ell} w_i \cdot d_{L_2}(Q_\ell, \widehat{P}_i)^2 \right\} \text{ where } Q_1, \ldots, Q_L \text{ follow a trend} \tag{3.9}$$

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 (3.6) with the weighted sum in (3.9). 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 (3.1), 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 (3.9). Defining $p_0 := 2p_1 - p_2$ and $p_P := 2p_{P-1} - p_{P-2}$, we employ the following midpoint-approximation of the integral

$$\min_{G_1^{-1},\ldots,G_L^{-1}} \left\{ \sum_{\ell=1}^{L} \sum_{i \in I_\ell} w_i \sum_{k=1}^{P-1} \left(\widehat{F}_i^{-1}(p_k) - G_\ell^{-1}(p_k) \right)^2 \left[\frac{p_{k+1} - p_{k-1}}{2} \right] \right\}$$
$$\text{where } G_1, \ldots, G_L \text{ must follow a trend} \tag{3.10}$$

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 (3.10). 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).

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

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 $y_\ell \geq \widehat{y}_{\ell-1}$, 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 $\ell = L$.

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 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 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 $\delta \in$ {"nonincreasing", "nondecreasing"} that determines which monotonic best-fitting sequence to find.

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 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. Relying on auxiliary lemmas presented in

Trend-Fitting Algorithm: Numerically solves (3.9) by optimizing (3.10)

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. $\widehat{F}_i^{-1}(p_k) := \textbf{quantile}(\widehat{F}_i, p_k)$ 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. $x_\ell[k] := \frac{1}{w_\ell^*} \sum_{i \in I_\ell} w_i \widehat{F}_i^{-1}(p_k)$ 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. $y_1, \ldots, y_L := \textbf{AlternatingProjections}\left(x_1, \ldots, x_L \,;\, \delta \,;\, \{w^*\}_{\ell=1}^L, \{p_k\}_{k=1}^{P-1}\right)$
7. $W[\delta] :=$ the value of (3.10) evaluated with $G_\ell^{-1}(p_k) = y_\ell[k] \quad \forall \ell, k$
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 $\min_\delta W[\delta]$ and return $\widehat{G}_\ell^{-1}(p_k) = y_\ell^*[k] \quad \forall \ell, k$ 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 $\forall \ell, k: \; y_\ell^{(t)}[k] \le y_\ell^{(t)}[k+1]$ and the sequence $y_1^{(t)}[k], \ldots, y_L^{(t)}[k]$ is monotone nonincreasing/nondecreasing as specified by $\delta[k]$, provided that $x_\ell^{(0)}[k] \le x_\ell^{(0)}[k+1]$ 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. $y_1^{(t)}[k], \ldots, y_L^{(t)}[k] := \textbf{PAVA}\left(\left(x_1^{(t)}[k] + r_1^{(t)}[k], w_1^*\right), \ldots, \left(x_L^{(t)}[k] + r_L^{(t)}[k], w_L^*\right); \delta[k]\right)$ 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. $\forall \ell \in \{1, \ldots, L\}: \quad x_\ell^{(t+1)}[1], \ldots, x_\ell^{(t+1)}[P-1] :=$ $\textbf{PAVA}\left(\left(y_\ell^{(t)}[1] + s_\ell^{(t)}[1], \frac{p_2 - p_0}{2}\right), \ldots, \left(y_\ell^{(t)}[P-1] + s_\ell^{(t)}[P-1], \frac{p_P - p_{P-2}}{2}\right); \text{"nondecreasing"}\right)$
6. $s_\ell^{(t+1)}[k] := y_\ell^{(t)}[k] + s_\ell^{(t)}[k] - x_\ell^{(t+1)}[k]$ for each ℓ, k

§3.8, our proof of Theorem 5 provides much intuition on the TF algorithm. 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.

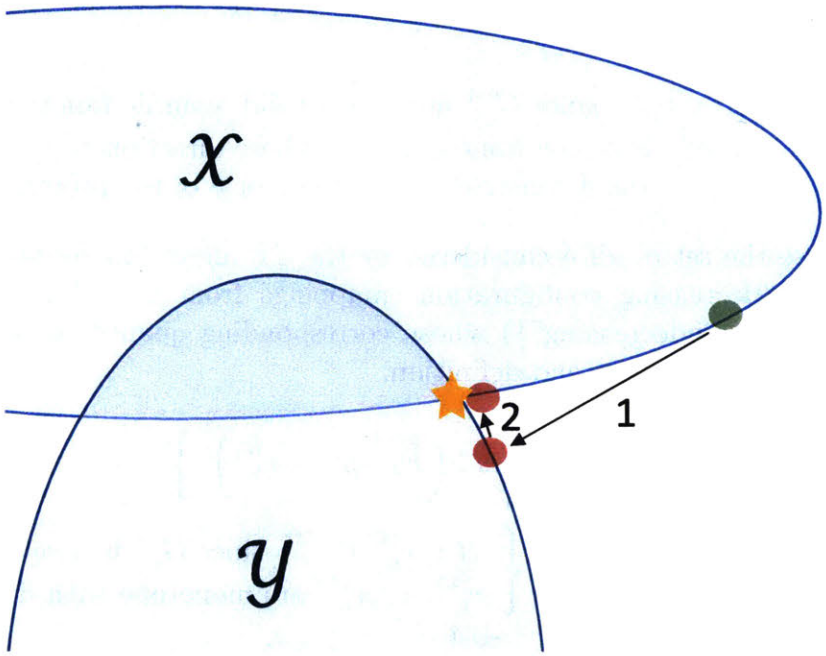


Figure 3-4: Visual example of the first two updates made by Dykstra's method of alternating projections to find the empirical Wasserstein-least-squares-fit. Each point depicts a $L \times (P-1)$ matrix, whose $(\ell, k)^{\text{th}}$ entry is supposed to numerically represent the p_k^{th} quantile of the ℓ^{th} distribution. $\mathcal{X}$ is the closed/convex set of matrices whose columns are nondecreasing (representing valid quantiles of a probability distribution), and $\mathcal{Y}$ is the closed/convex set of matrices whose rows are monotonic over ℓ and satisfy the trend criterion for each p_k^{th} quantile. Given some initial matrix depicted in green, whose columns contain the empirical quantiles of each observed distribution $\widehat{P}_\ell$, our goal is to find the closest matrix $\mathbf{A} \in \mathcal{X} \cap \mathcal{Y}$ (depicted by the star). The columns of $\mathbf{A}$ will thus be valid quantile functions of a sequence of distributions that follow a trend, ensuring $\mathbf{A}$ represents the Wasserstein-least-squares-fit (as distances between quantile functions correspond to Wasserstein distances between distributions).

Theorem 5. *The Trend-Fitting algorithm produces valid quantile-functions* $\widehat{G}_1^{-1}, \ldots, \widehat{G}_L^{-1}$ *which optimally solve the numerical version of the TRENDS objective given in (3.10).*

Proof. We have:

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

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

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

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

s.t. $\forall\ k < k' \in \{1, \ldots, P-1\}$:

$$\begin{cases} \forall \ell : \ v_k^{(\ell)} < v_{k'}^{(\ell)} & \text{since } G_\ell^{-1} \text{ must be a valid quantile function} \\ v_k^{(1)}, \ldots, v_k^{(L)} & \text{is a monotone sequence whose direction} = \delta[k] \text{ for one of} \\ & \text{the } \delta \text{ constructed in Step 6 or 8 of the procedure.} \end{cases}$$

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

$$= \underset{v^{(1)},\ldots,v^{(L)}}{\operatorname{argmin}} \left\{ \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2} \right) \sum_{\ell=1}^{L} w_\ell^* \left(\overline{\widehat{F}}_\ell^{-1}(p_k) - v_k^{(\ell)} \right)^2 \right\} \tag{3.11}$$

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

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

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

$$d_W\left([v^{(1)},\ldots,v^{(L)}],[w^{(1)},\ldots,w^{(L)}]\right)=\sum_{k=1}^{P-1}\left(\frac{p_{k+1}-p_{k-1}}{2}\right)\sum_{\ell=1}^{L}w_\ell^*\left(v_k^{(\ell)}-w_k^{(\ell)}\right)^2 \tag{3.12}$$

Lemmas 7 and 8 in §3.8 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 defined in (3.12)) exactly corresponds to the solution of the constrained optimization in (3.10) under the additional constraint imposed by δ. □

3.7 Theoretical results

Under the model given in (3.5), we establish some results regarding the statistical quality of the $\hat{Q}_1,\ldots,\hat{Q}_L$ estimates produced by the TF algorithm. The corresponding proofs are relegated to §3.8. To develop pragmatic theory, we use finite-sample bounds defined in terms of quantities encountered in practice rather than the true Wasserstein distance (3.1), 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 (3.10). In addition to the conditions of model (3.5), we make the following simplifications throughout for ease of exposition:

(A10) The number of batches at each level is the same, i.e. $N_\ell := N_1 = \cdots = N_L \geq 1$.

(A11) The same number of samples are drawn per batch, i.e. $n := n_i$ for all $1 \leq i \leq N$.

(A12) For $k = 1,\ldots,P-1$: the (k/P)th quantiles of each distribution are considered.

(A13) Uniform weights are employed, i.e. in (3.9): $w_i = 1$ for all i.

Theorem 6. *Under model (3.5) and additional conditions (A10)-(A13), 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:* $d_W(\widehat{G}_\ell^{-1}, G_\ell^{-1}) < \epsilon$ *for each* $\ell \in \{1, \ldots, L\}$

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

Thus, Theorem 6 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:

(A14) For $i = 1, \ldots, N$: quantiles are estimated from n i.i.d. samples $X_{1,i}, \ldots, X_{n,i} \sim P_i$.

(A15) 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 $k = 1, \ldots, P-1$.

(A16) The basic 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 7. *Under the assumptions of Theorem 6 and (A14)-(A16), 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:*

$$1 - 2PL\left[\exp\left(\frac{-\epsilon^2 N_\ell}{32\sigma^2 L}\right) + N_\ell \exp\left(-2n \cdot R\left(\frac{\epsilon}{4\sqrt{L}}\right)^2\right)\right] \tag{3.13}$$

where for $\gamma > 0$:

$$\begin{aligned} R(\gamma) &:= \min_{i,k}\{R(\gamma, i, k/P) : i = 1, \ldots, N, k = 1, \ldots, P-1\} \\ R(\gamma, i, p) &:= \min\left\{F_i\left(F_i^{-1}(p) + \gamma\right) - p\ ,\ p - F_i\left(F_i^{-1}(p) - \gamma\right)\right\} \end{aligned} \tag{3.14}$$

Theorem 7 is our most general result applying to arbitrary distributions P_i that satisfy basic condition (A15). 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 7 may be

significantly improved by instead adopting one of the following stronger assumptions:

(A17) The simple quantile-estimator defined in (A16) is used, and the support of each P_i is bounded and connected with non-neglible density. This implies that there exist constants $B, c > 0$ such that for each i:

$$f_i(x) = 0 \ \ \forall x \notin [-B, B] \ \text{ and } \ f_i(x) \geq c \ \ \forall x \in [-B, B]$$

where f_i is the density associated with CDF F_i.

(A18) 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 $\epsilon > 0$:

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

Theorem 8. *Under the assumptions of Theorem 6, conditions (A14), (A15), and one of either (A17) or (A18), the bound in (3.13) may be sharpened to ensure that for any $\epsilon > 0$:*

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

with probability greater than:

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

In Theorem 8, 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 7 assume use of the simple quantile-estimator from (A16), 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 (A17) is an example of a distribution and quantile-estimator combination which achieves the error required by (A18). Because the boundedness assumption is undesirably limiting, we also derive a similar result under weaker assumptions:

(A19) 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) $f_i(x) \geq c \ \forall x \in [-B, B]$ where f_i is the density function of CDF F_i.

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

(A20) Defining $r := \min\left\{2c^2\ ,\ \frac{2ab^2-1}{4PB^2}\right\}$, we have $r > 0$, or equivalently, $2ab^2 > 1$.

(A21) We avoid estimating extreme quantiles, i.e. $F_i^{-1}(k/P) \in (-B, B)$ for $k = 1, \ldots, P-1$.

Theorem 9. *Under the assumptions of Theorems 6 and 7 as well as conditions (A19)-(A21), the previous bound in (3.13) may be sharpened to ensure that for all $\epsilon > 0$:*

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

with probability greater than:

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

Theorem 9 again provides an exponential convergence 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 (A16) 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.

3.8 Auxiliary proofs and lemmas

Lemma 6 (de Leeuw (1977)). *Given weights $w_1, \ldots, w_N \geq 0$ and pairs $(\ell_1, y_1), \ldots, (\ell_N, y_N)$ where each $\ell \in \{1, \ldots, L\}$ appears at least once, the fitted values $\widehat{y}_1, \ldots, \widehat{y}_L$ produced by tertiary-variant of PAVA are guaranteed to be the best-fitting nondecreasing sequence in the least-squares sense, i.e.*

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

Lemma 7. *Recall the definitions from the TF algorithm and the proof of Theorem 5. Given any $[x^{(1)}, \ldots, x^{(L)}] \in \mathcal{X}$, its projection onto $\mathcal{Y}$ under metric d_W, $[y^{(1)}, \ldots, y^{(L)}]$,*

may be computed $\forall k \in \{1, \ldots, P-1\}$ *as*

$$y_k^{(1)}, \ldots, y_k^{(L)} = \mathbf{PAVA}\left((x_k^{(1)}, w*_1), \ldots, (x_k^{(L)}, w_L^*); \delta[k]\right)$$

Proof. Choose any $[z^{(1)}, \ldots, z^{(L)}] \in \mathcal{Y}$. By consequence of Lemma 6

$$\begin{aligned}
&\mathbf{PAVA}\left((x_k^{(1)}, w_1^*), \ldots, (x_k^{(L)}, w_L^*); \delta[k]\right) \\
&= \underset{\text{monotone } \lambda_1, \ldots, \lambda_L}{\operatorname{argmin}} \left\{ \sum_{\ell=1}^{L} w_\ell^* \left(x_k^{(\ell)} - \lambda_\ell\right)^2 \right\} \quad \text{where the } \lambda_\ell \text{ are only increasing if specified by } \delta[k] \\
\implies &\sum_{\ell=1}^{L} w_\ell^* \left(y_k^{(\ell)} - x_k^{(\ell)}\right)^2 \leq \sum_{\ell=1}^{L} w_\ell^* \left(z_k^{(\ell)} - x_k^{(\ell)}\right)^2 \quad \forall k \\
&\qquad \text{since } z_k^{(1)}, \ldots, z_k^{(L)} \text{ have monotonicity specified by } \delta \\
\implies &\sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \sum_{\ell=1}^{L} w_\ell^* \left(y_k^{(\ell)} - x_k^{(\ell)}\right)^2 \leq \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \sum_{\ell=1}^{L} w_\ell^* \left(z_k^{(\ell)} - x_k^{(\ell)}\right)^2
\end{aligned}$$

□

Lemma 8. *Recall the definitions from the TF algorithm and the proof of Theorem 5. 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*

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

Proof. Choose any $[z^{(1)}, \ldots, z^{(L)}] \in \mathcal{X}$. By Lemma 6:

$$\begin{aligned}
&\mathbf{PAVA}\left(\left(y_1^{(\ell)}, \frac{p_2 - p_0}{2}\right), \ldots, \left(y_{P-1}^{(\ell)}, \frac{p_P - p_{P-2}}{2}\right); \text{“nondecreasing”}\right) \\
&= \underset{\lambda_1 \leq \cdots \leq \lambda_{P-1}}{\operatorname{argmin}} \left\{ \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \left(y_k^{(\ell)} - \lambda_k\right)^2 \right\} \quad \text{for each } \ell \\
\implies &\sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \left(x_k^{(\ell)} - y_k^{(\ell)}\right)^2 \leq \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \left(z_k^{(\ell)} - y_k^{(\ell)}\right)^2 \quad \forall \ell \\
&\quad \text{since } [z^{(1)}, \ldots, z^{(L)}] \in \mathcal{X} \implies \forall \ell : z_1^{(\ell)} \leq \cdots \leq z_{P-1}^{(\ell)} \\
\implies &\sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \sum_{\ell=1}^{L} w_\ell^* \left(x_k^{(\ell)} - y_k^{(\ell)}\right)^2 \leq \sum_{k=1}^{P-1} \left(\frac{p_{k+1} - p_{k-1}}{2}\right) \sum_{\ell=1}^{L} w_\ell^* \left(x_k^{(\ell)} - z_k^{(\ell)}\right)^2
\end{aligned}$$

□

3.8.1 Proof of Theorem 6

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

$$\bar{F}_\ell^{-1}(p) := \frac{1}{N_\ell} \sum_{i \in I_\ell} F_i^{-1}(p) \tag{3.15}$$

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

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

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

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

$$\begin{aligned}
\Pr\left(\sum_{\ell=1}^{L} d_W\left(\bar{F}_\ell^{-1}, G_\ell^{-1}\right)^2 > \eta\right) &\leq \sum_{\ell=1}^{L}\sum_{k=1}^{P-1} \Pr\left(\frac{1}{P}\left(\bar{F}_\ell^{-1}(k/P) - G_\ell^{-1}(k/P)\right)^2 > \frac{\eta}{PL}\right) \\
&\qquad \text{by a union-bound} \\
&= L \cdot P \cdot \Pr\left(\left|\bar{F}_\ell^{-1}(k/P) - G_\ell^{-1}(k/P)\right| > \sqrt{\frac{\eta}{L}}\right) \\
&\leq 2PL\exp\left(-\frac{\eta N_\ell}{2\sigma^2 L}\right) \qquad (3.16)
\end{aligned}$$

Note that $\widehat{G}_1^{-1}, \ldots, \widehat{G}_L^{-1}$ form the best trending approximation to the F_i^{-1} by Theorem 5, and since $G_1^{-1}, \ldots, G_L^{-1}$ are valid quantile functions which also follow a trend, this implies:

$$\begin{aligned}
&\sum_{\ell=1}^{L}\sum_{i \in I_\ell} d_W\left(F_i^{-1}, \widehat{G}_\ell^{-1}\right)^2 \leq \sum_{\ell=1}^{L}\sum_{i \in I_\ell} d_W\left(F_i^{-1}, G_\ell^{-1}\right)^2 \\
\Rightarrow &\sum_{\ell=1}^{L} d_W\left(\bar{F}_\ell^{-1}, \widehat{G}_\ell^{-1}\right)^2 \leq \sum_{\ell=1}^{L} d_W\left(\bar{F}_\ell^{-1}, G_\ell^{-1}\right)^2 \quad \text{by Lemma 4}
\end{aligned}$$

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

Thus, by the triangle-inequality:

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

which implies $\forall \epsilon > 0$ we can combine this result with (3.16) setting $\eta := \epsilon^2/4$ to get:

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

□

3.8.2 Proof of Theorem 7

Proof. We proceed similarly as in the proof of Theorem 6. Defining

$$\overline{\widehat{F}}_\ell^{-1}(p) := \frac{1}{N_\ell} \sum_{i \in I_\ell} \widehat{F}_i^{-1}(p) \tag{3.17}$$

by Theorem 3.10 and Lemma 4, we have:

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

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

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

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

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

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

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

Therefore, for all $\epsilon > 0$:

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

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

and we can use (3.16) to bound the first summand, resulting in the following bound

$$\Pr\left(\exists \ell : d_W\left(\widehat{G}_\ell^{-1}, G_\ell^{-1}\right) > \epsilon\right) \leq 2PL\exp\left(\frac{-\epsilon^2 N_\ell}{32\sigma^2 L}\right) + \Pr\left(\sum_{\ell=1}^{L} d_W\left(\overline{\widehat{F}}_\ell^{-1}, \bar{F}_\ell^{-1}\right)^2 > \frac{\epsilon^2}{16}\right) \tag{3.18}$$

Finally, Lemma 10 implies:

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

which produces the desired bound when combined with (3.18). □

3.8.3 Proof of Theorem 8

Proof. By Lemma 11, (A17) ⇒ (A18), so we only need to show the result assuming (A18) holds. Lemma 12 then implies:

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

Note that the bound in (3.18) only requires the assumptions from Theorem 6, so we can combine it with the above expression to obtain the desired bound. □

3.8.4 Proof of Theorem 9

Proof.

$$\begin{aligned} \text{Consider} \quad & \Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) \\ & = \Pr\left(\widehat{F}_i\left(F_i^{-1}(k/P) + \epsilon\right) \leq \frac{k}{P}\right) \end{aligned}$$

$$= \Pr\left(\sum_{j=1}^{n} \mathbb{1}\left[X_{i,j} \leq F_i^{-1}(k/P) + \epsilon\right] \leq \frac{nk}{P}\right) \tag{3.19}$$

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 $\epsilon + F_i^{-1}(k/P) \geq B > 0$, which implies $n\widetilde{p} \geq \widetilde{x}$.
Letting $D(\alpha \mid\mid \beta)$ 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 (3.19) by

$$\leq \exp\left(-nD\left(\frac{\widetilde{x}}{n} \mid\mid \widetilde{p}\right)\right)$$

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

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

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

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

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

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

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

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

because $1 - k/P > 0$ and by (A19): $F_i\left(F_i^{-1}(k/P) + \epsilon\right) \geq z$
since we've assumed $F_i^{-1}(k/P) + \epsilon \geq B$

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

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

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

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

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

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

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

$$\leq \exp\left(-n\left(\frac{2a\left(1 - \frac{k}{P}\right)b^2 - 1}{4B^2}\right)\epsilon^2\right)$$ because by (A19) and (A21):

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

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

$$\Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) \leq \exp\left(-n\left(\frac{2ab^2 - 1}{4PB^2}\right)\epsilon^2\right) \tag{3.20}$$

Following the proof of Lemma 11, one can show that (A19) implies

$$\Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) \leq \exp(-2nc^2\epsilon^2) \text{ if } 0 < \epsilon < B - F_i^{-1}(k/P) \tag{3.21}$$

Combining (3.21) with (3.20), we thus have

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

where $r := \min\left\{2c^2\ ,\ \frac{2ab^2-1}{4PB^2}\right\} > 0$ by (A20).

One can show by an identical argument that

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

and therefore

$$\Pr\left(\left|\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P)\right| > \epsilon\right) \leq 2\exp\left(-nr\epsilon^2\right) \quad \forall \epsilon > 0 \tag{3.22}$$

$\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P)$ 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 (A16) is used. Following the proof of Lemma 12, $\forall \gamma > 0$:

$$\Pr\left(\sum_{\ell=1}^{L} d_W\left(\overline{\widehat{F}}_\ell^{-1}, \bar{F}_\ell^{-1}\right)^2 > \frac{\epsilon^2}{16}\right) \leq 2P \exp\left(-\frac{r}{16} N_\ell \, n \, \epsilon^2\right) \tag{3.23}$$

Note that the bound in (3.18) only requires the assumptions from Theorem 6, so we can combine it with the above inequality to obtain the desired bound. □

Lemma 9 (Serfling (1980): Theorem 2.3.2). *For $p \in (0,1)$: if there 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 for all $\gamma > 0$:*

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

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 10. *Under the assumptions of Theorem 7 and definitions (3.14), (3.15), (3.17)*

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

Proof of Lemma 10.

$$\begin{aligned}
&\Pr\left(\sum_{\ell=1}^{L} d_W\left(\overline{\widehat{F}}_\ell^{-1}, \bar{F}_\ell^{-1}\right)^2 > \gamma\right) \\
=&\Pr\left(\sum_{\ell=1}^{L} \frac{1}{N_\ell} \sum_{i \in I_\ell} \sum_{k=1}^{P-1} \frac{1}{P}\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P)\right)^2 > \gamma\right) \\
\leq& N_\ell L \sum_{k=1}^{P-1} \Pr\left(\left|\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P)\right| > \sqrt{\frac{\gamma}{L}}\right) \quad \text{by the union-bound} \\
\leq& 2N_\ell L \sum_{k=1}^{P-1} \exp\left(-2nR\left(\sqrt{\gamma/L}, i, k/P\right)^2\right) \quad \text{by (A15) and Lemma 9} \\
\leq& 2N_\ell L P \exp\left(-2nR\left(\sqrt{\gamma/L}\right)^2\right) \quad \text{by definition (3.14)}
\end{aligned}$$

□

Lemma 11. *If we assume (A14) and (A15), then condition (A17) implies condition (A18).*

Proof of Lemma 11. Assume WLOG that $F_i^{-1}(k/P) \geq 0$ and note that $F_i^{-1}(k/P) \leq B$ by (A17).
Then, by a bound established in the proof of Lemma 9 given in (Serfling 1980), $\forall \epsilon > 0$:

$$\Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) \leq \exp\left(-2n\, R(\epsilon, i, k/P)^2\right) \tag{3.24}$$

and

$$\Pr\left(F_i^{-1}(k/P) - \widehat{F}_i^{-1}(k/P) > \epsilon\right) \leq \exp\left(-2n\, R(\epsilon, i, k/P)^2\right) \tag{3.25}$$

By (A17): $f_i(x) = \frac{d}{dx}F_i(x) \geq c\ \forall x \in (-B, B)$ which implies

$$R(\gamma, i, p) \geq c\gamma > 0 \ \text{ if } F_i^{-1}(p) \pm \gamma \in (-B, B) \tag{3.26}$$

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 (3.26), (3.24) and (3.25) imply

$$\Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) \leq \exp(-2nc^2\epsilon^2) \ \text{ if } F_i^{-1}(k/P) + \epsilon < B \tag{3.27}$$

and

$$\Pr\left(F_i^{-1}(k/P) - \widehat{F}_i^{-1}(k/P) > \epsilon\right) \leq \exp(-2nc^2\epsilon^2) \ \text{ if } F_i^{-1}(k/P) - \epsilon > -B \tag{3.28}$$

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

$$\begin{aligned} &\Pr\left(\widehat{F}_i^{-1}(k/P) > F_i^{-1}(k/P) + \epsilon\right) = 0 \quad \text{if } \epsilon \geq B - F_i^{-1}(k/P) \\ \Longrightarrow\ &\Pr\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P) > \epsilon\right) = 0 \quad \text{if } \epsilon \geq B - F_i^{-1}(k/P) \end{aligned} \tag{3.29}$$

as well as

$$\begin{aligned} &\Pr\left(\widehat{F}_i^{-1}(k/P) < F_i^{-1}(k/P) - \epsilon\right) = 0 \quad \text{if } \epsilon \geq B + F_i^{-1}(k/P) \\ \Longrightarrow\ &\Pr\left(F_i^{-1}(k/P) - \widehat{F}_i^{-1}(k/P) > \epsilon\right) = 0 \quad \text{if } \epsilon \geq B + F_i^{-1}(k/P) \end{aligned} \tag{3.30}$$

Putting together (3.27), (3.28), (3.29), and (3.30), we thus have

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

and

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

which implies

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

□

Lemma 12. *Under condition (A18) and definitions (3.14), (3.15), (3.17)*

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

Proof of Lemma 12.

$$\begin{aligned}
&\Pr\left(\sum_{\ell=1}^{L} d_W\left(\overline{\widehat{F}}_\ell^{-1}, \bar{F}_\ell^{-1}\right)^2 > \gamma\right) \\
=&\Pr\left(\frac{1}{LN_\ell}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\sum_{k=1}^{P-1}\frac{1}{P}\left(\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P)\right)^2 > \frac{\gamma}{L}\right) \\
\leq&\sum_{k=1}^{P-1}\Pr\left(\left|\frac{1}{LN_\ell}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\widehat{F}_i^{-1}(k/P) - F_i^{-1}(k/P)\right| > \sqrt{\frac{\gamma}{L}}\right) \quad \text{by the union-bound} \\
\leq&2\sum_{k=1}^{P-1}\exp\left(-2nc^2 LN_\ell\sqrt{\frac{\gamma}{L}}^2\right) = 2P\exp\left(-2nc^2 N_\ell\gamma\right)
\end{aligned}$$

where in the last inequality, we have used the fact that (A18) 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). □

3.9 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 scRNA-seq models 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 $L = 5$:

(S$_1$) $Q_\ell \sim \text{NB}(r_\ell, p_\ell)$ 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 $\text{NB}(r = 5, p = 0.3)$ and $\text{NB}(r = 5, p = 0.7)$ 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$_3$) $Q_\ell \sim \text{NB}(r = 5, p = 0.5)$ 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). 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 capture various types of noise affecting scRNA-seq measurements (e.g. dropout, PCR amplification bias, transcriptional bursting), noise for the i^{th} 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 $\widetilde{r}_\ell = r_\ell + r_{\text{noise}}$ and $\widetilde{p}_\ell = p_\ell + p_{\text{noise}}$. Here, p_{noise}, r_{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}. In order to ensure we are sampling from valid distributions after the introduction of noise, we subsequently enforce the following additional constraints: $\widetilde{r}_\ell \geq 1$, $0.05 \leq \widetilde{p}_\ell \leq 0.95$ before drawing our observations. 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 §3.11) 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 3-5 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 §4.7) and shows that TRENDS can produce a much better picture of the underlying distributions than the (noisy) observed empirical distributions. As shown in Figure 3-5A, 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.

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 §3.5.1). In this analysis, we also apply the following alternative methods (detailed in §3.10): 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

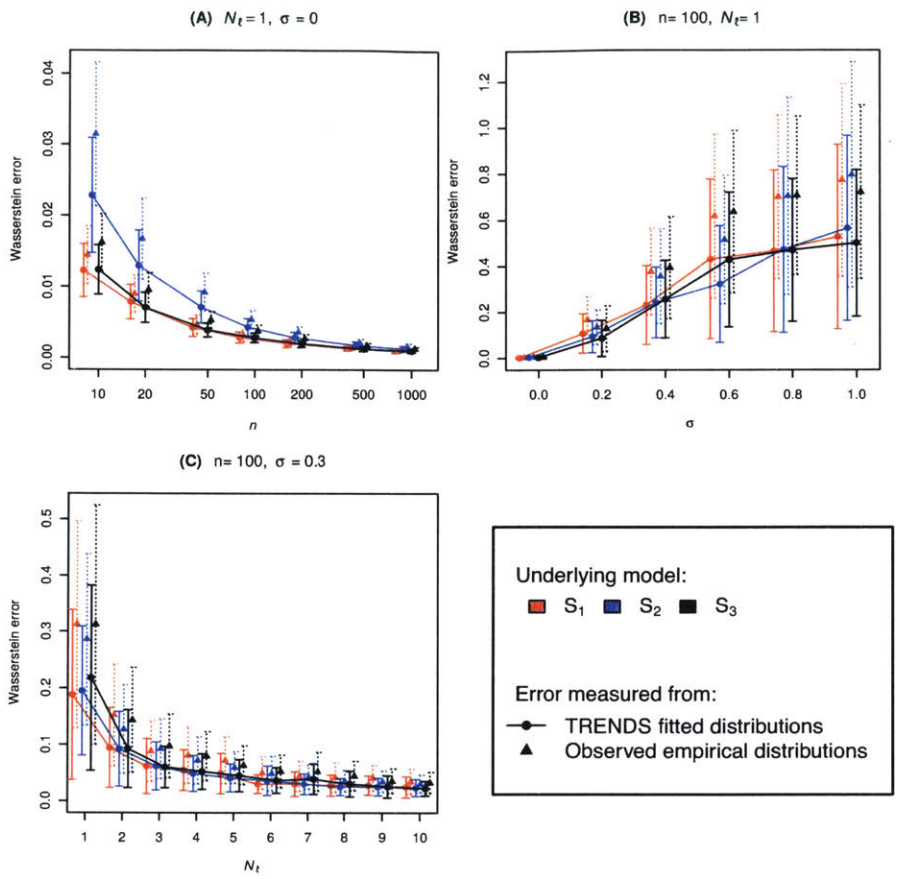


Figure 3-5: The Wasserstein error of the TRENDS fitted distributions vs. the observed empirical distributions, under models S_1 - S_3 with various settings of n, σ, and N_ℓ. Depicted is the average error (and standard deviation) over 100 repetitions.

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 in this chapter 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.

Table 3.1 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 gener-

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 3.1: 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).

ating 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 3.1 show that TRENDS is superior at drawing this distinction in these simulations.

3.9.1 Evaluating TRENDS p-values

Under the simulation setup of §3.9, 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 3.2 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 3.2: 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.

3.9.2 Determining whether TRENDS model is appropriate

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 3-6 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 5, we employ the

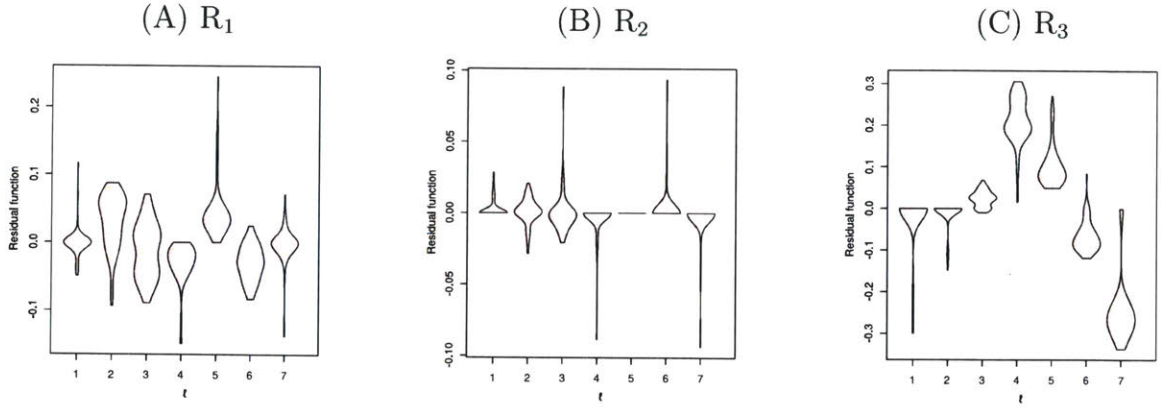


Figure 3-6: Diagnostic plot of the residual functions $\widehat{\mathcal{E}}_i(p)$ when TRENDS is fit to data from each underlying setting R_1, R_2, R_3 ($N_\ell = 1, n_i = 1000, \sigma = 0.1$). For each batch i, the plot depicts a kernel density estimate of the values taken by $\widehat{\mathcal{E}}_i(p)$ over $p = 0.01, 0.02, \ldots, 0.99$.

L_1 Wasserstein distance to define the true overall sequential-progression effect in this simulation as $\Delta_{\text{true}} = \sum_{\ell=2}^{L} d_{L_1}(Q_{\ell-1}, Q_\ell)$, 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 $\Delta_{\text{emp}} = \sum_{\ell=2}^{L} d_{L_1}(P_{\ell-1}, P_\ell)$ 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 3-7 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.

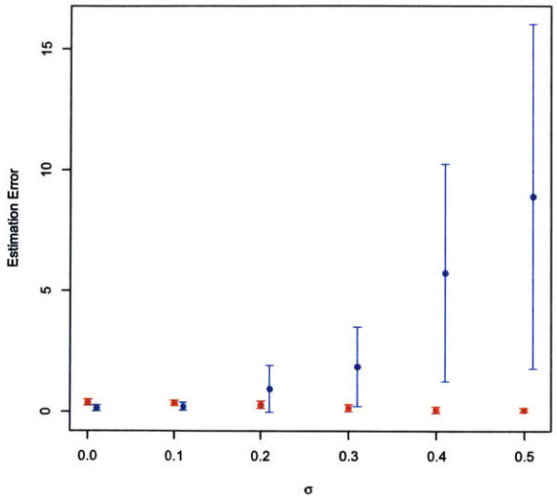


Figure 3-7: The mean/standard-deviation of the squared error of Δ_{emp} estimates (blue) and Δ_{TRENDS} estimates (red) over 100 datasets drawn from R_3 (under each value of σ, with $n_i = 100$ for each batch).

3.10 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).

3.10.1 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.

3.10.2 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.

3.10.3 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).

3.10.4 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 chapter.

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 5, 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.

3.11 TRENDS analysis of single cell RNA-seq data

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 §3.10. 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). Trapnell et al. (2014) 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 dif-

ferentiation 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 ($\geq$ 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[1] 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 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 ordinal levels. In this data, 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[2] in the sampled cells.

In all 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 between different genes. The methods we compare TRENDS against (§3.10) 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).

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 (3.8). 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 (Ge et al. 2003)). In these analyses, significance testing (which identifies high-confidence effects) and the Δ statistic (which identifies very large effects) both produce informative results.

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

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

To determine whether the TRENDS model appropriately fits this data, we first investigate the residual functions when TRENDS is applied to the scRNA-seq data from genes known to play a major role in regulating developmental processes. Figure 3-8 does not indicate any systematic pattern in the residuals that would suggest our model is inappropriate for these data. Table 3.3 details the most highly enriched terms identified in the significantly trending gene set from each dataset. Shown in Table 3.4 are 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 3.5 lists the highly enriched GO terms (found via the ConsensusPathDB tool of Kamburov et al. (2011)) in the 100 genes with largest Δ values in each dataset.

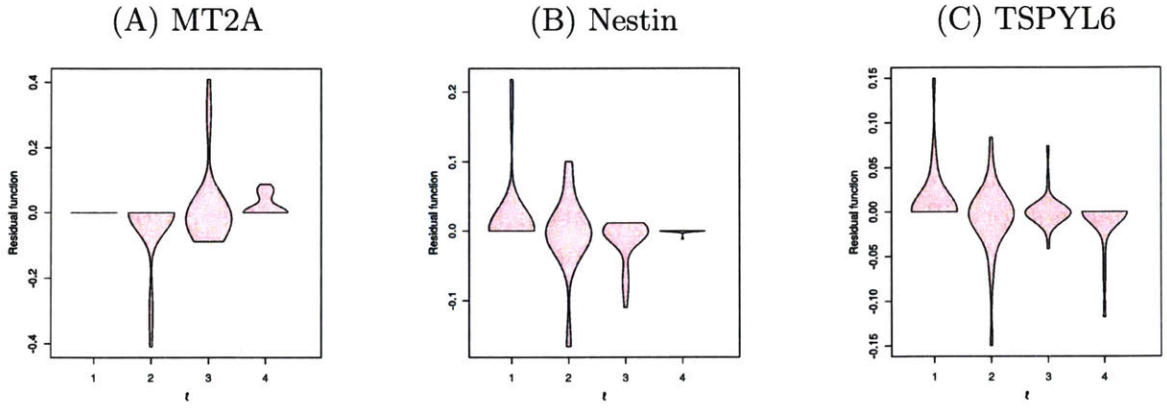


Figure 3-8: Diagnostic plot of the residual functions $\widehat{\mathcal{E}}_i(p)$ for TRENDS fit to scRNA-seq data from known regulatory genes of myoblast development. For each batch i, the plot depicts a kernel density estimate of the values taken by $\widehat{\mathcal{E}}_i(p)$ over $p = 0.01, 0.02, \ldots, 0.99$.

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 step-down minP multiple-testing correction (Ge et al. 2003)). 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 3-9A). Enriched biological process annotations include "anatomical structure development" and "cardiovascular system development" (Table 3.3A).

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 step-down minP multiple-testing correction). A search for GO enriched terms in the annotations of these genes shows a large subset to be developmentally relevant (Figure 3-9B), with enriched terms such as "neurogenesis" and "nervous system development" (Table 3.3B). 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 3-9: Word clouds of biological process 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). Each word cloud was produced using the Consensus-PathDB tool of 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 3.4A). Also, 7 of the top 10 genes in the cortex analysis have been previously implicated in brain development, particularly in sensory regions (Table 3.4B). Thus, TRENDS accurately assigns the largest inferred effects to clearly developmental genes (see also Table 3.5). 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 3-1A 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 ≥ 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 3.6 contains all GO annotations meeting this criterion. Figure 3-10A 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 §3.10). 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 3.7), and again find that TRENDS produces higher sensitivity than the other approaches (Figure 3-10B) 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.

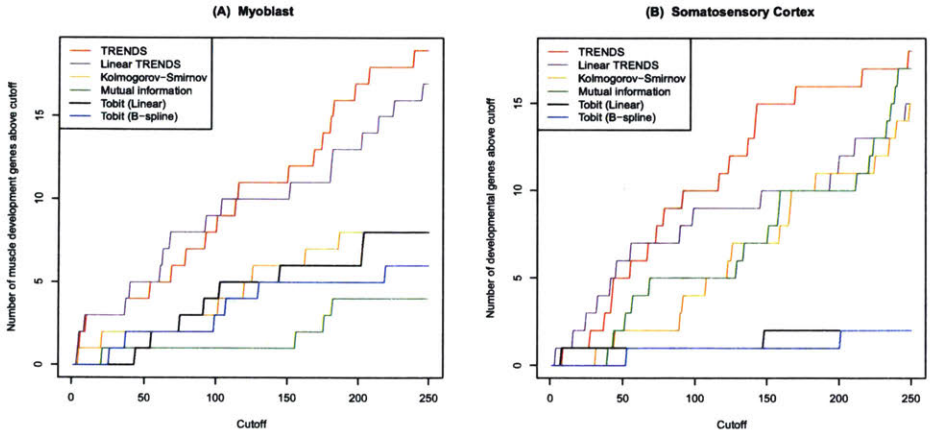


Figure 3-10: Pseudo-sensitivity of various methods based on their ability to identify known developmental genes. (A) the number of genes with a GO annotation containing both "muscle" and "development" found in the top K genes (ranked by the different methods for the myoblast data), over increasing K. (B) similar plot for the cortex data, where developmental genes are now those annotated with a relevant GO term from Table 3.7.

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 3-1B. TRENDS ranks Nestin 35th in terms of inferred developmental effect-size (with TRENDS p-value $= 0.02$ before step-down minP 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 3-1B 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 pseudo-sensitivity analysis (Figure 3-10), 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 ($R^2 = 0.68$). Figure 3-11 confirms that the TRENDS-fitted distributions for this gene lie very close to the observed expression distributions, so the vast majority of temporal variation in empirical TSPYL5 expression can be attributed to the presence of a consistent underlying trend.

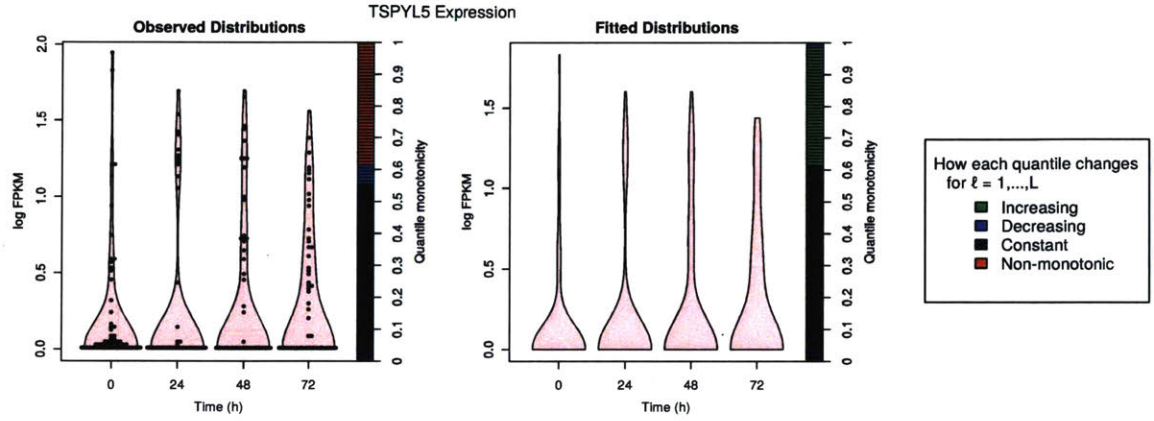


Figure 3-11: Violin plots depicting the empirical distribution of TSPYL5 expression measured in myoblast cells (on left), and the corresponding TRENDS fitted distributions (on right). Each point shows a sampled cell.

(A) Myoblast

Term	p-value	q-value
liver development	1e-4	6e-3
hepaticobiliary system development	1e-4	6e-3
anatomical structure development	3e-4	8e-3
gland development	3e-4	0.03
system development	2e-3	0.08
regulation of cyclin-dependent protein serine/threonine kinase activity	2e-3	0.08
single-multicellular organism process	3e-3	0.04
single-organism developmental process	4e-3	0.04
central nervous system development	5e-3	0.07
cardiovascular system development	5e-3	0.07
circulatory system development	5e-3	0.07
multicellular organismal development	5e-3	0.08
cellular nitrogen compound catabolic process	5e-3	0.07
response to hormone	5e-3	0.08
nervous system development	e-3	0.07
heart development	5e-3	0.08
regulation of cell cycle	6e-3	0.07
organ development	6e-3	0.08

(B) Somatosensory Cortex

Term	p-value	q-value
transmission of nerve impulse	6e-8	2e-5
multicellular organismal signaling	1e-7	3e-5
cell communication	6e-7	7e-5
neuron differentiation	1e-6	2e-4
cell development	3e-6	2e-4
ensheathment of neurons	3e-6	2e-4
axon ensheathment	3e-6	3e-4
single organism signaling	4-e6	3e-4
neurogenesis	1e-5	1e-3
regulation of biological quality	1e-5	4e-4
system development	1e-5	5e-4
neuron projection development	1e-5	1e-3
cell projection organization	1e-5	5e-4
single-organism cellular process	2e-5	4e-4
neuron development	2e-5	1e-3
anatomical structure development	3e-5	5e-4
nervous system development	3e-5	2e-3
cellular developmental process	5e-5	6e-4
cell differentiation	6e-5	2e-3
single-organism developmental process	7e-5	7e-4

Table 3.3: 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).

(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 3.4: 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 step-down minP 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.

(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 3.5: 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 adjustment 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 3.6: A list of all GO annotation terms containing both the words "muscle" and "development", used to produce the pseudo-sensitivity plots in Figure 3-10A.

	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 3.7: A list of the GO annotation terms relevant to the somatosensory cortex development, used to produce the pseudo-sensitivity plots in Figure 3-10B. This brain region is primarily composed of oligodendrocyte and pyramidal neuron cells (Zeisel et al. 2015).

3.12 ACS income distribution analysis

To demonstrate the broader utility of TRENDS beyond scRNA-seq analysis, we finally 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 3.8: 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).

Table 3.8 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. Interest-

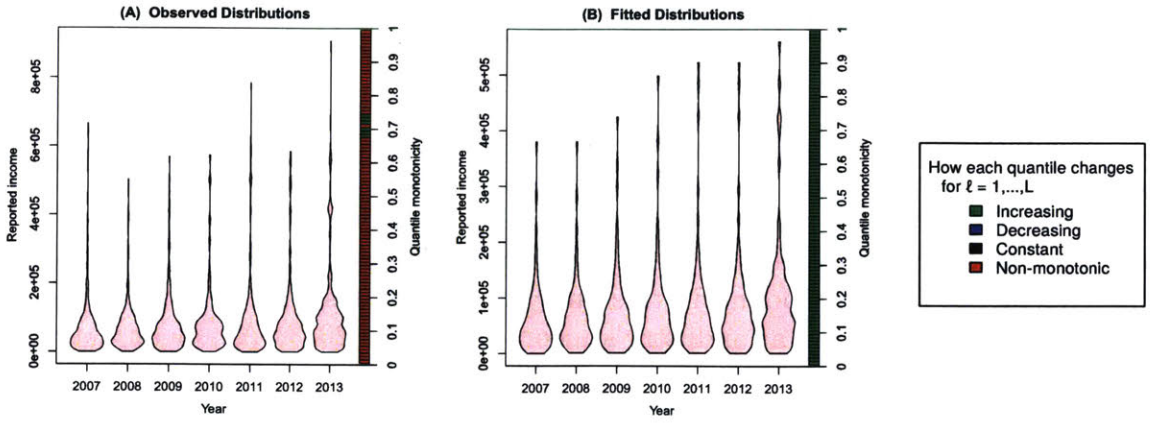


Figure 3-12: Distributions of reported income of individuals in the "other information services" industry. (A) kernel density estimates applied to the ACS survey results from each year (B) corresponding TRENDS fitted distributions.

ingly, many industries from the energy sector are also included in the table[3]. 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.

[3]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.

Chapter 4

Learning optimal interventions under uncertainty

In many data-driven applications, including medicine, the primary interest is identifying interventions that produce a desired change in some associated outcome. Because of experimental limitations, learning in such domains is commonly restricted to an observational dataset $\mathcal{D}_n := \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n}$ which consists of IID samples from a population with joint distribution P_{XY} over covariates (features) $X \in \mathbb{R}^d$ and outcomes $Y \in \mathbb{R}$. Typically, such data is analyzed using models which facilitate understanding of the relations between variables (e.g. assuming linearity/additivity). Based on conclusions drawn from this analysis, the analyst decides how to intervene in a manner they confidently believe will improve outcomes. Formalizing such beliefs via Bayesian inference, we develop an alternative framework that instead automatically identifies beneficial interventions directly from the data.

4.1 Causal assumptions

Under our setup, an intervention on an individual with pre-treatment covariates X produces post-treatment covariate values $\widetilde{X}$ that determine the resulting outcome Y (depicted as the graphical model: $X \to \widetilde{X} \to Y$). Each possible intervention results in a diffferent $\widetilde{X}$. More concretely, we make the following simplifying assumption:

$$Y = f(\widetilde{X}) + \varepsilon \quad \text{with } \mathbb{E}[\varepsilon] = 0, \varepsilon \perp\!\!\!\perp \widetilde{X}, X \tag{4.1}$$

for some underlying function f that encodes the effects of causal mechanisms (i.e. $\widetilde{X}$ represents a fair description of the system state, and some covariates in $\widetilde{X}$ causally affect Y, not vice-versa). The observed data is comprised of naturally occurring covariate values where we presume $\widetilde{x}^{(i)} = x^{(i)}$ for $i = 1, \ldots, n$ (i.e. the system state

remains static without intervention, so the observed covariate values directly influence the observed outcomes). Moreover, we assume the relationship between these covariate values and the outcomes remains invariant, following the same (unknown) function f for any $\widetilde{X}$ arising from one of our feasible interventions (or no intervention at all). Note that this assumption precludes the presence of hidden confounding. Peters et al. (2016) have also relied on this invariance assumption, verifying it as a reasonable property of causal mechanisms in nature.

4.2 Objectives

Given this data, we aim to learn an intervention policy defined by a covariate transformation $T : \mathbb{R}^d \to \mathbb{R}^d$, applied to each individual in the population. Here, $T(x)$ presents a desired setting of the covariates that should be reflected by subsequent intervention to actually influence outcomes. When T only specifies changes to a subset of the covariates, an intervention seeking to realize T may have unintended side-effects on covariates outside of this subset. We ignore such "fat hand" settings (Duvenaud et al. 2010) until §4.9. Instead, our methods assume interventions can always be carried out with great precision to ensure the desired transformation T is exactly reflected in the post-treatment values: $\widetilde{x} = T(x)$. Our goal is to identify the transformation T which produces the largest corresponding post-treatment improvement with high certainty. $T(x)$ can either represent a single mapping to be performed on all individuals (global policy) or encode a personalized policy where the intervened upon variables and their values may change with x.

Our strong assumptions are made to ensure that statistical modeling alone suffices to identify beneficial interventions. While many real-world tasks violate these conditions, there exist important domains in which violations are sufficiently minor that our methods can discover effective interventions (cf. Rojas-Carulla et al. (2016), Peters et al. (2016)). We use two applications to illustrate our framework. One is a writing improvement task where the data consists of documents labeled with associated outcomes (e.g. grades or popularity) and the goal is to suggest beneficial changes to the author. Our second example is a gene perturbation task where the expression of some regulatory genes can be up/down-regulated in a population (e.g. cells or bacteria) with the goal of inducing a particular phenotype or activation/repression of a downstream gene. In these examples, covariates are known to cause outcomes and our other assumptions may hold to some degree, depending on the type of external intervention used to alter covariate values.

The contributions of this chapter include:

1. A formal Bayesian definition of the optimal intervention that exhibits desirable characteristics under uncertainty due to limited data.

2. Widely applicable types of (sparse) intervention policy that are easily enacted across a whole population, and efficient algorithms to find the optimal intervention under practical constraints.

3. Theoretical insight regarding our methods' properties in Gaussian Process settings and certain misspecified applications, where interventions on a few covariates may unintentionally perturb additional non-intervened-upon covariates.

4.3 Related work

The same invariance assumption has been exploited by Peters et al. (2016) and Rojas-Carulla et al. (2016) for causal variable selection in regression models. Recently, researchers such as Duvenaud et al. (2010) and Kleinberg et al. (2015) have supported a greater role for predictive modeling in various decision-making settings. Zeevi et al. (2015) use gradient boosting to predict glycemic response based on diet (and personal/microbiome covariates), and found they can naively leverage their regressor to select personalized diets which result in superior glucose levels than the meals proposed by a clinical dietitian. As treatment-selection in high-impact applications (e.g. healthcare) grows increasingly reliant on supervised learning methods, it is imperative to properly handle uncertainty.

Nonlinear Bayesian predictive models have been employed by Hill (2011), Brodersen et al. (2015), and Krishnan et al. (2015) for quantifying the effects of a given treatment from observations of individuals who have been treated and those who have not. Rather than considering a single given intervention, we introduce the notion of an optimal intervention under various practical constraints, and how to identify such a policy from a limited dataset (in which no individuals have necessarily received any interventions).

Although our goals appear similar to Bayesian optimization and bandit problems (Shahriari et al. 2016, Agarwal et al. 2013), additional data is not collected in our setup. Since we consider settings where interventions are proposed based on all available data, acquisition functions for sequential exploration of the response-surface are not appropriate. As most existing data is not generated through sequential experimentation, our methods are more broadly applicable than iterative approaches like Bayesian optimization.

A greater distinction is our work's focus on the pre vs. post-intervention change in outcome for each particular individual, whereas Bayesian optimization seeks a single globally optimal configuration of covariates. In practice, feasible covariate transformations are constrained based on an individual's naturally occurring covariate-values, which stem from some underlying population beyond our control. For example in the writing improvement task, the goal is not to identify a globally optimal configuration

of covariates that all texts should strive to achieve, but rather to inform a particular author of simple modifications likely to improve the outcome of his/her existing article. Appropriately treating such constraints is particularly important when we wish to prescribe a global policy corresponding to a single intervention applied to all individuals from the population (there is no notion of an underlying population in Bayesian optimization).

4.4 Methods

Our strategy is to first fit a Bayesian model for $Y \mid X$ whose posterior encodes our beliefs about the underlying function f given the observed data. Subsequently, the posterior for $f \mid \mathcal{D}_n$ is used to identify a transformation of the covariates $T : \mathbb{R}^d \to \mathbb{R}^d$ which is likely to improve expected post-intervention outcomes according to our current beliefs. The posterior for $f \mid \mathcal{D}_n$ may be summarized at any points $x, x' \in \mathbb{R}^d$ by mean function $\mathbb{E}[f(x) \mid \mathcal{D}_n]$ and covariance function $\text{Cov}(f(x), f(x') \mid \mathcal{D}_n)$.

4.4.1 Intervening at the individual level

For $x \in \mathbb{R}^d$ representing the covariate-measurements from an individual, we are given a set $\mathcal{C}_x \subset \mathbb{R}^d$ that denotes constraints of possible transformations of x. Let $T(x) = \widetilde{x} \in \mathcal{C}_x$ denote the new covariate-measurements of this individual after a particular intervention on x which alters covariates as specified by transformation $T : \mathbb{R}^d \to \mathbb{R}^d$. Recall that we assume an intervention can be conducted to produce post-treatment covariate-values that exactly match any feasible transformation: $\widetilde{x} = T(x)$, and we thus write $f(T(x))$ in place of $\mathbb{E}_\varepsilon[Y \mid \widetilde{X} = T(x)]$.

We first consider *personalized interventions* in which T may be tailored to a particular x. Under the Bayesian perspective, $f \mid \mathcal{D}_n$ is randomly distributed according to our posterior beliefs, and we define the *individual expected gain* function:

$$G_x(T) := f(T(x)) - f(x) \mid \mathcal{D}_n \tag{4.2}$$

Since $f(x) = \mathbb{E}_\varepsilon[Y \mid \widetilde{X} = x]$, random function G_x evaluates the expected outcome-difference at the post vs. pre-intervention setting of the covariates (this expectation is over the noise ε, not our posterior). To infer the best personalized intervention (assuming higher outcomes are desired), we use optimization over vectors $T(x) \in \mathbb{R}^d$ to find:

$$T^*(x) = \underset{T(x) \in \mathcal{C}_x}{\text{argmax}} \; F^{-1}_{G_x(T)}(\alpha) \tag{4.3}$$

where $F^{-1}_{G(\cdot)}(\alpha)$ denotes the α^{th} quantile of our posterior distribution over $G(\cdot)$. We choose $0 < \alpha < 0.5$, which implies the intervention that produces $T^*(x)$ should

improve the expected outcome with probability $\geq 1 - \alpha$ under our posterior beliefs.

Defined based on known constraints of feasible interventions, the set $\mathcal{C}_x \subset \mathbb{R}^d$ enumerates possible transformations that can be applied to an individual with covariate values x. If the set of possible interventions is independent of x (i.e. $\mathcal{C}_x = \mathcal{C}\ \forall x$), then our goal is similar to the optimal covariate-configuration problem studied in Bayesian optimization. However, in many practical applications, x-independent transformations are not realizable through intervention. Consider gene perturbation, a scenario where it is impractical to simultaneously target more than a few genes due to technological limitations. If alternatively intervening on a quantity like caloric intake, it is only realistic to change an individual's current value by at most a small amount. The choice $\mathcal{C}_x := \{z \in \mathbb{R}^d : ||x - z||_0 \leq k\}$ reflects the constraint that at most k covariates can be intervened upon. We can denote limits on the amount that the s^{th} covariate may be altered by $\mathcal{C}_x := \{z \in \mathbb{R}^d : |x_s - z_s| \leq \gamma_s\}$ for $s \in \{1, \ldots, d\}$. In realistic settings, $\mathcal{C}_x$ may be the intersection of many such sets reflecting other possible constraints such as boundedness, impossible joint configurations of multiple covariates, etc.

For any $x, T(x) \in \mathbb{R}^d$: the posterior distribution for $G_x(T)$ has:

$$\text{mean} = \mathbb{E}[f(T(x) \mid \mathcal{D}_n] - \mathbb{E}[f(x) \mid \mathcal{D}_n \tag{4.4}$$

$$\begin{aligned}\text{variance} = {} & \text{Var}(f(T(x)) \mid \mathcal{D}_n) + \text{Var}(f(x) \mid \mathcal{D}_n) \\ & - 2\text{Cov}(f(T(x)), f(x) \mid \mathcal{D}_n)\end{aligned} \tag{4.5}$$

which is easily computed using the corresponding mean/covariance functions of the posterior $f \mid \mathcal{D}_n$. When $T(x) = x$, the objective in (4.3) takes value 0, so any superior optimum corresponds to an intervention we are confident will lead to expected improvement. If there is no good intervention in $\mathcal{C}_x$ (corresponding to a large increase in the posterior mean) or too much uncertainty about $f(x)$ given limited data, then our method simply returns $T^*(x) = x$ indicating no intervention should be performed.

Our objective exhibits these desirable characteristics because it relies on the posterior beliefs regarding both $f(T(x))$ and $f(x)$, which are tied via the covariance function. In contrast, a similarly-conservative lower confidence bound objective (i.e. an adaptation of the UCB acquisition function with lower rather than upper quantiles) would only consider $f(T(x))$, and could propose unsatisfactory transformations where we actually have $\mathbb{E}[f(x) \mid \mathcal{D}_n] > \mathbb{E}[f(T(x)) \mid \mathcal{D}_n]$.

4.4.2 Intervening on entire populations

The above discussion focused on personalized interventions tailored on an individual basis. In certain applications, policy-makers are interested in designing a single intervention which will be applied to all individuals from the same underlying population as the data. Relying on such a *global policy* is the only option in cases where we no

longer observe covariate-measurements of new individuals outside the data. In our gene perturbation example, gene expression may no longer be individually profiled in future specimens that receive the decided-upon intervention to save costs/labor.

Here, the covariates X are assumed distributed according to some underlying (pre-intervention) population, and we define the *population expected gain* function:

$$G_X(T) := \mathbb{E}_X[G_x(T)] = \mathbb{E}_X\big[f(T(x)) - f(x) \mid \mathcal{D}_n\big]$$

which is also randomly distributed based on our posterior ($\mathbb{E}_X$ is expectation with respect to the covariate-distribution X which is not modeled by $f \mid \mathcal{D}_n$). Our goal is now to find a single transformation $T : \mathbb{R}^d \to \mathbb{R}^d$ corresponding to a *population intervention* which will (with high certainty under our posterior beliefs) lead to large outcome improvements on average across the population:

$$T^* = \underset{T \in \mathcal{T}}{\operatorname{argmax}} \; F^{-1}_{G_X(T)}(\alpha) \tag{4.6}$$

Here, the family of possible transformations $\mathcal{T}$ is constrained such that $T(x) \in \mathcal{C}_x$ for all $T \in \mathcal{T}, x \in \mathbb{R}^d$. As a good model of our multivariate features may be unknown, we instead work with the empirical estimate:

$$T^* = \underset{T \in \mathcal{T}}{\operatorname{argmax}} \; F^{-1}_{G_n(T)}(\alpha) \tag{4.7}$$

where
$$G_n(T) := \frac{1}{n}\sum_{i=1}^{n} \big[f(T(x^{(i)})) - f(x^{(i)})\big] \mid \mathcal{D}_n$$

is the *empirical* population expected gain, whose posterior distribution has:

$$\text{mean} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(T(x^{(i)})) \mid \mathcal{D}_n] - \mathbb{E}[f(x^{(i)}) \mid \mathcal{D}_n] \tag{4.8}$$

$$\begin{aligned}\text{variance} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \Big[&\operatorname{Cov}\big(f(x^{(i)}), f(x^{(j)}) \mid \mathcal{D}_n\big) \\ &- \operatorname{Cov}\big(f(T(x^{(i)})), f(x^{(j)}) \mid \mathcal{D}_n\big) \\ &- \operatorname{Cov}\big(f(x^{(i)}), f(T(x^{(j)})) \mid \mathcal{D}_n\big) \\ &+ \operatorname{Cov}\big(f(T(x^{(i)})), f(T(x^{(j)})) \mid \mathcal{D}_n\big)\Big]\end{aligned} \tag{4.9}$$

The population intervention objective in (4.7) is again 0 for the identity mapping $T(x) = x$. Under excessive uncertainty or a dearth of beneficial transformations in $\mathcal{T}$, the policy produced by this method will again simply be to perform no intervention. In this population intervention setting, T is designed assuming future individuals will stem from the same underlying distribution as the samples in $\mathcal{D}_n$. Although T is a function of x, the form of the transformation must be agnostic to the specific values of x (so the intervention can be applied to new individuals without measuring their

covariates).

We consider two types of transformations that we find widely applicable. *Shift* interventions involve transformations of the form: $T(x) = x + \Delta$ where $\Delta \in \mathbb{R}^d$ represents a (sparse) shift that the policy applies to each individuals' covariates (e.g. always adding 3 to the value of the second covariate corresponds to $T(x) = [x_1, x_2 + 3, \ldots, x_d]$). *Covariate-fixing* interventions are policies which set certain covariates to a constant value for all individuals, and involve transformations $T_{\mathcal{I} \to z}(x) = [z_1, \ldots, z_d]$ such that for some covariate-subset $\mathcal{I} \subseteq \{1, \ldots, d\}$: $z_j = x_j \; \forall j \notin \mathcal{I}$ and for $j \in \mathcal{I}$: $z_j \in \mathbb{R}$ is fixed across all x (e.g. always setting the first covariate to 0, for example in gene knockout, corresponds to $T(x) = [0, x_2, \ldots, x_d] \; \forall x$). Figure 4-1 depicts examples of these different interventions. Under a sparsity constraint, we must carefully model the underlying population in order to identify the best covariate-fixing intervention (here, setting X_1 to a large value is superior to intervening on X_2).

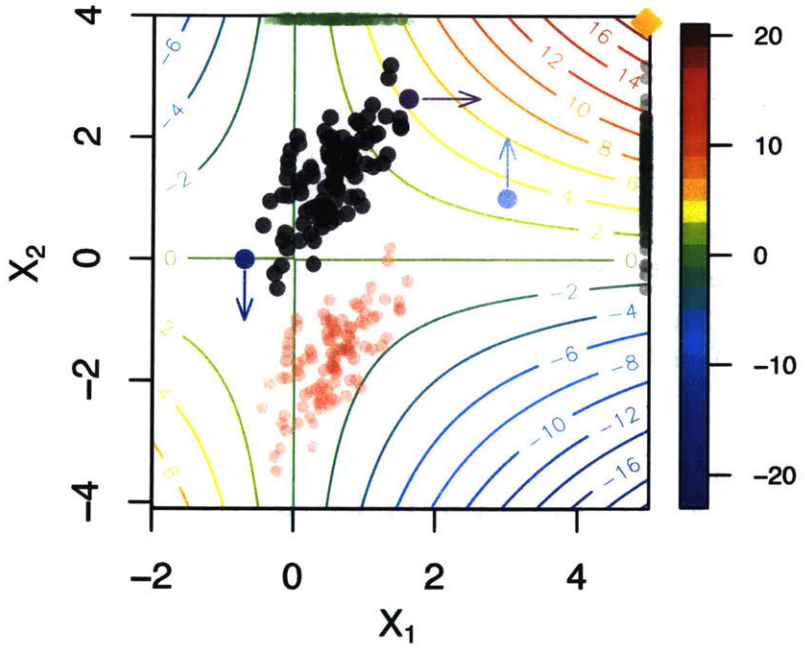


Figure 4-1: Contour plot of expected outcomes over feature space $[X_1, X_2]$ for relationship $Y = X_1 \cdot X_2 + \varepsilon$. Black points: the underlying population. Gold diamond: optimal covariate-setting if any transformation in the box were feasible. Red points: same population after shift intervention $\Delta = [-3, 0]$. Light (or dark) green points (along border): best covariate-fixing intervention which can only set X_2 (or only X_1) to a fixed value. Blue, purple, light blue points: individuals who receive a single-variable personalized intervention (arrows indicate direction of optimal transformation).

4.5 Gaussian process regression

Gaussian Process (GP) regression is a nonparametric Bayesian model that has enjoyed widespread success in supervised learning settings with limited data (Rasmussen 2006). This technique has been favored in many applications as it produces both accurate predictions and effective measures of uncertainty (with closed-form estimators available in the standard case). Furthermore, a variety of GP models exist for different settings including: non-Gaussian response variables (Rasmussen 2006), non-stationary relationships (Paciorek & Schervish 2004), deep representations (Daminaou & Lawrence 2013), measurement error (McHutchon & Rasmussen 2011), and heteroscedastic noise (Le et al. 2005). While these more advanced variants are not employed in this thesis, our intervention-methodology can be directly used in conjunction with such extensions (or more generally, any model which produces a useful posterior for $f \mid \mathcal{D}_n$).

The key idea of the GP approach is to adopt a Gaussian process prior over the space of possible functions mapping features to outcomes, under which $f(x^{(1)}), \ldots, f(x^{(n)})$ follow a multivariate Gaussian distribution $\mathrm{N}(\mathbf{m}_n, \mathbf{K}_{n,n})$ for any collection of data points $\{x^{(i)}\}_{i=1}^n$. This prior (and the resulting GP model) is specified by a prior mean function $m : \mathbb{R}^d \to \mathbb{R}$ and positive-definite covariance function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ which encodes our prior belief regarding properties of the underlying relationship between X and Y (such as smoothness or periodicity). Here, the vector $\mathbf{m}_n \in \mathbb{R}^n$ denotes the evaluation of function m at each point $\{x^{(i)}\}_{i=1}^n$, and $\mathbf{K}_{n,n}$ denotes the matrix whose i, j^{th} component is $k(x^{(i)}, x^{(j)})$. Given test input points $x_*^{(1)}, \ldots, x_*^{(n_*)} \in \mathbb{R}^d$ in addition to training data $\mathcal{D}_n$, we additionally define: $\mathbf{f}_* := [f(x_*^{(1)}), \ldots, f(x_*^{(n_*)})]$, $\mathbf{y}_n = [y^{(1)}, \ldots, y^{(n)}]$, matrix $\mathbf{K}_{n,*}$ with i, j^{th} entry $k(x^{(i)}, x_*^{(j)})$ (where $x^{(i)}$ is the i^{th} training input), and matrix $\mathbf{K}_{*,*}$ which contains pairwise covariances between test inputs.

The standard GP regression model assumes that the noise $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ is independently sampled for each observation. In this case, the posterior for f at the test inputs, $\mathbf{f}_* \mid \mathcal{D}_n$, follows a $\mathrm{N}(\mu_{\mathbf{n}*}, \mathbf{\Sigma}_{\mathbf{n}*})$ distribution with the following mean vector and covariance matrix:

$$\mu_{\mathbf{n}*} = \mathbf{m}_* + (\mathbf{K}_{n,n} + \sigma^2\mathbf{I})^{-1}(\mathbf{y}_n - \mathbf{m}_n), \;\; \mathbf{\Sigma}_{\mathbf{n}*} = \mathbf{K}_{*,*} - \mathbf{K}_{*,n}(\mathbf{K}_{n,n} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{n,*}$$

Note that our intervention-optimization framework is not specific to this GP model, but can be combined with any algorithm that learns a reasonable posterior for f. While employing a more powerful model should improve the results of our approach, comparing various regressors is not our focus. Thus, all practical results of our methodology are presented using only the standard GP regression model, under which the posterior distribution over f is given by the above expressions. In each application presented here, our GP uses the Automatic-Relevance-Determination (ARD)

covariance function, a popular choice for multi-dimensional data (Rasmussen 2006):

$$k(x, x') = \sigma_0^2 \cdot \exp\left[-\frac{1}{2}\sum_{s=1}^{d}\left(\frac{x_s - x'_s}{l_s}\right)^2\right] \tag{4.10}$$

The ARD kernel relies on length-scale hyperparameters $l_1, \ldots, l_d$ which determine how much f can depend on each dimension of the feature-space. All hyperparameters of our GP regression model (covariance-kernel parameters $l_1 \ldots, l_d$ and σ_0 (the output variance) as well as the variance of the noise σ^2) are empirically selected via marginal-likelihood maximization (i.e. Type II maximum likelihood: Rasmussen, 2006). In each application, we employ the 0.05^{th} posterior-quantile ($\alpha = 0.05$) in our method to ensure that with high probability, the intervention it infers to be optimal induces a nonnegative change in expected outcomes.

4.6 Algorithms to identify beneficial transformations

Throughout this chapter, we employ Gaussian Process regression to model $Y \mid X$, as described in the previous section. Under the basic GP model, $G_x(T)$ follows a Gaussian distribution and the α^{th} quantile of our personalized gain is simply given by:

$$F^{-1}_{G_x(T)} = \mathbb{E}[G_x(T)] + \Phi^{-1}(\alpha) \cdot \text{Var}[G_x(T)]^{1/2} \tag{4.11}$$

where Φ^{-1} denotes the $N(0,1)$ quantile function. The quantiles of the empirical population gain may be similarly obtained. When a smooth covariance kernel $k(\cdot, \cdot)$ is adopted in the GP prior, derivatives of our intervention-objectives are easily computed with respect to T.

In many practical settings, an intervention that only affects a small subset of variables is desired. Software to improve text, for example, should not overwhelm authors with a multitude of desired changes, but rather present a concise list of the most beneficial revisions in order to retain underlying semantics. Note that identifying a sparse transformation of the covariates is different from feature selection in supervised learning (where the goal is to identify dimensions along which f varies most). In contrast, we seek the dimensions $\mathcal{I} \subset \{1, \ldots, d\}$ along which one of our feasible covariate-transformations can produce the largest high-probability increase in f, assuming the other covariates remain fixed at their initial pre-treatment values (in the case of personalized intervention) or follow the same distribution as the pre-intervention population (in the case of a global policy).

For a shift intervention $T(x) = x + \Delta$, we introduce the convenient notation $G_n(\Delta) := G_n(T)$. In applications where shifting x_s (the s^{th} covariate for $s \in \{1, \ldots, d\}$) by one unit incurs cost γ_s, we account for these costs by considering the following

regularized intervention-objective:

$$J_\lambda(\Delta) := F^{-1}_{G_n(\Delta)}(\alpha) - \lambda \sum_{s=1}^{d} \gamma_s |\Delta_s| \tag{4.12}$$

By maximizing this objective over feasible set $\mathcal{C}_\Delta := \{\Delta \in \mathbb{R}^d : x + \Delta \in \mathcal{C}_x \text{ for all } x \in \mathbb{R}^d\}$, policy-makers can decide which variables to intervene upon (and how much to shift them), depending on the relative value of outcome-improvements (specified by λ).

The optimization of our regularized objective J_λ is performed using the proximal gradient method (Bertsekas 1995), where at each iterate: a step in the gradient direction is followed by a soft-thresholding operation (Bach et al. 2012) as well as a projection back onto the feasible set $\mathcal{C}_\Delta$. When $\lambda = 0$ and there is no penalty, we instead use the Sequential Least Squares Programming procedure of Kraft (1988), which empirically converged more quickly that the basic gradient method when applied to our unregularized objective.

4.6.1 Continuation method to avoid poor local optima

As our intervention objective J_λ is often highly nonconcave, first/second-order optimization methods may suffer from the presence of poor local optima. Nonconcavity arises primarily due to the fact that the Gaussian process uncertainty balloons in any region lacking data. To deal with local optima in acquisition functions, Bayesian optimization methods employ heuristics like combining the results of many local optimizers or operating over a fine partitioning of the feature space (Shahriari et al. 2016, Lizotte 2008).

We instead develop a continuation technique that performs a series of gradient-based optimizations over variants of this objective with tapering levels of added smoothness. Under this strategy, we solve a series of optimization problems, each of which operates on our objective under a smoothed posterior (and the amount of additional smoothing is gradually decreased to zero). Excessive smoothing of the posterior is achieved by simply considering GP models whose kernels are given overly large length-scale parameters. Each time the amount of smoothing is tapered, we initialize our local gradient optimizer using the solution found at the previously greater smoothing level. Intuitively, the highly smoothed GP model is primarily influenced by the global structure in the data, and thus our optimization with respect to the posterior of this model is far less susceptible to low-quality local maxima. Analysis of a similar homotopy strategy under radial basis kernels has been conducted by Mobahi et al. (2012).

4.6.2 Sparse shift intervention

In some settings, one may require that at most $k < d$ covariates are intervened upon. We identify the optimal k-sparse shift intervention via the Sparse Shift Algorithm below, which relies on ℓ_1-relaxation (Bach et al. 2012) and the regularization path of our penalized intervention objective. As the ℓ_1-norm provides the closest convex relaxation to the ℓ_0 norm, this is a a commonly adopted strategy to avoid combinatorial search in feature selection (Bach et al. 2012).

When applied to identify sparse shift interventions for populations, the Sparse Shift Algorithm computes the regularization path over different settings of the penalty $\lambda > 0$ for the following regularized objective:

$$J_\lambda(\Delta) := F^{-1}_{G_n(\Delta)}(\alpha) - \lambda||\Delta||_1 \tag{4.13}$$

which is maximized over the feasible set $\mathcal{C}_\Delta := \{\Delta \in \mathbb{R}^d : x + \Delta \in \mathcal{C}_x \text{ for all } x \in \mathbb{R}^d\}$ (recall we write $G_n(\Delta) := G_n(T)$ when $T(x) = x + \Delta$). Subsequently, we identify the regularization penalty which produces a shift of desired cardinality and select our intervention set $\mathcal{I}$ as the covariates which receive nonzero shift. Finally, we optimize the original unregularized objective ($\lambda = 0$) with respect to only the selected covariates in $\mathcal{I}$ to remove bias induced by the regularizer. Each inner maximization in both the Sparse Shift/Covariate-fixing algorithms is performed via the previously described proximal gradient methods combined with our continuation strategy.

Sparse Shift Algorithm: Finds the best k-sparse shift transformation vector Δ^* within feasible set $\mathcal{C}_\Delta \subset \mathbb{R}^d$.

1: Set $\gamma_s = 1$ for $s = 1, \ldots, d$

2: Perform binary search over λ to find:

$$\lambda^* \leftarrow \text{argmin}\left\{\lambda \geq 0 \text{ s.t. } \Delta^* := \underset{\Delta \in \mathcal{C}_\Delta}{\text{argmax}}\, J_\lambda(\Delta) \quad \text{has} \leq k \text{ nonzero entries}\right\}$$

3: Define $\mathcal{I} \leftarrow \text{support}(\Delta^*_{\lambda^*}) \subseteq \{1, \ldots, d\}$ where $\Delta^*_{\lambda^*} := \underset{\Delta \in \mathcal{C}_\Delta}{\text{argmax}}\, J_{\lambda^*}(\Delta)$

4: **Return:** $\Delta^* \leftarrow \underset{\Delta \in B}{\text{argmax}}\, J_0(\Delta)$ where $B := \mathcal{C}_\Delta \bigcap \{\Delta \in \mathbb{R}^d : \Delta_s = 0 \text{ if } s \notin \mathcal{I}\}$

Recall that in the case of personalized intervention, we simply optimize over vectors $T(x) \in \mathcal{C}_x$. Any personalized transformation can therefore be equivalently expressed as a shift in terms of $\Delta_x \in \mathbb{R}^d$ such that $T(x) = x + \Delta_x$. After substituting the individual gain $G_x(\Delta_x)$ in place of the population gain $G_n(\Delta)$ within our definition of J_λ in (4.13), we can thus employ the same algorithm to identify sparse/cost-sensitive personalized interventions. To find a covariate-fixing intervention which sets k of the

covariates to particular fixed constants across all individuals from the population, we instead employ a forward step-wise selection algorithm detailed in the next section, as the form of the optimization is no longer amenable to ℓ_1-relaxation.

4.6.3 Sparse covariate-fixing intervention

In other applications, one may wish to identify the optimal covariate-fixing intervention which sets k of the covariates to particular fixed constants uniformly across all individuals from the population. In this setting, it is not easy to leverage ℓ_1-relaxation, so we instead employ the forward step-wise selection algorithm described below. Recall $\mathcal{I} \subseteq \{1, \ldots, d\}$ denotes the subset of covariates which are intervened upon, and the covariate-fixing intervention produces vector $T_{\mathcal{I}\to z}(x) \in \mathbb{R}^d$ such that $T_{\mathcal{I}\to z}(x)_s = x_s$ if $s \notin \mathcal{I}$, otherwise $T_{\mathcal{I}\to z}(x)_s = z_s$ which is a constant chosen by the policy-maker. This same transformation is applied to each individual in the population, creating a more homogeneous group which share the same value for the covariates in $\mathcal{I}$. For a given $\mathcal{I}$, the objective function to find the best constants is:

$$J_{\mathcal{I}}^{\text{unif}}\big(\{z_s\}_{s\in\mathcal{I}}\big) := F^{-1}_{G_n(T_{\mathcal{I}\to z})}(\alpha) \tag{4.14}$$

$$\text{with } G_n(T_{\mathcal{I}\to z}) = \frac{1}{n}\sum_{i=1}^{n}\big[f(z^{(i)}) - f(x^{(i)})\big] \mid \mathcal{D}_n \text{ where } z_s^{(i)} = \begin{cases} x^{(i)} & \text{if } s \notin \mathcal{I} \\ z_s & \text{otherwise} \end{cases}$$

which is maximized over the constraints: $z_s \in \mathcal{C}_s \subseteq \mathbb{R}$ for $s \in \mathcal{I}$. Each maximization over a fixed set of $\{z_s\}_{s\in\mathcal{I}}$ is again performed via the previously described proximal gradient methods combined with our continuation strategy.

Sparse Covariate-fixing Algorithm: Identifies the best k-sparse covariate-fixing transformation, where sets $\mathcal{C}_1, \ldots, \mathcal{C}_d \subseteq \mathbb{R}$ encode feasible settings for each covariate.

Initialize $\mathcal{I} \leftarrow \varnothing$, $\mathcal{U} \leftarrow \{1, \ldots, d\}$, $J^* \leftarrow 0$

While $|\mathcal{I}| < k$:

  Set $J_s^* \leftarrow \max_{\mathcal{C}_r : r\in\mathcal{I}\cup\{s\}} J^{\text{unif}}_{\mathcal{I}\cup\{s\}}\big(\{z_r\}_{r\in\mathcal{I}\cup\{s\}}\big)$ **for** each $s \in \mathcal{U}$

  Find $s^* \leftarrow \text{argmax}_{s\in\mathcal{U}}\big\{J_s^*\big\}$

  If $J_{s^*}^* > J^*$: $J^* \leftarrow J_{s^*}^*$, $\mathcal{I} \leftarrow \mathcal{I}\cup\{s^*\}$, $\mathcal{U} \leftarrow \mathcal{U}\setminus s^*$

  Else: **break**

Return: $\{z_s^*\}_{s\in\mathcal{I}} \leftarrow \underset{\mathcal{C}_s : s\in\mathcal{I}}{\text{argmax}}\ J_{\mathcal{I}}^{\text{unif}}\big(\{z_s\}_{s\in\mathcal{I}}\big)$

4.7 Theoretical results

Here, we present some theoretical analysis of our methodology, in which, we presume the causal assumptions laid out in §4.1 hold throughout. For clarity, we rewrite the true underlying relationship as f^* in this section, letting f now denote arbitrary functions, and our theoretical results are with respect to the *true improvement* of an intervention $G_x^*(T) := f^*(T(x)) - f^*(x)$, $G_X^*(T) := \mathbb{E}_X[G_x^*(T)]$. This theory corresponds a frequentist analysis of our Bayesian approach, in which G_x^*, G_X^* are no longer viewed as random because the true underlying relationship is fixed at an arbitrary f^* (not sampled from some prior distribution). We also note that much of our theory relies on Gaussian Process results derived by Srinivas et al. (2010) and van der Vaart & van Zanten (2011).

Theorems 10 and 11 below characterize the rate at which our empirical personalized and population intervention objectives are expected to converge to the true improvement (due in part to contraction of the posterior as n grows). Since these results hold for all T, this implies the maximizer of each of our empirical intervention-objectives is asymptotically consistent, converging to the true optimal transformation as $n \to \infty$ (under a reasonable prior). Complementing these consistency results, Theorem 12 ensures that for any finite sample-size $n > 0$: optimizing our personalized intervention objective corresponds to improving a lower bound on the true improvement with high probability, when α is small and f^* belongs to the RKHS of our prior. Here, probability is used in the frequentist sense to refer to random draws of the data, so this property is non-obvious (from the Bayesian perspective, α exactly controls the probability of a harmful intervention).

In this case, the optimal transformation inferred by our approach only worsens the actual expected outcome with low probability.

Our consistency statements depend on the following additional assumptions:

(A22) There exist $\rho > 0$ such that the Hölder space $C^\rho[0,1]^d$ has probability one under our $\text{GP}\big(0, k(x,x')\big)$ prior (see van der Vaart & van Zanten (2011) for details).

(A23) f^* and any f supported by our Gaussian process prior are Lipschitz continuous over $\mathcal{C}$ with constant L.

(A24) All data lies in the compact subset $\mathcal{C} = [0,1]^d$.

(A25) We select a low quantile in our intervention objectives: $0 < \alpha \leq 0.5$.

(A26) The noise variables are Gaussian distributed: $\varepsilon^{(i)} \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, n$.

(A27) The density of our input covariates $p_X \in [a, b]$ is bounded above and below over compact domain $\mathcal{C}$.

Theorem 10. *If we adopt a* $GP\big(0, k(x,x')\big)$ *prior and conditions (A22)-(A27) are met, then for all* $x, T(x) \in \mathcal{C}$*:*

$$\mathbb{E}_{\mathcal{D}_n}\Big|F^{-1}_{G_x(T)}(\alpha) - G^*_x(T)\Big| \leq \frac{C}{\alpha}\Big(L + \frac{1}{a}\Big) \cdot \Psi_{f^*}(n)^{1/[2(d+1)]}$$

where constant C depends on the prior and density p_X and we define:

$$\Psi_f(n) := \begin{cases} \big[\psi_f^{-1}(n)\big]^2 & \text{if } \psi_f^{-1}(n) \leq n^{-d/(4\rho+2d)} \\ n \cdot [\psi_{f^*}^{-1}(n)]^{(4\rho+4d)/d} & \text{otherwise} \end{cases}$$

$\psi^{-1}_{f^*}(n)$ is the (generalized) inverse of $\psi_{f^*}(\epsilon) := \frac{\phi_{f^*}(\epsilon)}{\epsilon^2}$ which depends on the concentration function $\phi_{f^*}(\epsilon) = \inf_{h \in \mathcal{H}_k : ||h - f^*||_\infty < \epsilon} ||h||_k^2 - \log \Pi\big(f : ||f||_\infty < \epsilon\big)$. ϕ_{f^*} measures how well the RKHS of our GP prior $\mathcal{H}_k$ approximates f^* (see van der Vaart & van Zanten (2011) for more details). The expectation $\mathbb{E}_{\mathcal{D}_n}$ is over the distribution of the data $\{(X^{(i)}, Y^{(i)})\}_{i=1}^n$.

Importantly, Theorem 10 does not assume anything about the true relationship f^*, and the bound depends on the distance between f^* and our prior. When f^* is a ρ-smooth function, a typical bound is given by $\psi^{-1}_{f^*}(n) = \mathcal{O}(n^{-\min\{\nu,\rho\}/(2\nu+d)})$ if k is the Matérn kernel with smoothness parameter ν. When k is the squared exponential kernel and f^* is β-regular (in the Sobolev sense), $\psi^{-1}_{f^*}(n) = \mathcal{O}((1/\log n)^{\beta/2 - d/4})$.

Proof of Theorem 10. Recall that $G_x(T) := f(T(x)) - f(x) \mid \mathcal{D}_n$ depends on f. We fix $x_0, T(x_0) \in \mathcal{C}$ and adapt the bound provided by Theorem 13 to show our result. Let $\mathcal{B}_\delta(x) \subset \mathcal{C}$ denote the ball of radius $0 < \delta < \frac{1}{2}$ centered at $x \in \mathcal{C}$. We first establish the bound:

$$\begin{aligned}
&\int_{\mathcal{C}} \big|f(x) - f^*(x)\big| p_X(x)\, \mathrm{d}x \\
\geq &\int_{\mathcal{B}_\delta(x_0)} \big|f(x) - f^*(x)\big| p_X(x)\, \mathrm{d}x + \int_{\mathcal{B}_\delta(T(x_0))} \big|f(x) - f^*(x)\big| p_X(x)\, \mathrm{d}x \\
\geq &a \cdot \mathrm{Vol}(\mathcal{B}_\delta)\Big[\min_{x \in \mathcal{B}_\delta(x_0)} \big|f(x) - f^*(x)\big| + \min_{x \in \mathcal{B}_\delta(T(x_0))} \big|f(x) - f^*(x)\big|\Big] \\
\geq &a \cdot \mathrm{Vol}(\mathcal{B}_\delta) \cdot \Big[\Big|f(T(x_0)) - f(x_0) - \big[f^*(T(x_0)) - f^*(x_0)\big]\Big| - 8\delta L\Big] \\
\geq &a \cdot \mathrm{Vol}(\mathcal{B}_\delta) \cdot \Big[\Big|G_{x_0}(T) - G^*_{x_0}(T)\Big| - 8\delta L\Big] \qquad (4.15)
\end{aligned}$$

where $\mathrm{Vol}(\mathcal{B}_\delta) = \mathcal{O}(\delta^d)$. Theorem 13 below implies the following inequality (ignoring

constant factors):

$$
\begin{aligned}
[C \cdot \Psi_{f^*}(n)]^{1/2} \geq & \left[\mathbb{E}_{\mathcal{D}_n} \int \int_{\mathcal{C}} [f(x) - f^*(x)]^2 p_X(x) \, \mathrm{d}x \, \mathrm{d}\Pi_n(f \mid \mathcal{D}_n) \right]^{1/2} \\
\geq & \mathbb{E}_{\mathcal{D}_n} \int \int_{\mathcal{C}} \left| f(x) - f^*(x) \right| p_X(x) \, \mathrm{d}x \, \mathrm{d}\Pi_n(f \mid \mathcal{D}_n) \\
& \qquad \text{by Jensen's inequality} \\
\geq & a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int \left| G_{x_0}(T) - G^*_{x_0}(T) \right| - \delta L \;\; \mathrm{d}\Pi_n(f \mid \mathcal{D}_n) \\
& \qquad \text{via the bound from (4.15)} \\
= & -aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_0^\infty \Pr\left(\left| G_{x_0}(T) - G^*_{x_0}(T) \right| \geq r \right) \mathrm{d}r \\
= & -aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_0^1 F^{-1}_{|G_{x_0}(T) - G^*_{x_0}(T)|}(\widetilde{\alpha}) \;\; \mathrm{d}\widetilde{\alpha} \\
\geq & -aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_\alpha^1 F^{-1}_{G_{x_0}(T)}(\widetilde{\alpha}) - G^*_{x_0}(T) \;\; \mathrm{d}\widetilde{\alpha} \\
\geq & -aL\delta^{d+1} + a(1-\alpha)\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \left[F^{-1}_{G_{x_0}(T)}(\alpha) - G^*_{x_0}(T) \right] \qquad (4.16)
\end{aligned}
$$

We can similarly bound $G^*_{x_0}(T) - F^{-1}_{G_{x_0}(T)}(\alpha)$:

$$
\begin{aligned}
& -aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_0^1 F^{-1}_{|G^*_{x_0}(T) - G_{x_0}(T)|}(\widetilde{\alpha}) \; \mathrm{d}\widetilde{\alpha} \\
\geq & -aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_0^\alpha G^*_{x_0}(T) - F^{-1}_{G_{x_0}(T)}(\widetilde{\alpha}) \;\; \mathrm{d}\widetilde{\alpha} \\
\geq & -aL\delta^{d+1} + a\alpha\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \left[G^*_{x_0}(T) - F^{-1}_{G_{x_0}(T)}(\alpha) \right] \qquad (4.17)
\end{aligned}
$$

Choosing $\delta := [\Psi_{f^*}(n)]^{\frac{1}{2(d+1)}}$ and combining (4.16) and (4.17) produces the desired result, since $\alpha < 0.5$ implies $\alpha < 1 - \alpha$. □

Theorem 11. *Under the assumptions of Theorem 10, for any T such that $\Pr(T(X) \in \mathcal{C}) = 1$, we have:*

$$
\mathbb{E}_{\mathcal{D}_n} \left| F^{-1}_{G_n(T)}(\alpha) - G^*_X(T) \right| \leq \frac{C}{\alpha} \left[L\sqrt{\frac{d}{n}} + \left(L + \frac{1}{a} \right) \Psi_{f^*}(n)^{\frac{1}{2(d+1)}} \right]
$$

Proof. Combining the results of Lemmas 13 and 14 below, we obtain the desired upper bound through a straightforward application of the triangle inequality. Note that we've simplified the bound using the identity $-\log(1-\alpha) < 1/\alpha$ for $\alpha < 0.5$. □

For our final result, we adopt a few additional assumptions:

(A28) $f^* \in \mathcal{H}_k(\mathcal{C})$ which is the RKHS induced by our covariance function k with norm $||\cdot||_k$ (cf. Rasmussen (2006) §6.1).

(A29) Noise variables $\varepsilon^{(i)}$ form a uniformly bounded martingale difference sequence $\varepsilon^{(i)} \leq \sigma$ for $i = 1, \ldots, n$.

Theorem 12. *Suppose we adopt a $GP(0, k(x,x'))$ prior and conditions (A24)-(A29) are met. Then, for any $x, T(x) \in \mathcal{C}$:* $F^{-1}_{G_x(T)}(\alpha) \leq G^*_x(T)$

with probability (over the noise) greater than $1 - C(n+1)\cdot\exp\left(-\frac{[\Phi^{-1}(\alpha)]^2 - 2||f^*||_k^2}{\gamma_n}\right)$

In Theorem 12, $\gamma_n := \max_{A \subset \mathcal{C}: |A|=n} \frac{1}{2}\log\left|\mathbf{I} + \sigma^{-2}\mathbf{K}_A\right|$ is a kernel-dependent quantity ($\mathbf{K}_A := [k(x,x')]_{x,x' \in A}$) which, in the Gaussian setting, is the mutual information between f and observations of Y at the most informative choice of n points. When the kernel satisfies $k(x,x') \leq 1$, the following bounds are known (Srinivas et al. 2010): $\gamma_n = \mathcal{O}(d \log n)$ for the linear kernel, $\gamma_n = \mathcal{O}((\log n)^{d+1})$ for the squared exponential kernel, and $\gamma_n = \mathcal{O}(n^{d(d+1)/(2\nu + d(d+1))}(\log n))$ for the Matérn kernel with smoothness parameter ν. Note that while f^* is not required to be drawn from our prior and ε may be non-Gaussian, this result assumes the kernel k and noise-level σ are correctly set.

Proof of Theorem 12. Fix $x, T(x) \in \mathcal{C}$, and define $\delta := (n+1)\cdot\exp\left(-\frac{[\Phi^{-1}(\alpha)]^2 - 2||f^*||_k^2}{300\gamma_n}\right)$.
In this case, $-\sqrt{\beta_{n+1}} = \Phi^{-1}(\alpha)$ (see definition in Theorem 14).

Theorem 14 implies that with probability $\geq 1 - \delta$:
$|\mu_n(x) - f^*(x)| \leq -\Phi^{-1}(\alpha)\cdot\sigma_n(x)$ and $|\mu_n(T(x)) - f^*(T(x))| \leq -\Phi^{-1}(\alpha)\cdot\sigma_n(T(x))$

Since our posterior is Gaussian:

$$F^{-1}_{G_x(T)}(\alpha) = \mu_n(T(x)) - \mu_n(x) + \Phi^{-1}(\alpha)\left[\sigma_n^2(T(x)) + \sigma_n^2(x) - 2\sigma_n(x, T(x))\right]^{1/2}$$

Therefore:

$$\begin{aligned}
&f^*(T(x)) - f^*(x) - F^{-1}_{G_x(T)}(\alpha) \\
=&f^*(T(x)) - \mu_n(T(x)) + \mu_n(x) - f^*(x) - \Phi^{-1}(\alpha)\left[\sigma_n^2(T(x)) + \sigma_n^2(x) - 2\sigma_n(x, T(x))\right]^{1/2} \\
\leq&f^*(T(x)) - \mu_n(T(x)) + \mu_n(x) - f^*(x) - \Phi^{-1}(\alpha)\left[\sigma_n^2(T(x)) + \sigma_n^2(x) + 2\sqrt{\sigma_n^2(x)\sigma_n^2(T(x))}\right]^{1/2}
\end{aligned}$$

since we assume $\alpha \leq 0.5 \Rightarrow \Phi^{-1}(\alpha) \leq 0$, and $\left|\sigma_n(x, T(x))\right| \leq \sqrt{\sigma_n^2(x)\sigma_n^2(T(x))}$

$$= f^*(T(x)) - \mu_n(T(x)) + \mu_n(x) - f^*(x) - \Phi^{-1}(\alpha)\Big[\sigma_n(T(x)) + \sigma_n(x)\Big]$$
$$= \big[f^*(T(x)) - \mu_n(T(x)) - \Phi^{-1}(\alpha)\sigma_n(T(x))\big] + \big[\mu_n(x) - f^*(x) - \Phi^{-1}(\alpha)\sigma_n(x)\big]$$

which is less than 0 with probability at most δ. □

4.7.1 Auxiliary lemmas

Theorem 13 (van der Vaart & van Zanten (2011)). *The assumptions of Theorem 10 imply:*

$$\mathbb{E}_{\mathcal{D}_n} \int\int_{\mathcal{C}} [f(x) - f^*(x)]^2 p_X(x) \mathrm{d}x \; \mathrm{d}\Pi_n(f \mid \mathcal{D}_n) \;\leq\; C \cdot \Psi_{f^*}(n)$$

where $\Psi^{-1}_{f^*}(n)$ is defined as above. See van der Vaart & van Zanten (2011) for a detailed discussion about this function.

Theorem 14 (Srinivas et al. (2010)). *Assume conditions (A24)-(A29), fix $\delta \in (0,1)$, and define:*

$$\beta_n := 2||f^*||_k^2 + 300\gamma_n[\log(n/\delta)]^3$$

Then:
$$\Pr\Big[\forall x \in \mathcal{C} :\; |\mu_n(x) - f^*(x)| \leq \sqrt{\beta_{n+1}}\sigma_n(x)\Big] \geq 1 - \delta$$

Lemma 13. *Under the assumptions of Theorem 11, for any $x, T(x) \in \mathcal{C}$:*

$$\mathbb{E}_{\mathcal{D}_n}\Big|F^{-1}_{G_n(T)}(\alpha) - F^{-1}_{G_X(T)}(\alpha)\Big| \leq C \cdot \Big[\frac{-L^2 d}{n}\log(1-\alpha)\Big]^{1/2}$$

Proof. Define random variables $Z_i := f(T(x^{(i)}) - f(x^{(i)}) \mid \mathcal{D}_n$ for $i = 1, \ldots, n$.
Note that these variables all share the same expectation: $\mathbb{E}_X[Z] := \mathbb{E}_X[Z_i] = G_X(T)$ and $G_n(T) = \frac{1}{n}\sum_{i=1}^n Z_i$. The Lipschitz continuity of f combined with the fact that $\mathcal{C} = [0,1]^d$ implies: $Z_i \in [-L\sqrt{d}, L\sqrt{d}]$ for all i. Thus, Hoeffding's inequality ensures:

$$\Pr\left(\left|G_n(T) - G_X(T)\right| \geq t\right) \leq 2\exp\left(\frac{-nt^2}{2L^2 d}\right)$$
$$\Rightarrow F^{-1}_{\left|G_n(T) - G_X(T)\right|}(\alpha) \leq C \cdot \Big[\frac{-L^2 d}{n}\log(1-\alpha)\Big]^{1/2}$$

Because posteriors $G_n(T), G_X(T)$ follow a Gaussian distribution:

$$F^{-1}_{G_n(T)}(\alpha) - F^{-1}_{G_X(T)}(\alpha) \leq F^{-1}_{\left|G_n(T)-G_X(T)\right|}(\alpha)$$
$$\text{and } F^{-1}_{G_X(T)}(\alpha) - F^{-1}_{G_n(T)}(\alpha) \leq F^{-1}_{\left|G_n(T)-G_X(T)\right|}(\alpha) \qquad \square$$

Lemma 14. *Under the assumptions of Theorem 11, for any $x, T(x) \in \mathcal{C}$:*

$$\mathbb{E}_{\mathcal{D}_n}\left|F^{-1}_{G_X(T)}(\alpha) - G^*_X(T)\right| \leq \frac{C}{\alpha} \cdot \left(L + \frac{1}{a}\right) \cdot [\Psi_{f^*}(n)]^{1/[2(d+1)]}$$

Proof of Lemma 14. A similar argument as the proof of Theorem 10 applies here. We again first bound:

$$\begin{aligned}
&\int_{\mathcal{C}} |f(x) - f^*(x)| p_X(x)\, \mathrm{d}x \\
\geq & a \cdot \mathrm{Vol}(\mathcal{B}_\delta) \cdot \left[\int_{\mathcal{C}} |f(x) - f^*(x)| p_X(x)\, \mathrm{d}x + \int_{\mathcal{C}} |f(T(x)) - f^*(T(x))| p_X(x)\, \mathrm{d}x - 8\delta L\right] \\
\geq & a \cdot \mathrm{Vol}(\mathcal{B}_\delta) \cdot \left[\left|\mathbb{E}_X[f(x) - f^*(x)] + \mathbb{E}_X[f(T(x)) - f^*(T(x))]\right| - 8\delta L\right]
\end{aligned}$$

Following the same reasoning as in the proof of Theorem 10, we obtain (up to constant factors):

$$-aL\delta^{d+1} + a\alpha\delta^d \cdot \mathbb{E}_{\mathcal{D}_n}\left[G^*_X(T) - F^{-1}_{G_X(T)}(\alpha)\right] \leq [C \cdot \Psi_{f^*}(n)]^{1/2}$$

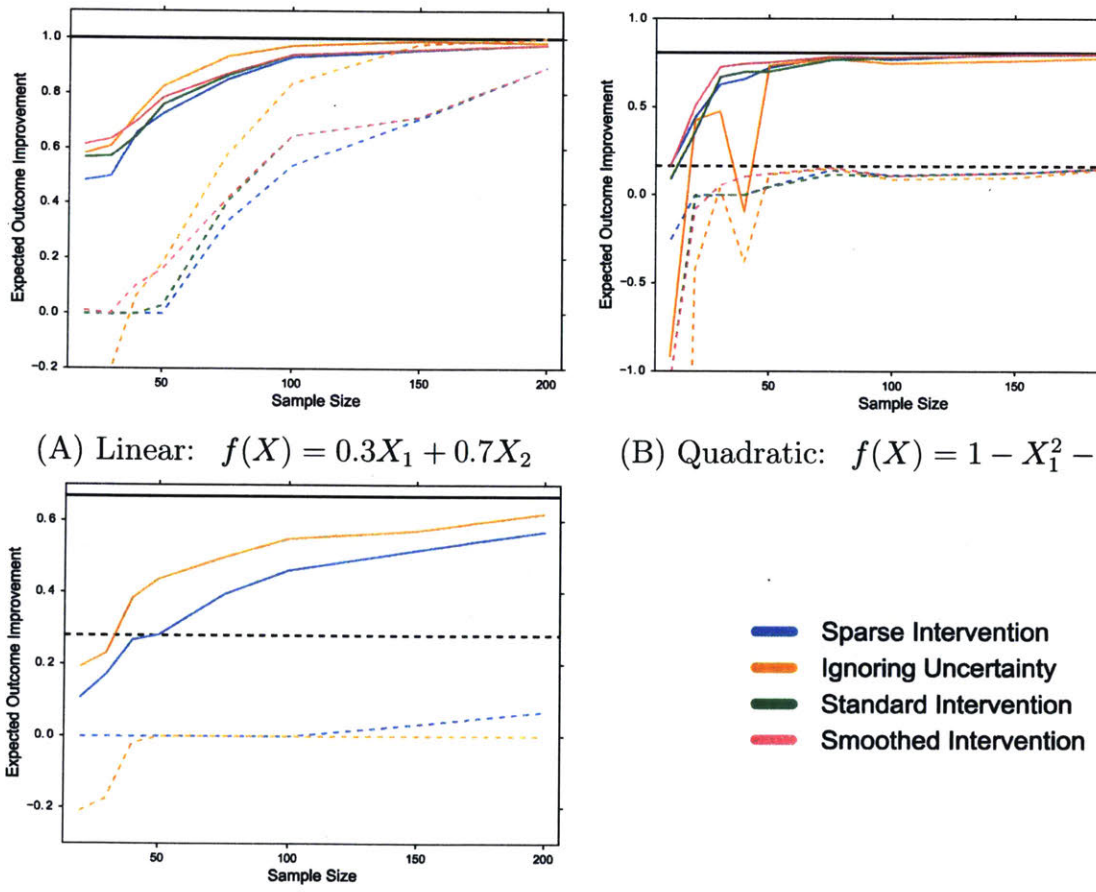
and we can use the same argument to similarly bound $\mathbb{E}_{\mathcal{D}_n}\left[F^{-1}_{G_X(T)}(\alpha) - G^*_X(T)\right]$. $\square$

4.8 Empirical results

4.8.1 Simulation study

We first apply our approach to simulated data from simple covariate-outcome relationships, where the average improvement produced by our chosen interventions rapidly converges to the best possible value with increasing n. In these experiments, sparse-interventions consistently alter the correct feature subset, and proposed transformations under our conservative $\alpha = 0.05$ criterion are much more rarely harmful than those suggested by optimizing the posterior mean function (which ignores uncertainty).

Over the simulated data summarized in Figure 4-2, we apply our basic personalized intervention method ($\alpha = 0.05$) with purely local optimization (standard) and our continuation technique (smoothed), which significantly improves results. For each of the 100 datasets, we randomly sampled a new point (from the same underlying distribution) to receive a personalized intervention. The magnitude of each intervention is bounded by 1, except for in data from the quadratic relationship. We also infer sparse interventions (with a cardinality constraint of 2 for the linear and quadratic relationships, 1 for the product relationship). When $Y = X_1 \cdot X_2 + \varepsilon$, the optimal (constrained) intervention may drastically vary depending upon the individual's covariate-values, and our algorithm is able to correctly infer this behavior



(A) Linear: $f(X) = 0.3X_1 + 0.7X_2$

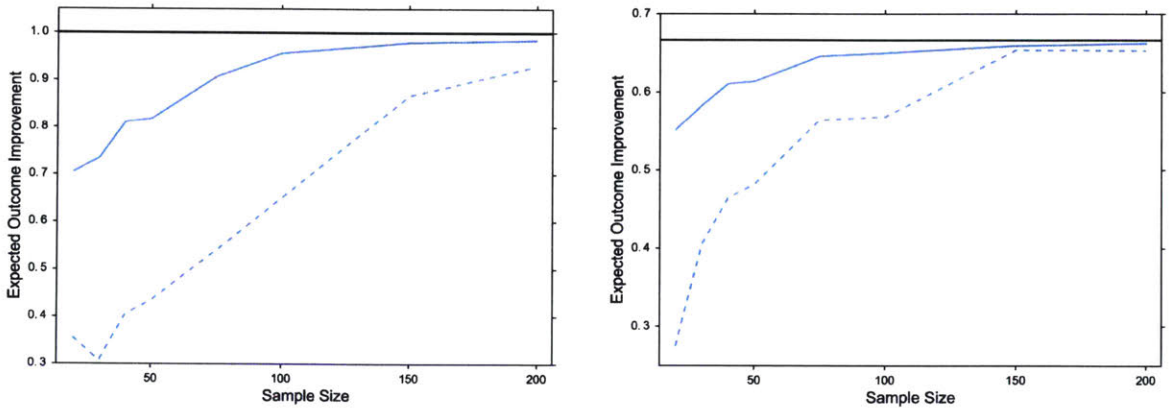
(B) Quadratic: $f(X) = 1 - X_1^2 - X_2^2$

(C) Product: $f(X) = X_1 \cdot X_2$

Figure 4-2: The mean (solid) and 0.05^{th} quantile (dashed) expected outcome change produced under personalized interventions suggested by various methods, over 100 datasets of each sample size. Each dataset contains 10-dimensional covariates, with $X_i \sim \text{Unif}[-1, 1]$, and Y is determined by the indicated relationships and additive Gaussian noise ($\sigma = 0.2$). The black lines indicate the best possible expected outcome change (when the best change depends on which individual received the intervention, the black solid/dashed lines indicates the mean and 0.05^{th} quantile over our 100 trials).

(Simulation C). Finally, we also apply a variant of our method which entirely ignores uncertainty ($\alpha = 0.5$). While this approach is on average better for larger sample sizes, highly harmful interventions are occasionally proposed, whereas our uncertainty-adverse method ($\alpha = 0.05$) is much less prone to producing damaging interventions (preferring to abstain by returning $T(x) = x$ instead). This is a critical property since interventions generally require effort and are only worth conducting when they are likely to produce a substantial benefit.

Figure 4-3 displays the behavior of both the population shift intervention in the linear setting, and the population covariate-fixing intervention under the quadratic relationship. The population intervention is notably safer than the individually tailored variants, producing no negative changes in our experiments.



(A) Linear: $f(X) = 0.3X_1 + 0.7X_2$ (B) Quadratic: $Y = 1 - X_1^2 - X_2^2$

Figure 4-3: The mean (solid) and 0.05^{th} quantile (dashed) expected outcome change produced by our population intervention method, over 100 datasets for each sample size (same setting as in Figure §4-2). The black line indicates the best possible expected outcome improvement.

4.8.2 Gene perturbation

Next, we applied our methodology to search for population interventions in observational yeast gene expression data from Kemmeren et al. (2014). We evaluated the effects of proposed interventions (restricted to single gene knockouts) over a set X of 10 transcription factors ($n = 161$) with the goal of down-regulating each of a set of 16 downstream small molecule metabolism genes (each of these is treated as a separate outcome Y). The data set used for this analysis contains gene expression levels for a set of wild type (i.e. 'observational') samples, $\mathcal{D}_{obs}$ ($n = 161$), as well as for a set of 'interventional' samples, $\mathcal{D}_{int}$, in which each individual gene was serially knocked out.

In our analysis, we search for potential interventions for affecting the expression of a desired target gene by training our GP regressor on $\mathcal{D}_{obs}$ and determining which knockout produces the best value of our empirical covariate-fixing population intervention objective (for down-regulating the target). Subsequently, we use $\mathcal{D}_{int}$ to evaluate the actual effectiveness of proposed interventions in the knockout experiments. We only search for interventions present in $\mathcal{D}_{int}$ (single gene knockouts) rather than optimizing to infer optimal covariate transformations. Each method evaluated in this analysis was to propose an intervention (single gene knockout) to down-regulate the expression of each target gene (separately). Once a gene to knock out was proposed, this intervention was evaluated by comparing the resulting expression of the target when the proposed knockout was actually performed in the experimental data $\mathcal{D}_{int}$. This expression level could then be compared to the 'optimal' choice of transcription factor from X to intervene upon (the transcription factor whose knockout produced the largest down-regulation of the target in $\mathcal{D}_{int}$). Note that this application represents a setting with complex underlying causal relationships, where it is likely that many of our stringent causal assumptions in §4.1 are severely violated.

We compared our approach against two methods commonly used to identify candidate genes for knockout by biologists (which are also more broadly used to draw conclusions about affecting outcomes across the sciences). First, we applied a multivariate regression analysis in which a linear regression model was fit to the observations of (X, Y) in $\mathcal{D}_{obs}$. The best gene to knockout was inferred on the basis of the regression coefficients and expression values (if no beneficial regression coefficient was found significant at the 0.05 level under the standard t-test, then no intervention was proposed). Second, we performed a marginal analysis in which separate univariate linear regression models were fit to $(X_1, Y), \ldots, (X_d, Y)$, and the best knockout was again inferred on the basis of the regression coefficients and expression values (again, no intervention was recommended if there was no statistically significant beneficial regression coefficient at the 0.05 level, after correcting for multiple testing via the False Discovery Rate).

Figure 4-4 compares the results produced by these methods to the optimal intervention over X for down-regulating each Y, as found in the experimental data

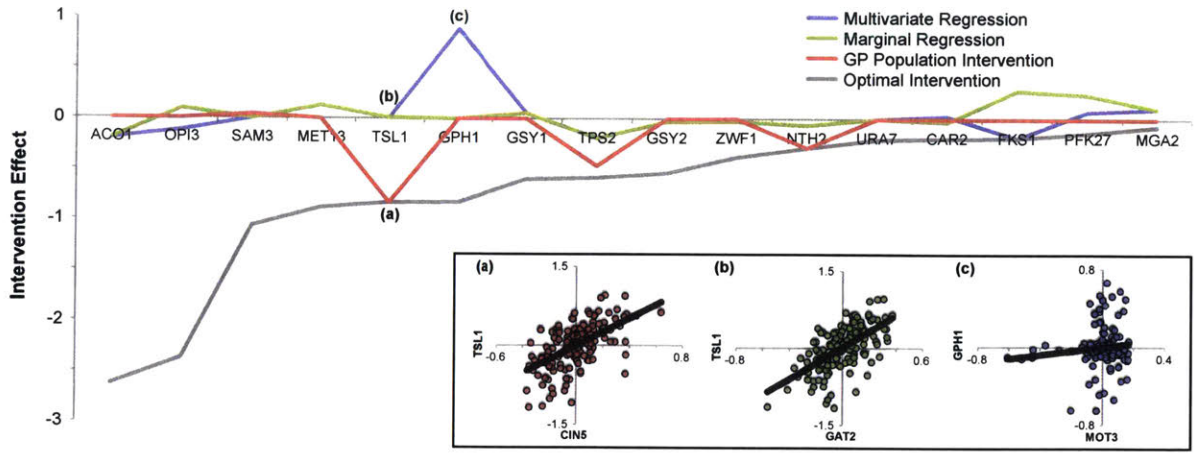


Figure 4-4: Actual effects of proposed interventions (single gene knockout) over a set transcription factors on down-regulation of each of a set of 16 small molecule metabolism target genes. Insets (a) and (b) show empirical marginal distributions between target gene *TSL1* and the transcription factors identified for knockout by our GP population intervention method (*CIN5*) and the marginal regression approach (*GAT2*). Inset (c) shows the empirical marginal distributions between target gene *GPH1* and *MOT3*, which was the transcription factor knockout proposed by the multivariate regression approach.

$\mathcal{D}_{int}$. The gray curve in the figure illustrates the maximal intervention effect on each downstream target gene that could have been achieved by selecting the right transcription factor to knock out. Of the 16 small molecule metabolism target genes tested, in three cases our method proposed an intervention which was found to be optimal or near optimal in $\mathcal{D}_{int}$, while in the remaining cases, the model uncertainty causes the method not to recommend any intervention (except for one very minorly harmful intervention for target *SAM3*). On the other hand, neither form of linear regression proposed effective interventions for any target other than *FKS1*, and in quite a few cases, the linear regressors proposed counterproductive interventions that up-regulated the target.

Compared to marginal linear regressions and multivariate linear regression, our method's uncertainty prevents it from proposing harmful interventions, and the interventions it proposes are optimal or near optimal (Figure 4-4). Insets (a) and (b) in Figure 4-4 depict the empirical marginal distributions between target gene *TSL1* and members of X identified for knockout by our method (*CIN5*) and the marginal regression approach (*GAT2*). From the linear perspective, these relationships are fairly indistinguishable, but only *CIN5* displays a strong inhibitory effect in the knockout experiments. Inset (c) shows the empirical marginal for a harmful intervention proposed by multivariate regression for down-regulating *GPH1*, where the overall correlation is significantly positive, but the few lowest expression values (which influence our GP intervention objective the most) do not provide strong evidence of a large knockdown effect. This highlights the importance of a model that properly accounts for uncertainty in the covariate space when evaluating potential transformations of

covariate-values, especially when the data are extremely noisy as in the case of these gene expression measurements.

4.8.3 Writing improvement

In this section, we apply our personalized intervention methodology to the task of transforming a given news article into one which will be more widely-shared on social media. We use a dataset from Fernandes et al. (2015) that consists of 39,000 news articles published by Mashable around 2013-15 (Fernandes et al. 2015). Each article is annotated with the number of shares it received in social networks (which we use as our outcome variable after log-transform and rescaling). A multitude of features have been extracted from each article (e.g. word count, the category such as "tech" or "lifestyle", keyword properties), many of which Fernandes et al. (2015) produced using natural language processing algorithms (e.g. subjectivity, polarity, alignment with topics found by Latent Dirichlet Allocation). After removing many highly redundant covariates, we center and rescale all variables to unit-variance (see Table 4.2 for a complete description of the 29 covariates used in this analysis).

We randomly partition the articles into 3 disjoint groups: a *training* set (5,000 articles on which scaling-factors are computed and our GP regressor is trained), an *improvement* set (300 articles we find interventions for), and a *held-out* set (over 34,000 articles used for evaluation). A large group is left out for validation to ensure there are many near-neighbors for any specified setting of article-covariates. Subsequently, a basic GP regression model is fitted to the training set. Over the held-out articles, the Pearson correlation between the observed popularity and the GP (posterior mean) predictions is 0.35. Furthermore, there is a highly significant ($p < 8 \cdot 10^{-41}$) positive correlation of 0.07 between the model's predictive variance and the actual squared errors of GP predictions over this held-out set. Our model is thus able to make reasonable predictions of popularity based on the available covariates, and its uncertainty estimates tend to be larger in areas of the feature-space where the posterior mean lies further from actual popularity values.

In this analysis, we compare our personalized intervention methodology which *rejects* uncertainty (using $\alpha = 0.05$) with a variant of the this approach that *ignores* uncertainty (using the same objective function with $\alpha = 0.5$). Both methods share the same GP regressor, optimization procedure, and set of constraints. Note that this latter approach corresponds to the popular strategy of fitting a supervised learning model to predict outcomes and then optimizing features with respect to the predicted outcomes without regard for predictive uncertainty. For the 300 articles in the intervention set (not part of the training data) we allow intervening upon all covariates except for the article category which presumably is fixed from an author's perspective. All covariate-transformations are constrained to lie within [-2,2] standard deviations of the original covariate value, and we impose a sparsity constraint that at most 10 covariates can be intervened upon for a given article.

Unfortunately, no pre-and-post-intervention articles are available for us to ascertain a ground truth evaluation. To crudely measure performance, we estimate the underlying expected popularity of a given covariate-setting using *benchmark popularity*: the (weighted) average observed popularity amongst 100 nearest neighbors (in the feature-space) from the set of held-out articles (with weights based on inverse Euclidean distance). Note that such a matching strategy is widely employed for inferring treatment effects in the causal inference literature. Over our improvement set, the Pearson correlation between articles' observed popularity and benchmark popularity is 0.28 (highly significant: $p \leq 2 \cdot 10^{-10}$). This approach thus appears to be, on average, a reasonable way to benchmark performance (even though nearest-neighbor held-out articles can individually differ from the text of a particular pre/post-intervention article despite sharing similar values of our 29 measured covariates).

Figure 4-5 depicts the results produced by our personalized intervention for each article in our intervention set. The expected improvement produced by a particular intervention is defined as the difference between the benchmark popularity of the post-intervention covariate-settings and the original covariate-settings of the article receiving the personalized intervention. Table 4.1 summarizes these results. A paired-sample t-test suggests our method is significantly superior on average ($p < 2 \cdot 10^{-6}$).

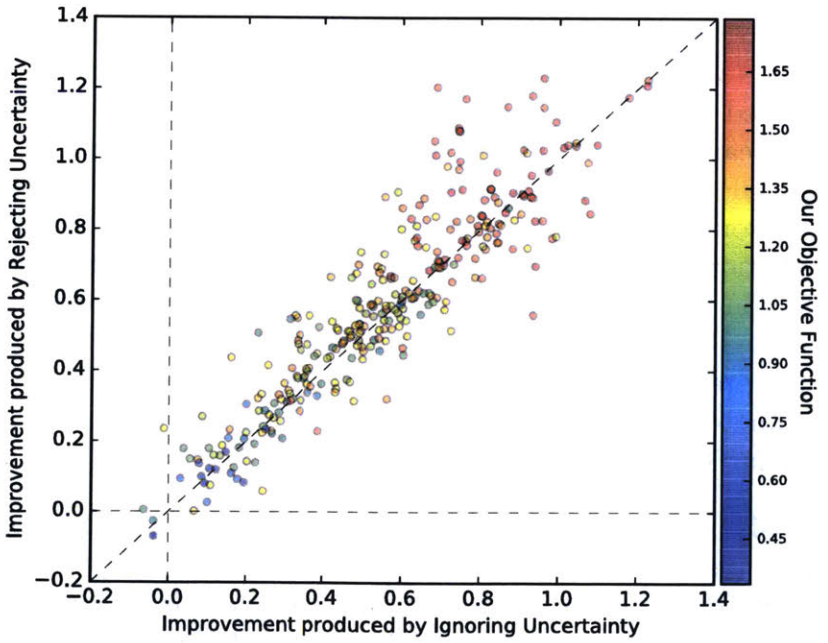


Figure 4-5: Benchmark popularity changes produced by the personalized interventions for 300 articles suggested by our method with $\alpha = 0.05$ (Rejecting Uncertainty) vs. $\alpha = 0.5$ (Ignoring Uncertainty). The points (i.e. articles) are colored according to the value of our personalized intervention objective with $\alpha = 0.05$. Using $\alpha = 0.05$ outperforms $\alpha = 0.5$ in this analysis in 177/300 articles in the improvement set.

When $\alpha = 0.05$, the average benchmark popularity increase produced by our personalized intervention methodology is 0.59, whereas it statistically significantly

decreases to 0.55 if $\alpha = 0.5$ is chosen. Thus, even given this large sample size, ignoring uncertainty appears detrimental for this application, and $\alpha = 0.5$ results in 4 articles whose benchmark popularity worsens post-intervention (compared to only 2 for $\alpha = 0.05$). Nonetheless, both methods generally produce very beneficial improvements in this analysis, as seen in Figure 4-5.

Method	Mean	Median	0.05^{th} Quantile	Num. Negative
Rejecting Uncertainty	0.586	0.578	0.126	2
Ignoring Uncertainty	0.552	0.555	0.105	4

Table 4.1: Summary statistics for the benchmark popularity change produced by each method over the 300 articles of the intervention set. The last column counts the number of harmful interventions (with change < 0).

As an example of the personalization of proposed interventions, our method ($\alpha = 0.05$) generally proposes different sparse interventions for articles in the Business category vs. the Entertainment category. On average, the sparse transformation for business articles uniquely advocates decreasing global sentiment polarity and increasing word count (which are not commonly altered in the personalized interventions found for entertainment articles), whereas interventions to decrease title subjectivity are uniquely prevalent throughout the entertainment category. These findings appear intuitive (e.g. critical business articles likely receive more discussion, and titles of popular entertainment articles often contain startling statements written non-subjectively as fact). Interestingly, the model also tends to advise shorter titles for business articles, but increasing the length for entertainment articles. Articles across all categories are universally encouraged to include more references to other articles and keywords that were historically popular.

To provide concrete examples, we present some articles of the Business and Entertainment categories (taken from our improvement set). For this business article: `http://mashable.com/2014/07/30/how-to-beat-the-heat/`, our method proposes shifting the following 10 covariates (see Table 4.2 for feature descriptions):

num_hrefs: +2, num_self_hrefs: -1.25, average_token_length: -1.771,
kw_avg_min: +1.71, kw_avg_avg: +2,
self_reference_min_shares: +2, self_reference_max_shares: +1.68,
self_reference_avg_sharess: +2, global_subjectivity: +1.57,
global_sentiment_polarity: -2

For this entertainment article: `http://mashable.com/2014/07/30/how-to-beat-the-heat/`, our method proposes shifting the following 10 covariates:

average_token_length: -1.55, kw_avg_min: + 1.63, kw_avg_avg: +2,
self_reference_min_shares: +2 self_reference_max_shares: +1.85,

self_reference_avg_shares: +2.0, LDA_00: +1.63, LDA_01: -2, LDA_04: +0.82, global_subjectivity: +1.62

Indifferent to uncertainty, the method with $\alpha = 0.5$ advocates shifting all these covariates by the ± 2 maximal allowed amounts, which leads to a 0.04 worse improvement in benchmark popularity compared with the covariate-changes specified above for this article.

Feature	Description
n_tokens_title	Number of words in the title
n_tokens_content	Number of words in the content
n_unique_tokens	Rate of unique words in the content
n_non_stop_words	Rate of non-stop words in the content
num_hrefs	Number of links
num_self_hrefs	Number of links to other articles published by Mashable
average_token_length	Average length of the words in the content
num_keywords	Number of keywords in the metadata
data_channel_is_lifestyle	Is the article category "Lifestyle"?
data_channel_is_entertainment	Is the article category "Entertainment"?
data_channel_is_bus	Is the article category "Business"?
data_channel_is_socmed	Is the article category "Social Media"?
data_channel_is_tech	Is the article category "Tech"?
data_channel_is_world	Is the article category "World"?
kw_avg_min	Avg. shares of articles with the least popular keyword used for this article
kw_avg_max	Avg. shares of articles with the most popular keyword used for this article
kw_avg_avg	Avg. shares of the average-popularity keywords used for this article
self_reference_min_shares	Min. shares of referenced articles in Mashable
self_reference_max_shares	Max. shares of referenced articles in Mashable
self_reference_avg_shares	Avg. shares of referenced articles in Mashable
LDA_00	Closeness to first LDA topic
LDA_01	Closeness to second LDA topic
LDA_02	Closeness to third LDA topic
LDA_03	Closeness to fourth LDA topic
LDA_04	Closeness to fifth LDA topic
global_subjectivity	Subjectivity score of the text
global_sentiment_polarity	Sentiment polarity of the text
title_subjectivity	Subjectivity score of title
title_sentiment_polarity	Sentiment polarity of title

Table 4.2: The 29 covariates of each article (dimensions of X in this analysis). Features involving the share-counts of other articles and LDA were based only on data known before the publication date.

4.9 Misspecified interventions

Our methodology heavily relies on the assumption that the outcome-determining covariate values $\widetilde{x}$ produced through intervention exactly match the desired covariate transformation $T(x)$. When transformations are only allowed to alter at most $k < d$ covariates, this requires that we can intervene to alter only this subset without affecting the values of other covariates. If T specifies a sparse change affecting only a subset of the covariates $\mathcal{I} \subset \{1, \ldots, d\}$, our methods assume the post-treatment value of any non-intervened-upon covariate remains at its initial value (i.e. $\widetilde{x}_s = x_s \; \forall s \notin \mathcal{I}$).

In some domains, the covariate-transformation induced via sparse external intervention can only be roughly controlled (e.g. our gene perturbation example when the profiled genes belong to a common regulatory network). Let $T_{\mathcal{I} \to z}$ denote a covariate-fixing transformation which sets a subset of covariates in $\mathcal{I} \subset \{1, \ldots, d\}$ to constant values $z_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ across all individuals in the population. In this section, we consider an alternative assumption under which the intervention applied in hopes of achieving $T_{\mathcal{I} \to z}$ propagates downstream to affect other covariates outside $\mathcal{I}$ (so there may exist $s \notin \mathcal{I}$: $\widetilde{x}_s \neq x_s$), which we formalize as the *do*-operation in the causal calculus of Pearl (2000). Here, we suppose the underlying population of X, Y follows a *structural equation model* (SEM) (Pearl 2000). The outcome Y is restricted to be a sink node of the causal DAG, so we can still write $Y = f^*(\widetilde{X}) + \varepsilon$ and maintain the other conditions from §4.1. Rather than exhibiting covariate-distribution $T_{\mathcal{I} \to z}(X)$ with $Y = f^*(T_{\mathcal{I} \to z}(X)) + \varepsilon$ (as presumed in our methods), the post-treatment population which arises from an intervention seeking to enact transformation $T_{\mathcal{I} \to z}$ is now assumed to follow the distribution specified by $p(X, Y \mid do(X_{\mathcal{I}} = z_{\mathcal{I}}))$. Note that the *do*-operation here is only applied to some nodes in the DAG (variables in subset $\mathcal{I}$) as discussed by Peters et al. (2014), but its effects can alter the distributions of non-intervened-upon covariates outside of $\mathcal{I}$ which lie downstream in the DAG.

For settings where sparse interventions elicit unintentional *do*-effects and the causal DAG meets condition (A31), Theorems 15 and 16 below imply that, under complete certainty about f^*, the (minimum cardinality) maximizer of our covariate-fixing intervention objective corresponds to an transformation that produces an equally good outcome change when the corresponding intervention is actually realized as a *do*-operation in the underlying population. Combined with Theorem 11, our results ensure that, even in this misspecified setting, the empirical maximizer of our sparse covariate-fixing intervention objective (4.7) produces (in expectation as $n \to \infty$) beneficial interventions for populations whose underlying causal relationships satisfy certain conditions.

We let $\text{pa}(Y)$ denote the variables which are parents of outcome Y in the underlying causal DAG, and $\text{desc}(\mathcal{I})^C$ denotes the set of variables which are *not* descendants

of variables in subset $\mathcal{I}$. We also define:

$$\mathcal{I}^* := \operatorname{argmin}\left\{|\mathcal{I}'| \text{ s.t. } \exists\ T_{\mathcal{I}'\to z} \in \operatorname*{argmax}_{T_{\mathcal{I}\to z}:|\mathcal{I}|\le k} \mathbb{E}_X\left[f^*(T_{\mathcal{I}\to z}(x)) - f^*(x)\right]\right\} \tag{4.18}$$

as the intervention set corresponding to the optimal k-sparse covariate-fixing transformation (where in the case of ties, the set of smallest cardinality is chosen), if transformations were exactly realized by our interventions (which is no longer assumed to be the case in this section). Our theory considers the following structural properties of the underlying causal DAG:

(A30) For a given $\mathcal{I} \subseteq \{1,\dots,d\}$: $pa(Y) \subseteq \mathcal{I} \bigcup \operatorname{desc}(\mathcal{I})^C$
which ensures that each variable which is parent of Y either belongs to the chosen intervention-set $\mathcal{I}$, or is otherwise not a descendant of the variables in this set.

(A31) No variable in $\operatorname{pa}(Y)$ is a descendant of other parents, ie. $\nexists\ j \in \operatorname{pa}(Y)$ s.t. $j \in \operatorname{desc}(\operatorname{pa}(Y) \setminus \{j\})$.

Theorem 15. *For any covariate-fixing transformation $T_{\mathcal{I}\to z}$ where $\mathcal{I}$ satisfies (A30):*

$$\mathbb{E}_X\left[f^*(T_{\mathcal{I}\to z}(x)) - f^*(x)\right] = \mathbb{E}_{\widetilde{x}\sim do(X_{\mathcal{I}}=z_{\mathcal{I}})}\left[f^*(\widetilde{x})\right] - \mathbb{E}_X\left[f^*(x)\right]$$

Proof. We employ subscripts to index particular covariates of X. The notation $[a_R, a_S] = a \in \mathbb{R}^d$ is used to denote a vector assembled from disjoint subsets of dimensions $R, S \subseteq \{1,\dots,d\}$. Regardless of the ordering of these partitions in our notation, we assume they are correctly arranged in the assembled vector based on their subscript-indices (i.e. $a = [a_R, a_S] = [a_S, a_R]$). We have:

$$\begin{aligned}
&\mathbb{E}_{\operatorname{do}(X_{\mathcal{I}}=z_{\mathcal{I}})}\left[f^*(x)\right] \\
&= \int f^*\left([x_{\mathcal{I}^C}, z_{\mathcal{I}}]\right)\, p\left(x_{\mathcal{I}^C} \mid do(X_{\mathcal{I}} = z_{\mathcal{I}})\right)\, \mathrm{d}x_{\mathcal{I}^C} \\
&= \iint f^*\left([x_{\operatorname{pa}(Y)\setminus\mathcal{I}}, z_{\mathcal{I}\cap\operatorname{pa}(Y)}, a_{\mathcal{I}^C\setminus\operatorname{pa}(Y)}]\right) \cdot p\left(x_{\mathcal{I}^C\setminus\operatorname{pa}(Y)} \mid x_{\operatorname{pa}(Y)\setminus\mathcal{I}}, do(X_{\mathcal{I}} = z_{\mathcal{I}})\right) \\
&\qquad\qquad \cdot p\left(x_{\operatorname{pa}(Y)\setminus\mathcal{I}} \mid do(X_{\mathcal{I}} = z_{\mathcal{I}})\right)\, \mathrm{d}x_{\mathcal{I}^C\setminus\operatorname{pa}(Y)}\, \mathrm{d}x_{\operatorname{pa}(Y)\setminus\mathcal{I}} \\
&\quad \text{where covariate-subset } a_{\mathcal{I}^C\setminus\operatorname{pa}(Y)} \text{ can take arbitrary values} \\
&\quad \text{since } f^* \text{ is constant along covariates } \notin \operatorname{pa}(Y) \\
&= \int f^*\left([x_{\operatorname{pa}(Y)\setminus\mathcal{I}}, z_{\mathcal{I}\cap\operatorname{pa}(Y)}, a_{\mathcal{I}^C\setminus\operatorname{pa}(Y)}]\right)\, p\left(x_{\operatorname{pa}(Y)\setminus\mathcal{I}} \mid do(X_{\mathcal{I}} = z_{\mathcal{I}})\right)\, \mathrm{d}x_{\operatorname{pa}(Y)\setminus\mathcal{I}} \\
&= \int f^*\left([x_{\operatorname{pa}(Y)\setminus\mathcal{I}}, z_{\mathcal{I}\cap\operatorname{pa}(Y)}, a_{\mathcal{I}^C\setminus\operatorname{pa}(Y)}]\right)\, p\left(x_{\operatorname{pa}(Y)\setminus\mathcal{I}}\right)\, \mathrm{d}x_{\operatorname{pa}(Y)\setminus\mathcal{I}} \\
&\quad \text{since the marginal distribution over } X_{\operatorname{pa}(Y)\setminus\mathcal{I}} \text{ equals the } do\text{-distribution by (A30)} \\
&= \iint f^*\left([x_{\operatorname{pa}(Y)\setminus\mathcal{I}}, z_{\mathcal{I}\cap\operatorname{pa}(Y)}, x_{\mathcal{I}^C\setminus\operatorname{pa}(Y)}]\right) p\left(x_{\mathcal{I}^C\setminus\operatorname{pa}(Y)} \mid x_{\operatorname{pa}(Y)\setminus\mathcal{I}}\right) p\left(x_{\operatorname{pa}(Y)\setminus\mathcal{I}}\right) \mathrm{d}x_{\mathcal{I}^C\setminus\operatorname{pa}(Y)} \mathrm{d}x_{\operatorname{pa}(Y)\setminus\mathcal{I}}
\end{aligned}$$

$$= \mathbb{E}_X\left[f^*(T_{\mathcal{I}\to z}(x))\right]$$ □

In the absence of extremely strong interactions between variables in $\text{pa}(Y)$, the equality of Theorem 15 will also hold for $\mathcal{I}^*$ if $|\text{pa}(Y)| \leq k$.

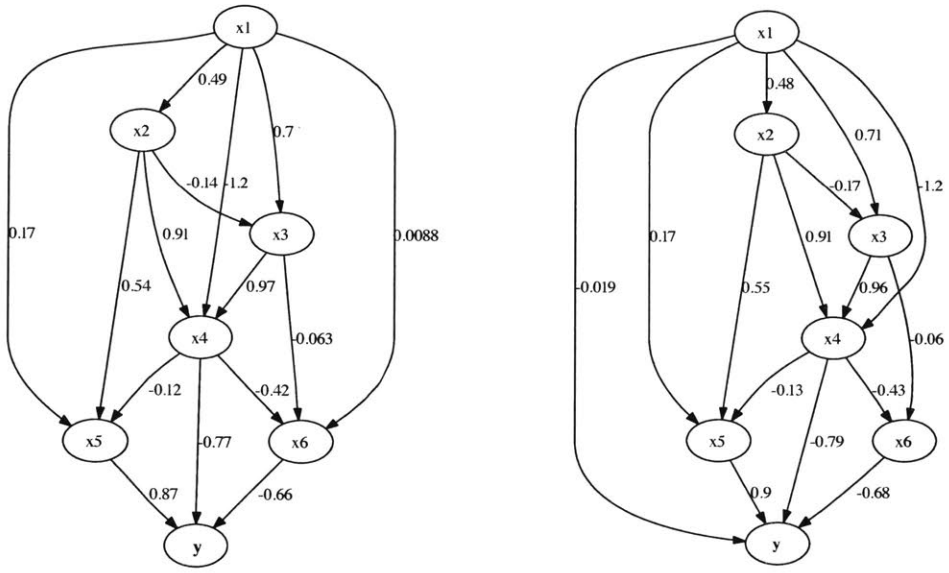
Theorem 16. *If the underlying DAG satisfies (A31), then $\mathcal{I}^*$ will satisfy (A30).*

Proof. Since $\mathbb{E}_X[f^*(T_{\mathcal{I}\to z}(x))]$ does not change when $z_j := [T_{\mathcal{I}\to z}(x)]_j$ is altered for any $j \notin \text{pa}(Y)$, including variables outside of the parent set in $\mathcal{I}$ does not improve this quantity. Thus, either $\text{pa}(Y) \subseteq \mathcal{I}^*$, or $\mathcal{I}^* \subset \text{pa}(Y)$. The first case immediately implies (A30). When $\mathcal{I}^* \subset \text{pa}(Y)$: our assumption that no variable in $\text{pa}(Y)$ is a descendant of other parents implies the other parents must belong the complement of $\text{desc}(\mathcal{I}^*)$, since this is a subset of $\text{desc}\big(\text{pa}(Y)\big)$. □

Finally, we empirically investigate how effective our methods are in this misspecified SEM setting, where a proposed sparse population transformation is actually realized as a *do*-operation and can therefore unintentionally affect other covariates in the post-intervention population. We generate data from an underlying linear *non*-Gaussian SEM, and where Y is a sink node in the corresponding causal DAG. Here, we suppose that a desired transformation upon variable $s \in \{1, \ldots, d\}$ cannot be enacted exactly and the Y which arises post-treatment is distributed according to $do(X_s = \mathbb{E}[X_s] + \Delta)$, where $\mathbb{E}[X_s]$ is the mean of the pre-treatment marginal distribution of the sth covariate. In this case, *do*-effects can propagate to other covariates which are descendants of s in the DAG because the values of descendant variables are redrawn from the *do*-distribution which arises as a result of shifting $\mathbb{E}[X_s]$. Because all relationships are linear in our SEMs, the actual expected outcome change resulting from a particular shift (resulting from the corresponding *do*-operation) is easily obtained in closed form.

Our GP intervention framework (with $\alpha = 0.05$) is applied to the data to infer an optimal 1-sparse shift population intervention (only interventions on a single variable are allowed). The maximal allowed magnitude of the shift is constrained to ensure the optimum is well-defined (to ± 1 times the standard deviation of each variable in the underlying SEM distribution). An alternative approach to improve outcomes in contrast to our black-box approach is to apply a causal inference method like LinGAM (Shimizu et al. 2006) to estimate the SEM from the data, and then identify the optimal single-variable shift Δ_s^* in the LinGAM-inferred SEM (since all inferred relationships are also linear, the optimal single-variable shift will be either 0 or the lower/upper allowed shift and we simply search over these possibilities). We compare our approach against LinGAM by evaluating the actual expected outcome change produced by the shift Δ_s^* proposed by each method (where the actual expected outcome change is found by analytically performing the $do(X_s = x_s + \Delta_s^*)$ operation in the true underlying SEM). Note that LinGAM is explicitly designed for this setting, while both our method and the relied-upon Gaussian Process model are severely misspecified.

In our experiment, two underlying SEM models are considered which were used by Shimizu et al. (2006) to demonstrate the utility of their LinGAM method (albeit with impractically large sample size = 10,000). SEM_A and SEM_B are respectively used to refer to the models depicted on the lefthand and righthand sides of Figure 4-6. In each SEM, the outcome Y is always fixed at a sink node of the causal DAG (to ensure the covariates cause the outcomes and not vice versa), while the remainder of the variables are adopted as our observed covariates X. Note that neither data-generating SEM considered here satisfies the structural condition (A31) required by our previously described theoretical guarantees.

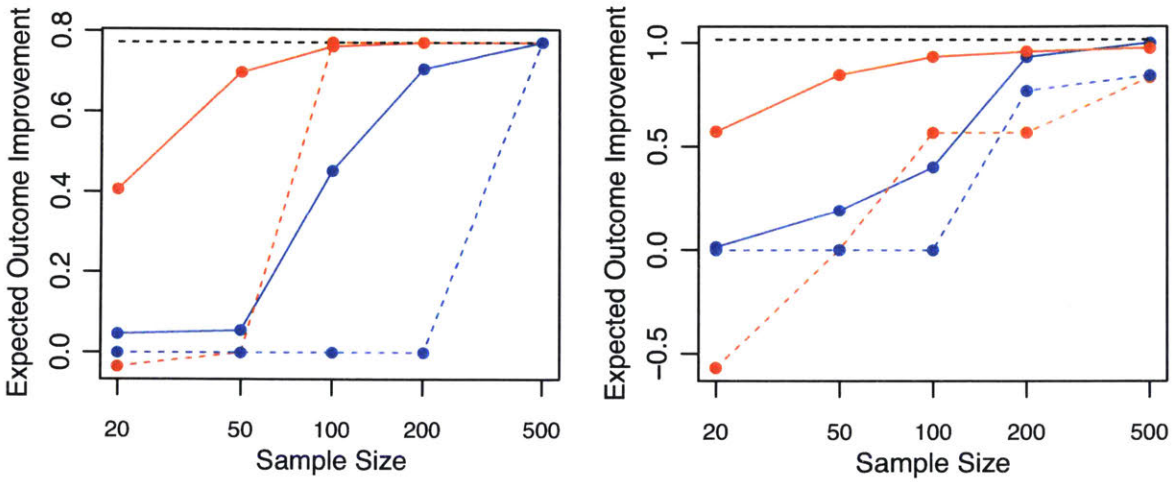


(A) SEM_A (B) SEM_B

Figure 4-6: Two linear non-Gaussian SEM structures depicted in Figure 3 of Shimizu et al. (2006).

This experiment represents an application of our method in a highly misspecified setting. The true data-generating mechanism differs significantly from assumptions of our GP regressor (output noise is now fairly non-Gaussian, the underlying relationships are all linear while we use an ARD kernel). Furthermore, an intervention to transform a single covariate incurs a multitude of unintentional off-target effects resulting from the *do*-effects propagating to downstream covariates in the SEM, whereas our method believes only the chosen covariate is changed. In contrast, this data exactly follows the special assumptions required by LinGAM, and we properly account for inferred downstream *do*-operation effects when identifying the best inferred intervention under LinGAM. The only disadvantage of the LinGAM method is that it does not a priori know the direction of the causal relationship $X \to Y$ (although we found it always estimated this direction correctly except on rare occasions with tiny sample sizes of $n = 20$).

Since LinGAM only estimates linear relations, the best inferred shift-intervention found by this approach will always be 0 or the minimal/maximal shift allowed for a particular covariate. Searching over these three values for each covariate ensures the actual optimal shift will be recovered if the LinGAM SEM-estimate were correct. However, under our approach, identifying the optimal population shift-intervention requires solving an optimization problem. Even if the GP regression posterior were to exactly reflect the true data-generating mechanism, our approach might get stuck in a suboptimal local maximum or avoid the minimal/maximal allowed shift due to too much uncertainty about f in the resulting region of feature-space. In practice, these potential difficulties do not pose much of an issue for our approach.



(A) SEM_A (B) SEM_B

Figure 4-7: The average (solid) and 0.05^{th} quantile (dashed) expected outcome change produced by our method (red) vs LinGAM (blue) over 100 datasets drawn from two underlying SEMs chosen by Shimizu et al. (2006). The black dashed line indicates the best possible improvement in each case.

Figures 4-7A and 4-7B demonstrate that the inferred best single-variable shift population intervention (under constraints on the magnitude of the shift) matches the performance the interventions suggested by LinGAM (except for in rare cases with tiny sample size) when the proposed interventions are evaluated as *do*-operations in the true underlying SEM. Thus, we believe a supervised learning approach like ours is preferable in practical applications where interpreting the underlying causal structure is not as important as producing good outcomes (especially for higher dimensional data where estimation of the causal structure becomes difficult (Peters et al. 2014)).

The assumption of sparse interventions realized as a *do*-operation (as defined by Peters et al. (2014)) may also be an inappropriate in many domains, particularly if off-target effects of interventions are explicitly mitigated via external controls. To appreciate the intricate nature of assumptions regarding non-intervened-upon variables, consider our example of modeling text documents represented using two features: polarity and word count. A desired transformation to increase the text's polarity can be accomplished by inserting additional positive adjectives, but such an intervention also

increases articles' word count. Alternatively, polarity may be identically increased by replacing words with more positive alternatives, an external intervention which would not affect the word count (and thus follows the assumptions of our framework). These two different external interventions seek to enact the same desired transformation, but neither necessarily alters the covariate-distribution in the manner presumed by the *do*-calculus).

Chapter 5

Discussion

This thesis presented various data-driven methodologies for addressing the key tasks of understanding and shaping a specific population of interest. In particular, we are able to interpretably characterize changes and identify beneficial interventions in a nonparametric fashion, without having to specify a restrictive generative model for the (often high-dimensional) distribution of measurements in the underlying (often heterogeneous) population. While primarily used for analysis of biological populations throughout this document, each of our statistical methods comprises a broader framework of general ideas and objectives that are widely applicable across a number of diverse domains, including demographic studies and business analytics.

Chapter 2 introduced the idea of principal differences analysis for interpretable characterization of differences between distributions, as well as efficient optimization algorithms for the corresponding estimation problem and its semidefinite relaxation. The PDA approach demonstrated numerous empirical benefits in tasks ranging from feature-selection, high-dimensional two-sample testing, and identifying differential gene-gene interactions between cell populations. Although we focused on algorithms for PDA & SPARDA tailored to the Wasserstein distance, which we favor for its interpretability when the profiled variables are measured on a meaningful scale, different statistical divergences may be better suited in other applications.

PDA is a useful method for contrasting two populations, but other applications (such as developmental scRNA-seq analysis) call for characterizing changes in an evolving population. 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 proposed TF algorithm resembles quantile-modeling techniques, our ideas are grounded under the unifying lens of the Wasserstein distance,

which we use to measure effects (3.8), goodness-of-fit (3.7), and a distribution-based least-squares fit (3.6). 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. 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. As simultaneously-profiled cell numbers grow to the many-thousands thanks to technological advances (Macosko et al. 2015), significant developmental discoveries may be made by studying the evolution of population-wide expression distributions, and TRENDS provides a principled framework for this analysis.

Chapter 4 focused on identifying how to best shape populations via external intervention rather than merely improving our understanding of their underlying characteristics. We proposed a Bayesian framework for directly learning beneficial transformations from observational data. While this objective is, strictly speaking, only possible under stringent causal assumptions, our approach performs well in both intentionally-misspecified and complex real-world settings. As supervised learning algorithms grow ever more popular, we expect intervention-decisions in many domains will increasingly rely on predictive models. Here, we introduced a "first do no harm" philosophy which formalizes the role of uncertainty within our definition of the optimal action. Our conservative definition provides a principled approach to handle the inherent uncertainty in these settings due to finite data. Able to employ any Bayesian regressor, the ideas presented in this chapter are widely applicable, considering easily-implemented forms of transformations that can either be personally tailored or enacted uniformly over a population.

5.1 Future work

Our work has opened up a number of interesting lines of future research. First, further theoretical investigation of the SPARDA framework is of interest, particularly in the high-dimensional $d = O(n)$ setting. Here, rich theory has been derived for compressed sensing and sparse PCA by leveraging ideas such as restricted isometry or spiked covariance (Amini & Wainwright 2009). A natural question is then which analogous properties of P_X, P_Y theoretically guarantee the strong empirical perfor-

mance of SPARDA observed in our high-dimensional applications. Finally, we also envision extensions of the methods presented here which employ multiple projections in succession, or adapt the approach to non-pairwise comparison of multiple populations.

The TRENDS framework presented in Chapter 3 introduces many theoretical questions, including further examination of the interplay between convergence rates and types of distributions, noise, and quantile-estimators. While our trend definition produces good empirical results in these scRNA-seq analyses (and encompasses various conceptually interesting effects discussed in §3.3.1), 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 5 provides an alternative definition of a trend which also applies to multidimensional distributions, and thus may be useful for applications such as spatiotemporal modeling.

In Chapter 4, we presented methods for identifying interventions which can be tailored on an individual basis, or globally enacted across an entire population. However, there exist numerous applications (particularly in the field of medicine), in which one wishes to identify distinct subpopulations, where all individuals within a subgroup uniformly receive the same intervention. The optimal strategy for such stratification of a heterogeneous population remains an open question, which might be addressed through clustering techniques or a tree-based partitioning of the covariate-space. Chapter 4 also relied on the strong assumption that any desired transformation within the feasible set can be precisely realized via external intervention, which is not the case in many practical settings. While our approach is still able to achieve strong empirical performance even in settings where this condition is violated, it remains an important task to explicitly address the issue of imprecisely enacted transformations. A key question is how to the propagate uncertainty in the post-intervention change in covariates into our objective function which defines the optimal intervention. Here, one might seek wide and flat optima of our intervention-objectives, such that even if the realized noisy transformation does not precisely achieve the optimal objective value, it nonetheless should produce a substantial outcome improvement (with high confidence).

Finally, we point out that Chapter 4 solely considered transformations of vector-valued data which can be represented in a standard tabular format, where each column represents a meaningful covariate and each row contains the measurements of these different variables observed within a single individual from the population. It remains of interest to extend this methodology to settings where, rather than being interpretably featurized, the data only consist of structured objects. For example, individual molecules in chemical applications may be represented as graphs, and sen-

tences in natural language processing are often viewed as a discrete sequences. When faced with a population consisting of structured objects, it becomes less obvious how to best identify maximally beneficial transformations. For the case of sequence data, we have introduced one way to apply the same type of gradient-based optimization utilized in Chapter 4, which involves leveraging a latent variable generative model of the structures (Mueller et al. 2017). Generalizing these ideas to encompass a broader class of structures remains an open avenue for future research.

Bibliography

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L. & Schapire, R. (2013), 'Taming the monster: A fast and simple algorithm for contextual bandits', *International Conference on Machine Learning* .

Altman, N. & Leger, C. (1995), 'Bandwidth selection for kernel distribution function estimation', *Journal of Statistical Planning and Inference* **46**, 195–214.

Amini, A. A. & Wainwright, M. J. (2009), 'High-dimensional analysis of semidefinite relaxations for sparse principal components', *The Annals of Statistics* **37**(5B), 2877–2921.

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.

Bach, F., Jenatton, R., Mairal, J. & Obozinski, G. (2012), 'Optimization with sparsity-inducing penalties', *Foundations and Trends in Machine Learning* **4**(1), 1–106.

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.

Beck, A. & Teboulle, M. (2009), 'A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems', *SIAM Journal on Imaging Sciences* **2**(1), 183–202.

Bertsekas, D. (1995), *Nonlinear Programming*, Athena Scientific.

Bertsekas, D. P. (1998), *Network Optimization: Continuous and Discrete Models*, Athena Scientific.

Bertsekas, D. P. (2011), Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey, *in* S. Sra, S. Nowozin & S. J. Wright, eds, 'Optimization for Machine Learning', MIT Press, pp. 85–119.

Bertsekas, D. P. & Eckstein, J. (1988), 'Dual coordinate step methods for linear network flow problems', *Mathematical Programming* **42**, 203–243.

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.

Bormuth, I., Yan, K., Yonemasu, T., Gummert, M., Zhang, M., Wichert, S., Grishina, O., Pieper, A., Zhang, W., Goebbels, S., Tarabykin, V., Nave, K.-A. & Schwab, M. H. (2013), 'Neuronal basic helix-loop-helix proteins Neurod2/6 regulate cortical commissure formation before midline interactions', *Journal of Neuroscience* **33**(2), 641–651.

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.

Bradley, P. S. & Mangasarian, O. L. (1998), 'Feature selection via concave minimization and support vector machines', *International Conference on Machine Learning* .

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N. & Scott, S. L. (2015), 'Inferring causal impact using bayesian structural time-series models', *Annals of Applied Statistics* **9**, 247–274.

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', *Nature Biotechnology* **33**(2), 155–60.

Clemmensen, L., Hastie, T., Witten, D. & Ersbøll, B. (2011), 'Sparse Discriminant Analysis', *Technometrics* **53**(4), 406–413.

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', New York, pp. 185–212.

Cramer, H. & Wold, H. (1936), 'Some Theorems on Distribution Functions', *Journal of the London Mathematical Society* **11**(4), 290–294.

Cuesta-Albertos, J. A., Fraiman, R. & Ransford, T. (2007), 'A sharp form of the Cramer–Wold theorem', *Journal of Theoretical Probability* **20**(2), 201–209.

Cuturi, M. (2013), 'Sinkhorn distances: Lightspeed computation of optimal transport', *Advances in Neural Information Processing Systems* **26**.

Daminaou, A. & Lawrence, A. (2013), 'Deep Gaussian processes', *International Conference on Artificial Intelligence and Statistics* .

D'Aspremont, A., El Ghaoui, L., Jordan, M. I. & Lanckriet, G. R. (2007), 'A direct formulation for sparse PCA using semidefinite programming', *SIAM Review* pp. 434–448.

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.

Dobrushin, R. (1970), 'Definition of a system of random variables by conditional distributions', *Teor. Verojatnost. i Primenen* **15**, 469–497.

Duchi, J., Hazan, E. & Singer, Y. (2011), 'Adaptive Subgradient Methods for Online Learning and Stochastic Optimization', *Journal of Machine Learning Research* **12**, 2121–2159.

Duvenaud, D., Eaton, D., Murphy, K. & Schmidt, M. (2010), 'Causal learning without DAGs', *JMLR: Workshop and Conference Proceedings* **6**, 177–190.

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', *Nature Cell Biology* **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.

Fernandes, K., Vinagre, P. & Cortez, P. (2015), 'A proactive intelligent decision support system for predicting the popularity of online news', *EPIA Portuguese Conference on Artificial Intelligence* .

Gallopin, T., Geoffroy, H., Rossier, J. & Lambolez, B. (2006), 'Cortical sources of CRF, NKB, and CCK and their effects on pyramidal cells in the neocortex.', *Cerebral Cortex* **16**(10).

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.

Gibbs, A. L. & Su, F. E. (2002), 'On Choosing and Bounding Probability Metrics', *International Statistical Review* **70**(3), 419–435.

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

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

Goodfellow, S. J., Rebello, M. R., Toska, E., Zeef, L. A. H., Rudd, S. G., Medler, K. F. & Roberts, S. G. E. (2011), 'WT1 and its transcriptional cofactor BASP1 redirect the differentiation pathway of an established blood cell line', *Biochemical Journal* **435**, 113–125.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Scholkopf, B. & Smola, A. (2012), 'A Kernel Two-Sample Test', *The Journal of Machine Learning Research* **13**, 723–773.

Guyon, I., Gunn, S., Nikravesh, M. & Zadeh, L. A. (2006), *Feature Extraction: Foundations and Applications*, Springer-Verlag, Secaucus, NJ, USA.

Hagemann, T. L., Jobe, E. M. & Messing, A. (2012), 'Genetic Ablation of Nrf2/Antioxidant Response Pathway in Alexander Disease Mice Reduces Hippocampal Gliosis but Does Not Impact Survival', *PLoS ONE* **7**(5), e37304.

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.

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.

Haque, A., Engel, J., Teichmann, S. A. & Lonnberg, T. (2017), 'A practical guide to single-cell rna-sequencing for biomedical research and clinical applications', *Genome Medicine* **9**, 75.

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

Hill, J. L. (2011), 'Bayesian nonparametric modeling for causal inference', *Journal of Computational and Graphical Statistics* **20**(1), 217–240.

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

Jara, J. H., Genç, B., Cox, G. A., Bohn, M. C., Roos, R. P., Macklis, J. D., UlupÄśnar, E. & Hande Özdinler, P. (2015), 'Corticospinal Motor Neurons Are Susceptible to Increased ER Stress and Display Profound Degeneration in the Absence of UCHL1 Function', *Cerebral Cortex* .

Jasnow, A. M., Ressler, K. J., Hammack, S. E., Chhatwal, J. P. & Rainnie, D. G. (2009), 'Distinct subtypes of cholecystokinin (CCK)-containing interneurons of the basolateral amygdala identified using a CCK promoter-specific lentivirus.', *Journal of Neurophysiology* **101**(3), 1494–1506.

Jirak, M. (2011), 'On the maximum of covariance estimators', *Journal of Multivariate Analysis* **102**, 1032–1046.

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.

Kemmeren, P., Sameith, K., van de Pasch, L. A., Benschop, J. J., Lenstra, T. L., Margaritis, T., O'Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C. W. et al. (2014), 'Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors', *Cell* **157**(3), 740–752.

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

Kleinberg, J., Ludwig, J., Mullainathan, S. & Obermeyer, Z. (2015), 'Prediction policy problems', *American Economic Review: Papers & Proceedings* **105**(5), 491–495.

Kraft, D. (1988), *A software package for sequential quadratic programming*, DLR German Aerospace Center - Institute for Flight Mechanics, Koln, Germany.

Krishnan, R. G., Shalit, U. & Sontag, D. (2015), 'Deep kalman filters', *Advances in Neural Information Processing Systems* **28**.

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).

Le, Q. V., Smola, A. J. & Canu, S. (2005), 'Heteroscedastic Gaussian process regression', *International Conference on Machine Learning* .

Lederer, C. W., Torrisi, A., Pantelidou, M., Santama, N. & Cavallaro, S. (2007), 'Pathways and genes differentially expressed in the motor cortex of patients with sporadic amyotrophic lateral sclerosis.', *BMC Genomics* **8**, 26.

Lesage, S. & Brice, A. (2009), 'Parkinson's disease: from monogenic forms to genetic susceptibility factors', *Human Molecular Genetics* **18**(R1), R48–R59.

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

Li, X., Yu, K., Zhang, Z., Sun, W., Yang, Z., Feng, J., Chen, X., Liu, C.-H., Wang, H., Guo, Y. P. & He, J. (2014), 'Cholecystokinin from the entorhinal cortex enables neural plasticity in the auditory cortex', *Cell Research* **24**(3), 307–330.

Linnertz, C., Lutz, M. W., Ervin, J. F., Allen, J., Miller, N. R., Welsh-Bohmer, K. A., Roses, A. D. & Chiba-Falek, O. (2014), 'The genetic contributions of SNCA and LRRK2 genes to Lewy Body pathology in Alzheimer's disease.', *Human Molecular Genetics* **23**(18), 4814–4821.

Lizotte, D. J. (2008), Practical Bayesian Optimization, PhD thesis, University of Alberta.

Lopes, M., Jacob, L. & Wainwright, M. (2011), 'A More Powerful Two-Sample Test in High Dimensions using Random Projection', *Advances in Neural Information Processing Systems* pp. 1206–1214.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A. & McCarroll, S. A. (2015), 'Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets', *Cell* **161**(5), 1202–1214.

Mallows, C. (1972), 'A note on asympototic joint normality', *Annals of Mathematical Statistics* pp. 508–515.

McHutchon, A. & Rasmussen, C. E. (2011), 'Gaussian process training with input noise', *Advances in Neural Information Processing Systems* **24**.

Mobahi, H., L, Z. C. & Ma, Y. (2012), 'Seeing through the blur', *IEEE Conference on Computer Vision and Pattern Recognition* .

Molyneaux, B. J., Arlotta, P., Menezes, J. R. L. & Macklis, J. D. (2007), 'Neuronal subtype specification in the cerebral cortex', *Nature Reviews Neuroscience* **8**(6), 427–437.

Mueller, J., Gifford, D. & Jaakkola, T. (2017), 'Sequence to better sequence: Continuous revision of combinatorial structures', *International Conference on Machine Learning* .

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.

Oh, D. H., Park, Y. C. & Kim, S. H. (2010), 'Increased glycogen synthase kinase-3beta mRNA level in the hippocampus of patients with major depression: a study using the stanley neuropathology consortium integrative database.', *Psychiatry Investigation* **7**(3).

Paciorek, C. J. & Schervish, M. J. (2004), 'Nonstationary covariance functions for Gaussian process regression', *Advances in Neural Information Processing Systems* **17**.

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.

Pearl, J. (2000), *Causality: Models, Reasoning and Inference*, Cambridge Univ. Press.

Pele, O. & Werman, M. (2009), 'Fast and robust earth mover's distances', *IEEE International Conference on Computer Vision* .

Peters, J., Bühlmann, P. & Meinshausen, N. (2016), 'Causal inference using invariant prediction: identification and confidence intervals', *Journal of the Royal Statistical Society: Series B* **78**, 1–42.

Peters, J., Mooij, J. M., Janzing, D. & Schölkopf, B. (2014), 'Causal discovery with continuous additive noise models', *Journal of Machine Learning Research* **15**, 2009–2053.

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**.

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.', *Journal of Cell Biology* **151**(6), 1295–1304.

Rasmussen, C. E. (2006), *Gaussian processes for machine learning*, MIT Press.

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.

Ritchie, M., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. & Smyth, G. K. (2015), 'limma powers differential expression analyses for RNA-sequencing and microarray studies', *Nucleic Acids Research* **43**(7), e47.

Rojas-Carulla, M., Schölkopf, B., Turner, R. & Peters, J. (2016), 'Causal transfer in machine learning', *arXiv:1507.05333* .

Rosenbaum, P. R. (2005), 'An exact distribution-free test comparing two multivariate distributions based on adjacency', *Journal of the Royal Statistical Society, Series B* **67**, 515–530.

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.

Sandler, R. & Lindenbaum, M. (2011), 'Nonnegative Matrix Factorization with Earth Mover's Distance Metric for Image Analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(8), 1590–1602.

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.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. (2016), 'Taking the human out of the loop: A review of Bayesian optimization', *Proceedings of the IEEE* **104**(1), 148–175.

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

Shimizu, S., Hoyer, P., Hyvärinen, A. & Kerminen, A. J. (2006), 'A linear non-Gaussian acyclic model for causal discovery', *Journal of Machine Learning Research* **7**, 2003–2030.

Shiota, J., Ishikawa, M., Sakagami, H., Tsuda, M., Baraban, J. M. & Tabuchi, A. (2006), 'Developmental expression of the SRF co-activator MAL in brain: role in regulating dendritic morphology', *Journal of Neurochemistry* **98**, 1778–1788.

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

Srinivas, N., Krause, A., Kakade, S. M. & Seeger, M. (2010), 'Gaussian process optimization in the bandit setting: No regret and experimental design', *International Conference on Machine Learning* .

Szekely, G. & Rizzo, M. (2004), 'Testing for equal distributions in high dimension', *InterStat* **5**.

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

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society, Series B* pp. 267–288.

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', *Nature Biotechnology* **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 and Data Analysis* **51**(8), 3885–3897.

van der Vaart, A. & van Zanten, H. (2011), 'Information rates of nonparametric Gaussian process methods', *Journal of Machine Learning Research* **12**, 2095–2119.

van der Vaart, A. W. & Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer.

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

Villani, C. (2008), *Optimal transport, old and new*, Springer.

Wang, Z., Lu, H. & Liu, H. (2014), 'Tighten after Relax: Minimax-Optimal Sparse PCA in Polynomial Time', *Advances in Neural Information Processing Systems* **27**, 3383–3391.

Wei, S., Lee, C., Wichers, L. & Marron, J. S. (2015), 'Direction-Projection-Permutation for High Dimensional Hypothesis Tests', *Journal of Computational and Graphical Statistics* .

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

Wright, S. J. (2010), 'Optimization Algorithms in Machine Learning', *NIPS Tutorial* .

Wu, J. Q., Wang, X., Beveridge, N. J., Tooney, P. A., Scott, R. J., Carr, V. J. & Cairns, M. J. (2012), 'Transcriptome Sequencing Revealed Significant Alteration of Cortical Promoter Usage and Splicing in Schizophrenia', *PLoS ONE* **7**(4), e36351.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M. et al. (2015), 'Personalized nutrition by prediction of glycemic responses', *Cell* **163**(5), 1079–1094.

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.

Zhang, L., Ju, X., Cheng, Y., Guo, X. & Wen, T. (2011), 'Identifying Tmem59 related gene regulatory network of mouse neural stem cell from a compendium of expression profiles', *BMC Systems Biology* **5**(1), 152.

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

Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A. M. & Mazutis, L. (2017), 'Single-cell barcoding and sequencing using droplet microfluidics', *Nature Protocols* **12**, 44–73.

Zou, H., Hastie, T. & Tibshirani, R. (2005), 'Sparse Principal Component Analysis', *Journal of Computational and Graphical Statistics* **67**(2), 301–320.