Probabilistic Models and Statistical Tools for Gene Expression Analysis

by

Đan Daniel Erdmann-Pham

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Applied Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yun S. Song, Co-chair
Professor Steven N. Evans, Co-chair
Professor Rasmus Nielsen

Summer 2021

Probabilistic Models and Statistical Tools for Gene Expression Analysis

Abstract

Probabilistic Models and Statistical Tools for Gene Expression Analysis

by

Đan Daniel Erdmann-Pham

Doctor of Philosophy in Applied Mathematics

University of California, Berkeley

Professor Yun S. Song, Co-chair

Professor Steven N. Evans, Co-chair

The recent explosion of novel experimental protocols in the life sciences is providing glimpses into fundamental biological processes that previously remained inaccessible. The mechanisms underlying these processes are unique enough that understanding them often requires approaches beyond those traditionally associated with data-driven statistics. This thesis explores three such instances, in which borrowing ideas from geometry, probability theory and partial differential equations can lead to tangible improvements over existing methods, as well as new frameworks for future tasks. Our first analysis revolves around protein synthesis: the conversion of genes into viable polypeptides through what is known as translation. By exploiting the so-called Totally Asymmetric Simple Exclusion Process (TASEP) as a model of translation, we rephrase questions of biological interest in terms of Markov chains properties, which in turn we successfully tackle by deriving an adequate continuum limit of the TASEP. Analysis of this limiting process reveals a handful of key parameters that govern translation efficiency, whose roles we summarize in a concise set of design principles, and confirm on ribosome profiling data of yeast. Secondly, we direct attention to the task of gene expression deconvolution: recovering individual cell type contributions to the transcript abundances of an entire tissue. By embedding our deconvolution procedure into a full-likelihood framework, we not only provide provably optimal error guarantees, but also enable convenient model evaluation, adaptation and uncertainty quantification. We demonstrate this improved performance and flexibility on a variety of simulated and experimental bulk samples. And thirdly, motivated by detecting differential expression of genes across tissues, individuals or conditions, we investigate non-parametric two-sample testing. After identifying a broad family of statistics that includes as special cases Mann-Whitney's U, Greenwood's and Dixon's, we employ combinatorial tools to quickly compute their null distributions' moments to arbitrary precision. Combined with an equally fast and provably accurate solution to the related moment problem we thus arrive at a well-calibrated, versatile goodness-of-fit test with applicability beyond the gene expression setting. We showcase its power in various direct comparisons with a number of tests commonly used in practice.

“The journey’s only just begun”, said the tortoise to the meerkat and lifted a leg to air out its shell. *“Adventures await.”*

Contents

List of Figures

List of Tables

Acknowledgments

Graduate school is a journey filled with ups, downs and long flats. That the ups prevailed ultimately is in large measure due to the phenomenally kind and supportive people who have pushed, prodded and reversed me in the right direction. To them I wish to express my profound gratitude.

It is Yun's curiosity and passion that motivated my first excursions into mathematical biology; and his generous advisorship and continual encouragement that helped me embrace it as a home for my research. Yun deeply cares about his advisees' success in all matters academic, professional and personal and frequently lends more hands, ears and time to them than I imagined any single human could have. I truly am fortunate to have come under his aegis. If Yun showed me exciting research paths to tread on, then it is Khanh, "the ribosome-whisperer", who early on helped ensure that I walked them without stumbling or veering off-track. As my first collaborator in Berkeley, he ever so patiently and kindly showed me the academic ropes, while offering—and continuing to do so—invaluable guidance and perspective in and outside the office, which I much cherish. Chapter 2 was written under his counsel. Similarly, I greatly appreciate what my co-authors "single-cell" Jonathan[1] and "tree-topology" Jonathan have taught me both within our collaborations and beyond, much of which is assembled in Chapters 3 and 4. Indeed, a substantial part of the pleasure of working interdisciplinarily springs from the exposure to an inexhaustibly diverse palette of individuals and the scientific ideas that drive their academic zeal. I thank Joe DeRisi, Valentina Garcia, Dirk Hockemeyer, Greg Huber, and Daniel Rokhsar for providing inspiration and sharing their perspective on various joint projects; although not included in this thesis, our collaborations significantly helped shape me into the scientist I am now. Jeff Calder, Craig Evans, Steve Evans, Jim Pitman, and Fraydoun Rezakhanlou, who through their friendliness turned my initial hesitation of approaching professors into excitement, equipped me with the mathematical tools that ultimately tie together every one of my research strands. Their clarity in teaching and conversation helped me build the mathematical lens through which I view most scientific questions today, and for which I thank them sincerely. Additionally, Steve's kind assistance in more pragmatic matters outside of mathematics is warmly appreciated; as is that of Barb, Vicky and Isabel, who in ways akin to magic airlifted me out of bureaucratic jungles that I had lost myself in on more than one occasion—your work makes our work that much easier. Last but certainly not least in this list of mentors and academic guides, I would like to thank Alistair Sinclair, whose thoughtful approach to teaching is inspiring, instructive and modern all at once; serving as a teaching assistant for his and Yun's class broadened my teaching toolkit immeasurably.

Life in Berkeley outside the ups, downs and flats of research can be hilly too, and I am glad to have had fantastic companions to traverse these hills with me. Sanjit and I began as lab-mates three years ago and over time turned into workout-mates, ping-pong mates, housemates and finally just mates[2]. His hearty laughter flattens any hill, disperses every worry, and even resolves much philosophical dispute; and I am deeply grateful for the countless times he has shared it with me. Milind H is one of the few individuals with whom I feel at ease talking sillily of the solemn and

[1] Any added nicknames are descriptors of expertise rather than bodily composition.

[2] from the Cambridge Dictionary: mate (*noun, chiefly British*)—friend or companion

solemnly of the silly; productively. Our midnightly lunches are a source of inspiration and fiber, and I am looking forward to their continuation across different time zones. Ben has an outstandingly colourful character, and getting glimpses of his palette of ideas, interactions and life experiences substantially enriched my own time and again. I will treasure what I learned from him both mathematically and extra-mathematically. My fellow I-Housees Crispin, Ezra, Jonathan, Makoto, Nina, and Uyen made the foreign familiar, the familiar fantastic, and fire alarms fun. Thank you for all the stories—both past and future. Jennifer's encouragement means more to me than she is willing to admit, Linnan's whirlwind energy is a sustainable power source in all my endeavours, Alexandra's Falafel-Philosophising provides delectable nourishment for body, mind and heart alike, and Jane's fusion of honesty, playfulness and thoughtfulness is something everyone needs in their lives; my heartfelt gratitude goes out to you for entwining your journeys with mine. Satyaki is an infinite-dimensional friend projected onto the three we live in; witnessing him draw energy from the dimensions hidden has been invigorating and refreshing. Nick is fully harnessing everything our three dimensions have to offer and his balancedness despite overwhelming breadth of experience is inspiring. His many wise words on any situation academic or non-academic (or, indeed, neither) have helped me climb daunting hills, which I am greatly thankful for. Climbing the hill to my office in Stanley Hall would not have been as enjoyable without the eclectic group of lab-mates not mentioned so far: "Squash"-Jeff, "Frisbee"-Jeff, "Frisbee-and-probably-also-Squash"-Neil, "More-German-than-me"-Yutong, "polytonal"-Alan, and Antoine, Gonzalo, Miaoyan, Milind, Sebastian, Yun and Zvi. Their creativity and jolliness both in and outside of work provided the ideal focus on, as well as distraction from, research; thank you for the excellent company. To Yining, Yuxi, Xurui, John, Wonjun, Justin and Frank I was in some form or another assigned as mentor or instructor, and I can only hope that I was able to teach them nearly as much as they taught me. Their enthusiasm for mathematics or statistics contributed to fueling my own, and I am excited to see where it will lead them. And finally, though most of this thesis was composed in Berkeley, part of the reason I remained composed throughout are the wonderful, distance-defying connections back home in Germany and across the world. My heartfelt appreciations go out to Anca for being full of surprises, Jakob for maturing me into the person and ping pong player I am now, Mai-Ly for dancing through life even when there is no music, Xuxu for always seeing both the trees and the forest (and often the flora and fauna hiding between them too), Hathu for her love of reading, listening and watching, but not baking, Thomas for teaching me maths without ever calling it such, and Andreas for his insatiable thirst for knowledge and hunger for my mom's food. What our interactions lack in frequency, they add in consistency, duration and meaning.

Of course last, and not only not least, but most profoundly, I would like to thank my parents. While physically in Germany, they are on my mind wherever I am, and provide a constant supply of energy, happiness, insights, advice, laughs, clothes, surprises, stories, love, and all other positive emotions. Mama, Papa, this thesis is in many ways dedicated to you (but please don't feel obliged to read through the entire thing; it's a bit lengthy and sort of technical at times—I'll give you a more digestible account of it over dinner soon).

Chapter 1

Introduction

The content of this thesis lies broadly in the field of mathematical biology: It attempts to understand biological processes through the lens of probability theory, geometry and mathematical analysis. More specifically, we will be interested in modeling quantitatively the processes that govern what is called the gene expression ecosystem on all its scales—from the large-scale descriptions of DNA-to-RNA transcription and genome evolution to the molecular dynamics of protein synthesis. Insights into these processes are primarily afforded through data, and consequently much of our energy will be focused on the development of rigorous and transparent data analysis schemes that tie the observables we measure to the mechanisms that generate them. Despite this technical and quantitative emphasis of our work, we hope that ultimately it serves to aid illuminate very real biological phenomena whose inner workings have remained opaque. Thus, we open this introductory chapter with a motivational prologue outlining the kind of biological questions that will guide our theoretical considerations throughout the thesis.

1.1 Background

Figure 1.1 provides a schematic representation of what is known as the *central dogma of molecular biology*, or perhaps less intimidatingly, the directed flow of genetic information that occurs in all[1] living cells, and is essential for any organism's proper functioning: Starting out as geometrically rigid, highly localized double-helix structures on the left-hand side of Figure 1.1, DNA molecules are converted into the considerably more nimble and flexible RNA molecules depicted in the middle panel through a process known as transcription. These messenger molecules in turn are mobile enough to freely traverse the cell, reaching even its remotest corners, whereupon they themselves, by way of the so-called translation step, are turned into protein (illustrated on the far right); the geometric machines whose folding and configuration allow performance of the various processes necessary to keep the cell alive. This conversion of rigid yet easily duplicated and maintained DNA to elaborate yet functional protein is mediated primarily by two classes of

[1] Throughout this thesis the word *all* is taken to mean *all, with possibly very few exceptions* when encountered in a biological context; when used in a mathematical setting, we truly mean *all*.

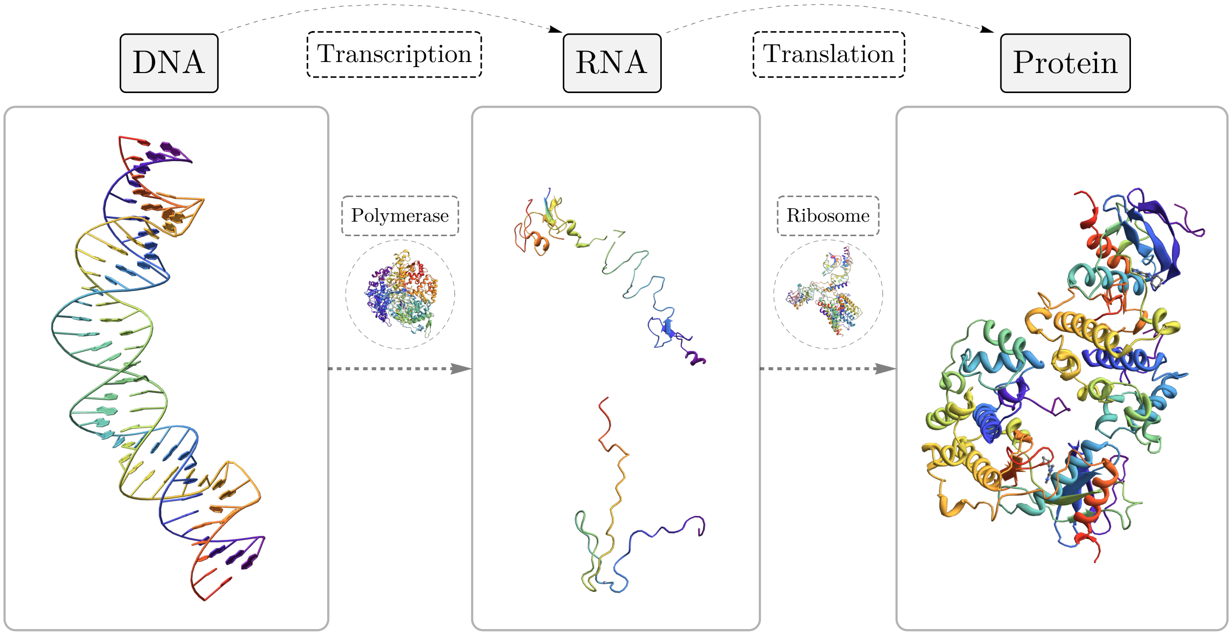


Figure 1.1: **The central dogma of molecular biology.** The directed flow of genetic information describes the conversion of rigid DNA molecules (left) into functional protein (right) via RNA molecules (middle). This conversion is mediated by an abundance of particle-like objects (named RNA polymerase and ribosomes) that traverse and travel between the various involved structures in a highly stochastic manner.

small, particle-like[2] organelles: RNA polymerases and ribosomes. These molecules perpetually meander through the cell, attaching to, detaching from and traversing the various DNA and RNA structures involved while working to transform them into their respective molecular successors. Being subject to and conditional on the various diffusive, thermal and chemical fluctuations that shape the cellular landscape, their motion is to all intents and purposes random in nature. As a result, the aforedescribed, macroscopic gene expression ecosystem is, on a molecular scale, driven by a stochastic interacting particle process, which naturally renders its (macroscopic) observables of interest stochastic as well. Attempting to capture and make sense of this complexity is what turns Figure 1.1 into an inviting playground for probabilistic and statistical modeling stories to grow up in—of which this thesis is presenting three.

[2] at least macroscopically; on a microscopic level, polymerases and ribosomes exhibit delicate geometric structure

1.2 Outline

A theoretical model of translation

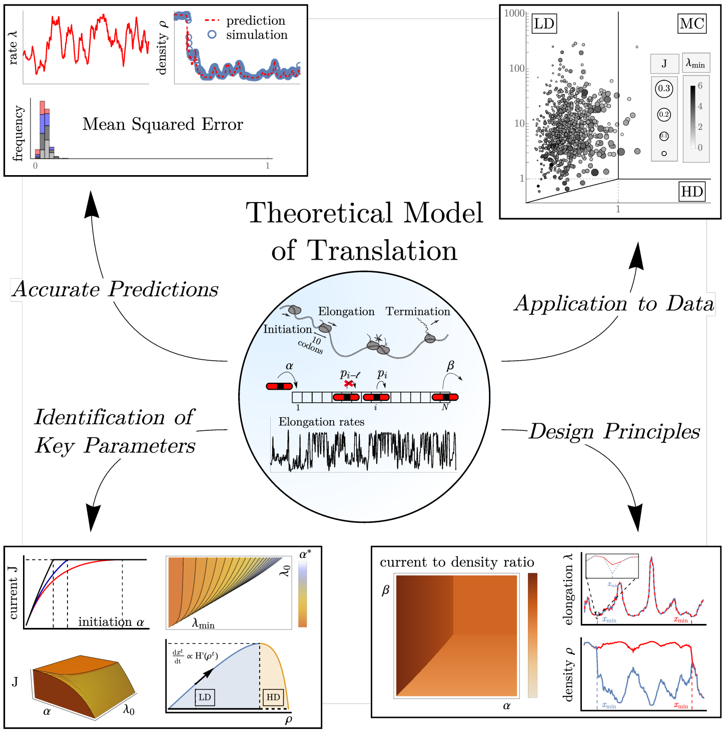


Figure 1.2: **Modeling translation dynamics.**

We begin by zooming into the rightmost sandbox of this playground in Chapter 2: The conversion of RNA into protein by way of translation. As one of the final major steps of synthesizing viable protein, translation has received intense scrutiny both as a target for interrupting pathogenic pathways (see, e.g., [W+03] for a comprehensive overview of translation as a target for antibiotics, and [BRH+15] for more recent developments) and as a tool for amplifying the efficacy of biotechnological and pharmaceutical applications (with the most recent mRNA-vaccines providing a case in point). Harnessing it most fruitfully requires pinpointing the key players and key determinants that govern translation efficiency; the knobs that need turning in order to enhance or throttle protein production. We carry out such identification by resorting to a mathematical model known as the inhomogeneous ℓ-TASEP (**T**otally **A**symmetric **S**imple **E**xclusion **P**rocess with particles of size $\boldsymbol{\ell}$), which constitutes a Markov chain popular in both the theoretical and applied sciences. In the context of the former, it has established itself as a model system in non-equilibrium statistical mechanics, and recently garnered notable attention in the field of integrable probability and random surface growth models for its unexpected connections to random matrix theory and display of a nascent universality class called KPZ (see, e.g., [Cor16] for a survey). In more applied settings, it provides an interacting particle process at the heart of various transport phenomena arising in traffic flow, molecular signaling and sedimentation among others (see, e.g., [SCN10] for an extensive overview). Despite its ubiquity, characterizing key quantities of the general inhomogeneous ℓ-TASEP like its stationary distribution or particle current has largely remained elusive, with quantitative and qualitative insights supplied mostly through simulation and approximation only. In light of recent advances in the field of experimental ribosome profiling, which affords unprecedented snapshots of translation in vivo, the intractability and inaccuracy that come with simulations and approximations are infeasible for faithful parameter inference from such data. Chapter 2 is devoted to overcoming these limitations by deriving *exact* expressions of various key quantities of interest in a continuum regime called the hydrodynamic limit, employing them to formulate a concise set of design principles that guarantee optimal translation efficiencies, and applying them to data to confirm their utility in practical biological settings. Figure 1.2 displays a condensed graphical abstract.

Full-likelihood deconvolution

The translation process described above requires as essential input RNA transcripts, and is thus fundamentally constrained by total transcript abundances in the cell. These abundances almost always exceed the two copies depicted in Figure 1.1, typically ranging in the thousands or hundreds of thousands instead, and are chiefly determined by the stochastic interplay between polymerases and DNA molecules, and therefore random variables themselves. Since their magnitude directly affects translational activity and thereby protein concentrations, it is not surprising that their distributional properties ought to depend on the function and environment of the cell they reside in; e.g., genes related to pigmentation are likely expressed at larger levels in cells of the iris than in the liver, while transcripts encoding various metabolic proteins are expected in more considerable quantities in the latter; with both fluctuating significantly throughout the various stages of development. This observation is the primary inspiration for the contents of Chapters 3 and 4, which interrogate its inverse: Can we distinguish cells of distinct cell types or in distinct environments based solely on their RNA-count profiles? And if we can, is it possible to tell the cell type composition of entire tissues or organs from observing (samples from) the total transcript tally aggregated across all its constituent cells? Despite its seemingly simpler structure, the first question calls for a statistically rather more complex answer and so is postponed to Chapter 4. The follow-up question is known as the problem of cell-type deconvolution[3] and is regularly asked in clinical settings: While practitioners are unlikely to encounter organs that are half-iris half-liver, it is common for tissues' cell type compositions to drift away from their (frequently difficult to assess) baseline throughout different stages of development or disease progression, and their analysis thus promises diagnostic potential. Chapter 3 is devoted to furnishing tools for precisely this analysis: By exploiting reference panels of known cell type profiles and positing explicit generative models for both their and the target tissues' associated noise, we are able to construct an inference scheme that (under mild conditions) supplies asymptotically optimal proportion estimates combined with provably well-calibrated confidence regions. We demonstrate our technique's favorable performance by benchmarking against a zoo of previously proposed deconvolution methods, and showcase the flexibility and versatility that result from our likelihood formulation through deconvolving and interpreting a sequence of real biological datasets collected to examine the impact of aging on tissue composition.

[3] not to be confused with the inverse operation of convolution, which in mathematics literature is typically referred to as deconvolution too; see, e.g., [Ria86]

Non-parametric hypothesis testing

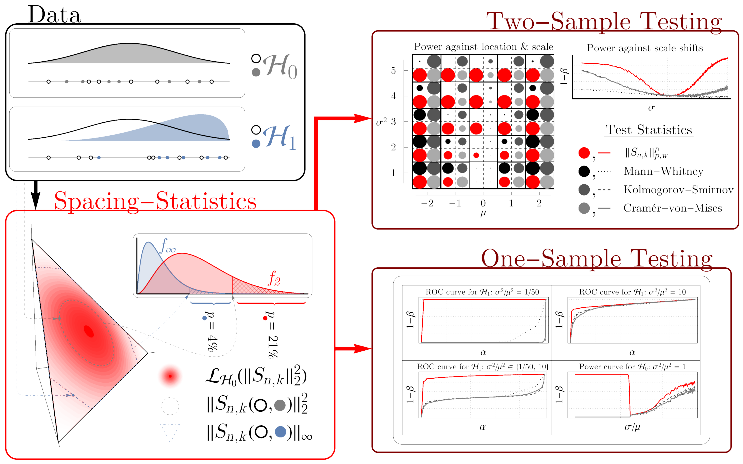

Data
$\mathcal{H}_0$
$\mathcal{H}_1$
Spacing–Statistics
f_∞
f_2
$p = 4\%$
$p = 21\%$
$\mathcal{L}_{\mathcal{H}_0}(\|S_{n,k}\|_2^2)$
$\|S_{n,k}(\bigcirc, ●)\|_2^2$
$\|S_{n,k}(\bigcirc, ●)\|_\infty$
Two–Sample Testing
Power against location & scale
σ^2
5
4
3
2
1
−2
−1
0
1
2
μ
Power against scale shifts
$1-\beta$
σ
Test Statistics
$\|S_{n,k}\|_{p,w}^p$
Mann–Whitney
Kolmogorov–Smirnov
Cramér–von–Mises
One–Sample Testing
ROC curve for $\mathcal{H}_1$: $\sigma^2/\mu^2 = 1/50$
ROC curve for $\mathcal{H}_1$: $\sigma^2/\mu^2 = 10$
ROC curve for $\mathcal{H}_1$: $\sigma^2/\mu^2 \in \{1/50, 10\}$
Power curve for $\mathcal{H}_0$: $\sigma^2/\mu^2 = 1$
$1-\beta$
α
σ/μ

Figure 1.3: **Hypothesis tests via spacings.**

The deconvolution method just described requires access to RNA-count distributions of individual cell types or proxies thereof; indeed, it is straightforward to convince oneself that absent such access faithful proportion inference is essentially impossible[4]. Interestingly, the aforementioned question of distinguishing cell types and conditions based on gene expression, rather than reconstructing their proportions in mixtures, is meaningful even without such reference profiles: It seems reasonable that two sets of transcript abundances can be declared distributionally distinct despite no explicit understanding of how precisely they are so. This is the task of so-called two-sample tests: given two collections of random variables $\{X_1,\dots,X_n\}$ and $\{Y_1,\dots,Y_m\}$ drawn *i.i.d.* from distributions μ and ν, respectively; is it possible to reliably decide whether $\mu = \nu$, or more broadly, if $\nu \in \mathscr{A}$ and $\mu \notin \mathscr{A}$ for some class of measures $\mathscr{A}$? In clinical settings, answering such question provides a dimension separate from, and more general than, cell type composition through which otherwise difficult to probe departures from normalcy can be captured, and has consequently spurred the development of a diverse array of statistical tools to aid its analysis. This wealth of methodology notwithstanding, satisfactory resolutions of this question have mostly been attained under comparatively narrow assumptions on $\mathscr{A}$: Procedures like Student's t-test or the Mann-Whitney test rest on rigid parametric or semi-parametric constraints (like normality of the involved distributions or a parametrization of $\mathscr{A}$ by locations shifts, i.e., $\mathscr{A} = \{\mu(\cdot - a), a \in \mathbb{R}\}$), while schemes akin to the Kolmogorov-Smirnov test or Cramér-von-Mises criterion trade power for generality (that is, $\mathscr{A} = \{\text{all } \nu : \nu \neq \mu\}$ is maximal, but $\nu\,\{\text{decide } \mu \neq \nu\}$ behaves poorly even for ν far away from μ). From a perspective of analyzing differential gene expression, neither such behavior is desirable—transcription levels may aberrate due to a variety of distinct biological conditions, rendering any absolute assumptions on $\mathscr{A}$ infeasible; while any sacrifice in power risks misdiagnosis of potential disease state. Chapter 4 attempts to demonstrate that this trade-off between generality and power is, in fact, surmountable: instead of relying on any individual test statistic, it devises an entire family $\mathscr{F}$ of summary statistics that is *(i)* compact enough to characterize null distributions explicitly, yet *(ii)* rich enough to contain at least one member of substantial power against any generic, fixed class of alternatives $\mathscr{A}$. By

[4] unless, of course, other strong assumptions are imposed; e.g., the availability of marker genes or a large number of adequately sampled bulk tissues can compensate for missing references; such approaches are aptly dubbed reference-free deconvolutions

designing an efficient algorithm to identify this particularly reliable member, it thus provides a general, non-parametric two-sample test capable of detecting a wide class of alternatives. As is often the case when analyzing hypothesis tests, the principal technical challenge lies in adequately describing the null distributions of $\mathscr{F}$, which our proposed method overcomes by means of deriving explicit formulae for the involved moment sequences coupled with a fast and provably accurate scheme to resolve the resultant moment problem. Conveniently, following such route not only provides compelling performance when applied to configurations typically encountered in the gene expression setting and compared with various test in common use (see, e.g., Figure 1.3 for a schematic impression), but also helps to illuminate one particular member of $\mathscr{F}$ which had sparked considerable theoretical work over the years prior: the so-called Greenwood statistic. By examining its moment sequence more closely, we are able to contribute towards addressing questions surrounding its tail behavior, regularity and monotonicity that had remained previously open, providing a pleasing example of how application-driven endeavors may yield results that are of interest to theorists as well. Indeed, on a less concrete level this dissertation hopes to convince the reader that instances of such symbiotic interplay between theory and application—with both profiting from each other—abound in nature and occasionally can enable insights that may not have been attained from studying either in isolation.

Chapter 2

Modeling Translation Dynamics

This chapter is joint work with Khanh Dao Duc and Yun S. Song, and was published as [EPDS20] in *Cell Systems*.

2.1 Introduction

Being a major determinant of gene expression and protein abundance levels [LVW+07, KGF13], translation of mRNA into polypeptides is one of the most fundamental biological processes underlying life. The extent to which this process is regulated and shaped by the sequence landscape has been widely studied over the past decades [DKP16, HC18, QCSvdO15], revealing many intricate mechanisms that may affect translation dynamics. From a more global perspective, however, it has been challenging to integrate these findings to elucidate the key factors that govern translation efficiency. Indeed, translation is a complex process that depends on many parameters, including the initiation rate, site-specific elongation rates (which can vary substantially along a given transcript), and the termination rate. How does the overall rate of protein synthesis depend on these parameters? To make the problem more concrete, suppose that the goal is to achieve the fastest rate of protein production while minimizing the cost. Would choosing the "fastest" synonymous codon at each site do the job? If the local elongation rate changes at a particular site, would it necessarily affect the overall rate of protein synthesis? If not, then which parameters actually matter? Aside from achieving a desired protein production rate, how does a translation system make efficient use of available resources, particularly the ribosomes? These are important questions in molecular and evolutionary biology, as well as synthetic biology, but challenging to answer because there are many parameters involved – for a transcript consisting of N codons, one has to analyze a model with about N parameters, which is seemingly intractable when N is large.

In this chapter, we develop a theoretical tool to answer the above questions. Our work hinges on analyzing a mathematical model that describes the traffic flow of ribosomes, which mediate translation by moving along the mRNA transcript. Beginning with [MGP68], most mechanistic studies of translation dynamics have been based on the so-called Totally Asymmetric Simple Exclusion Process (TASEP), a probabilistic model that explicitly describes the flow of particles along a lattice [ZDS11, ZT16]. As a classical model of transport phenomena in non-equilibrium, the TASEP has attracted wide interest from mathematicians and physicists [BE07]. To describe translation realistically, however, a generalized version of the model needs to be employed, taking into account the extended size of the ribosome and the heterogeneity of the elongation rate along the transcript. Under such general conditions, critical questions have hitherto remained open; in particular, identifying the parameters most crucial to the current and particle density has proven elusive.

Here we carry out a theoretical analysis of a generalized version of the TASEP and obtain analytic results that provide practical insights into translation dynamics. Our approach is to study the process in a continuum limit called the hydrodynamic limit, which leads to a general PDE satisfied by the density of particles. Upon solving this PDE, we obtain exact closed-form expressions for stationary currents and particle densities that agree very well with Monte Carlo simulations of the original TASEP model. Furthermore, we provide a complete characterization of phase transitions in the system. These results allow us to identify the key parameters that govern translation dynamics, and to formulate a set of specific design principles for optimizing translation efficiency in terms of protein production rate and resource usage. Using experimental ribosome profiling data of *S. cerevisiae*, we show that the translation system of this organism is generally efficient according to the design principles we found.

2.2 Results

We first present our theoretical results on a mathematical model of translation and identify the key parameters that govern its dynamics. We then apply our theoretical results to formulate four simple design principles that detail how to tune these parameters to optimize the overall rate of protein synthesis and efficiency of ribosome usage. We then analyze ribosome profiling data of *S. cerevisiae* and demonstrate that its translation system is generally efficient, consistent with the design principles we found.

Theoretical Results on a Stochastic Model of Translation

Model description of the inhomogeneous ℓ-TASEP

At a high level, translation of mRNA involves three types of movement of the ribosome, as illustrated in Figure 2.1A: 1) Initiation – a small ribosomal subunit enters the open reading frame so that its A-site is positioned at the second codon and then a large ribosomal subunit binds with the small subunit. 2) Elongation – the nascent peptide chain gets elongated by one amino acid and the ribosome moves forward by one codon. 3) Termination – the ribosome with its A-site at the stop codon unbinds from the transcript. An important point to note is that more than one ribosome can translate the same mRNA transcript simultaneously, so the movement of a ribosome can be obstructed by another ribosome in front, similar to what happens in a traffic flow on a one-lane road. Such interaction is what makes the dynamics difficult to analyze.

We model the flow of ribosomes on mRNA using a generalized TASEP, called the inhomogeneous ℓ-TASEP, on a one-dimensional lattice with N sites (see Figure 2.1B). In this process, each particle (corresponding to a ribosome in mRNA translation) is of a fixed size $\ell \in \mathbb{N}$ and is assigned a common reference point (e.g., the midpoint in the example illustrated in Figure 2.1B). The position of a particle is defined as the location of its reference point on the lattice. A configuration of particles is denoted by the vector $\boldsymbol{\tau} = (\tau_1, \dots, \tau_N)$, where $\tau_i = 1$ if the i^{th} site is occupied by a particle reference point and $\tau_i = 0$ otherwise. The jump rate at site i of the lattice is denoted by $p_i > 0$. During every infinitesimal time interval dt, each particle located at position $i \in \{1, \dots, N-1\}$ has probability $p_i dt$ of jumping exactly one site to the right, provided that the next ℓ sites are empty; particles at positions between $N - \ell + 1$ and N, inclusive, never get obstructed. Additionally, a new particle enters site 1 with probability αdt if $\tau_i = 0$ for all $i = 1, \dots, \ell$. If $\tau_N = 1$, the particle at site N exits the lattice with probability βdt. The parameter α is called the entrance (or initiation) rate, while β is called the exit (or termination) rate.

The hydrodynamic limit

The key quantities of interest are the stationary probability $\langle \tau_i \rangle$ of any individual site i being occupied or not, and the current (or flux) J of particles in the system. In the corresponding translation process, these quantities reflect the local ribosomal density and the protein production rate, respectively.

In the special case of the homogeneous 1-TASEP ($p_i = p$ for all i and $\ell = 1$), the stationary distribution of the process decomposes into matrix product states, which can be treated analytically [DEHP93]. Unfortunately, in the general case this approach is intractable, necessitating alternative methods such as the hydrodynamic limit. When $\ell > 1$, deriving the hydrodynamic limit is not straightforward, however, as the process does not possess stationary product measures [SS04]. To tackle this problem, we mapped the ℓ-TASEP to another interacting particle system called the zero range process (ZRP, see Section A.1 of the appendix and Figure A.1), whose hydrodynamic limit, assuming it exists, can be derived from the associated master equation. More precisely, we obtained the hydrodynamic limit through Eulerian scaling of time and space by a factor $a = N^{-1}$, and by following its dynamics on scale x such that $k = \left\lfloor \frac{x}{a} \right\rfloor$, for $1 < k < N$ [Rez91]. Implementing this limiting procedure for the ZRP and mapping it back to the inhomogeneous ℓ-TASEP, we found that the limiting occupation density $\rho(x,t) := \mathbb{P}(\tau_k(t) = 1)$, assuming its existence, satisfies the nonlinear PDE

$$\partial_t \rho = -\partial_x [\lambda(x) \rho\, G(\rho)] + \frac{a}{2} \partial_{xx} [\lambda(x) G(\rho)] + O(a^2), \tag{2.1}$$

where $G(\rho) = \dfrac{1 - \ell\rho}{1 - (\ell - 1)\rho}$ and λ is a differentiable extension of $(p_1, \dots, p_N)$, such that $\lambda(x) = \lambda(ka) = p_k$. More generally, this PDE takes the form of a conservation law with systematic and diffusive currents J and J_D, given by

$$J(\rho, x) = \lambda(x) \rho G(\rho) \quad \text{and} \quad J_D(\rho, x) = \frac{\lambda(x)\rho}{1 - (\ell - 1)\rho}. \tag{2.2}$$

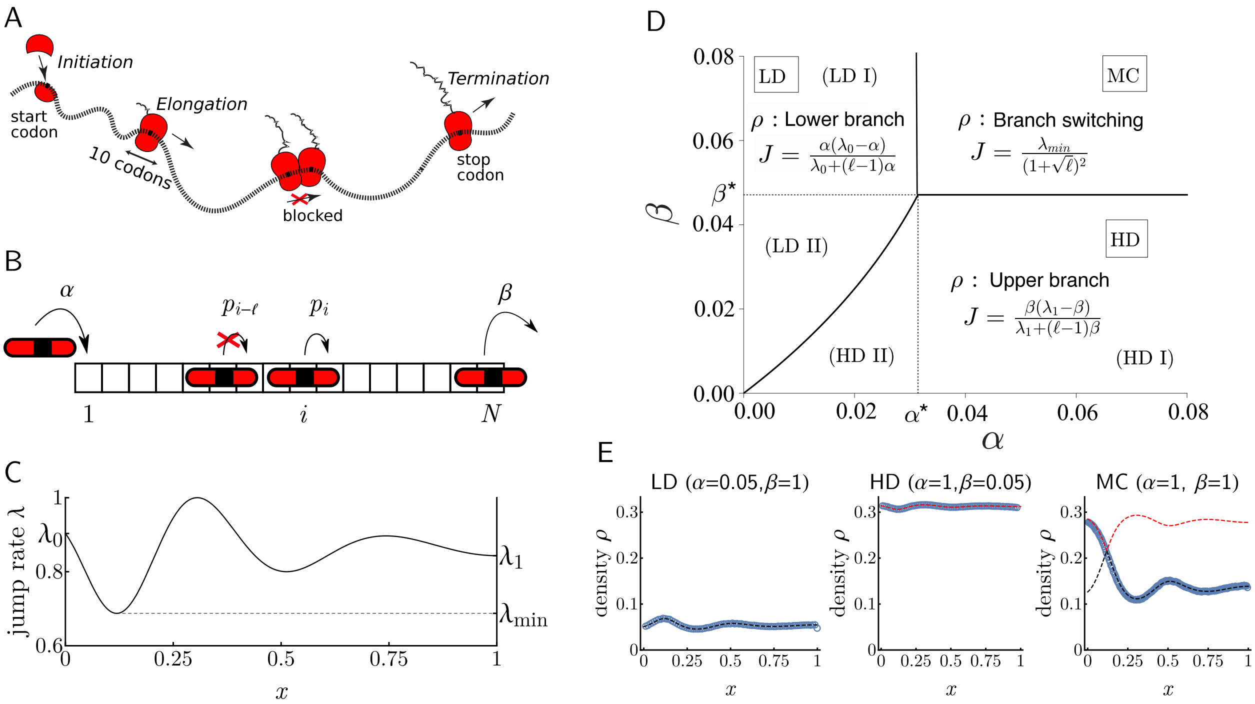


Figure 2.1: **Illustration of the translation process, the inhomogeneous ℓ-TASEP with open boundaries, and its phase diagram. A:** Ribosomes initiate translation at the mRNA $5'$ end, elongate the polypeptide by decoding one codon at a time, and eventually terminate the process by detaching from the transcript. **B:** Particles (of size $\ell = 3$ here) enter the lattice at rate α and a particle at position i (here defined by the position of the midpoint of the particle) moves one site to the right at rate p_i, provided that the move is not blocked by another particle in front. **C:** Example rate function with key parameters shown. **D:** The phase diagram is completely determined by $\lambda_0, \lambda_1, \lambda_{\min}$ and ℓ. In this example, $(\lambda_0, \lambda_1, \lambda_{\min}, \ell) = (0.9, 0.3, 0.1, 10)$. All phase transitions are continuous in J and, unless $\lambda_{\min}$ coincides with λ_0 or λ_1, discontinuous in ρ. **E:** Simulated results for $\ell = 3, N = 800$, and λ as in **C** are compared with theoretical predictions. Dashed black and red lines represent upper and lower branches of solutions to (2.1). Circles are averaged counts over 5×10^7 Monte-Carlo steps after 10^7 burn-in cycles.

As $a \ll 1$, the systematic current dominates and solutions of (2.1) generically converge locally uniformly on $(0, 1)$ to so-called entropy solutions of

$$\partial_t \rho = -\partial_x \left[\lambda(x) \rho \, G(\rho)\right]. \tag{2.3}$$

Further details and relevant calculations are provided in Section A.1 of the appendix.

Particle densities, currents and phase transitions

The first order nonlinear PDE given by (2.3) can be solved using the method of characteristics [Eva10], which describes the evolution of differently dense "patches" of particles over time. Solving for the characteristics yields two branches of solutions, which we call "upper" and "lower" branches, while the boundary conditions imposed by α and β determine which branch is taken by the stationary density of particles (see Section A.2 of the appendix). As a consequence, the behavior of the system is characterized by a phase diagram in α and β. Moreover, this phase diagram depends on only few parameters of the system (see Figure 2.1C): the size of particles ℓ, the jump rates at the boundaries, $\lambda_0 := \lambda(0)$ and $\lambda_1 := \lambda(1)$, and the minimum jump rate $\lambda_{\min} := \min\{\lambda(x) : x \in [0,1]\}$. In particular, these parameters determine the critical initiation and termination rates, α^* and β^*, that are associated with phase transitions. More precisely, the critical initiation rate α^* is given by

$$\alpha^* = \frac{\lambda_0 - (\ell-1)J_{\max}}{2}\left[1 - \sqrt{1 - \frac{4\lambda_0 J_{\max}}{[\lambda_0 - (\ell-1)J_{\max}]^2}}\right], \tag{2.4}$$

where $J_{\max} = \frac{\lambda_{\min}}{(1+\sqrt{\ell})^2}$. Note that α^* is determined by the jump rates λ_0 and $\lambda_{\min}$. In the context of translation dynamics, this means that α^* will be specific to each gene, as different genes will likely have different values of λ_0 and $\lambda_{\min}$. For a fixed λ_0, the critical rate α^* increases as $\lambda_{\min}$ increases. For a fixed $\lambda_{\min}$, it turns out that α^* satisfies

$$\frac{\lambda_{\min}}{(1+\sqrt{\ell})^2} \le \alpha^* \le \frac{\lambda_{\min}}{1+\sqrt{\ell}}, \tag{2.5}$$

where the lower bound is achieved as $\lambda_0 \to \infty$, while the upper bound is achieved when $\lambda_0 = \lambda_{\min}$. More generally, for a fixed $\lambda_{\min}$, the critical initiation rate α^* decreases as λ_0 increases. The critical termination rate β^* is obtained from (2.4) by replacing λ_0 with λ_1. Hence, for mRNA translation, β^* is also gene-specific, determined by the key elongation rates λ_1 and $\lambda_{\min}$.

The resulting phase diagram, which generalizes previous formulas for the homogeneous 1-TASEP [DEHP93], is summarized as follows (see Figure 2.1D):

1. If $\alpha < \alpha^*$ and $\beta > \beta^*$ (LD I): In this regime the flux is limited by the initiation rate, leading to a *low density* profile. The corresponding current assumed by the system is

$$J_L = \frac{\alpha(\lambda_0 - \alpha)}{\lambda_0 + (\ell-1)\alpha}, \tag{2.6}$$

while the site-specific particle density is

$$\rho_L(x) = \frac{1}{2\ell} + \frac{J_L(\ell-1)}{2\ell\lambda(x)} - \sqrt{\left[\frac{1}{2\ell} + \frac{J_L(\ell-1)}{2\ell\lambda(x)}\right]^2 - \frac{J_L}{\ell\lambda(x)}}. \tag{2.7}$$

2. If $\alpha > \alpha^*$ and $\beta < \beta^*$ (HD I): Now the flux is limited by the particle exit rate, resulting in a *high density* regime. The associated current J_R and density ρ_R are identical to J_L ((2.6)) and ρ_L ((2.7)), respectively, with λ_0 and α replaced by λ_1 and β.

3. If $\alpha < \alpha^*$ and $\beta < \beta^*$ (LD II and HD II): The steady state is determined by the sign of $J_L - J_R$ (computed as above). If it is positive ($J_L > J_R$), the system is in a low density regime with current and density given by J_L and ρ_L, respectively. Conversely, if it is negative, the system is in a high density regime with J_R and ρ_R as the current and density.

4. If $\alpha > \alpha^*$ and $\beta > \beta^*$ (MC): The system carries the *maximum possible current* (also referred to as the transport capacity of the system)

$$J_{\max} = \frac{\lambda_{\min}}{(1+\sqrt{\ell})^2}, \tag{2.8}$$

which is limited only by the minimum elongation rate $\lambda_{\min}$. Its density is characterized by qualitatively different profiles to the left and right of $x_{\min} = \arg\min_x \lambda(x)$: For $x < x_{\min}$, $\rho(x)$ is described by the upper branch (obtained by replacing J_R with $J_{\max}$ in the equation for ρ_R), while for $x > x_{\min}$, $\rho(x)$ is described by the lower branch (obtained by replacing J_L with $J_{\max}$ in ρ_L). That is, a branch switch occurs at $x_{\min}$ (where $\rho(x_{\min}) = (1+\sqrt{\ell})^{-2}$). We proved more generally that every global minimum of λ regulates the traffic of particles (like a toll reducing the traffic flow) in this fashion: incoming densities to the left of it are always described by the upper branch whereas outgoing particles on the right follow the lower branch. In particular, this implies that in the case of multiple global minima, the density between two consecutive minima must undergo a discontinuous jump from lower to upper branch (for more details, see Section A.2 of the appendix and Figure A.3).

Novel phenomena and applicability to discrete lattices

As shown in Figure 2.1E, for smooth rate functions the densities predicted by our analysis agree well with Monte Carlo simulations in all regimes of the phase diagram. In the context of translation dynamics, however, elongation rates are typically less regular, exhibiting substantial fluctuations throughout the entire transcript (see Figure 2.2A). Despite this lack of regularity, the hydrodynamic limit can still be employed to describe local averages of such a system. In particular, smoothing particle profiles by windows of length ℓ reproduces parameters that closely match hydrodynamic predictions (see Section A.2 of the appendix and Figure A.5). Hence, all subsequent analyses described below will pertain to elongation rate profiles smoothed by a ten-codon moving average. A noteworthy consequence of the above results is that local averages of elongation rates are more predictive of overall translation dynamics than their non-smoothed counterparts. In particular, the location at which branch switching occurs in the MC regime is governed by $x_{\min} = \arg\min_x\{\overline{p}_x\}/N$ which may be, and in many cases is, considerably different from $\arg\min_x\{p_x\}/N$ (cf. Figure A.4).

We highlight a few novel phenomena in our generalization of the homogeneous 1-TASEP: First, extending particles to size $\ell > 1$ and lowering the limiting jump rate $\lambda_{\min}$ reduces both the transport capacity $J_{\max}$ and the critical rates (α^* and β^*) for entrance and exit, leading to an enlarged MC phase region. This is expected as fewer particles are needed to saturate the lattice, and distances between particles are larger, which in turn limits the number of particles able to cross a site per given time. This phenomenon is quantified precisely using our explicit expressions for α^*, β^*, and $J_{\max}$ (see (2.4) and (2.8)). Second, the inhomogeneity in λ may deform the LD-HD phase separation

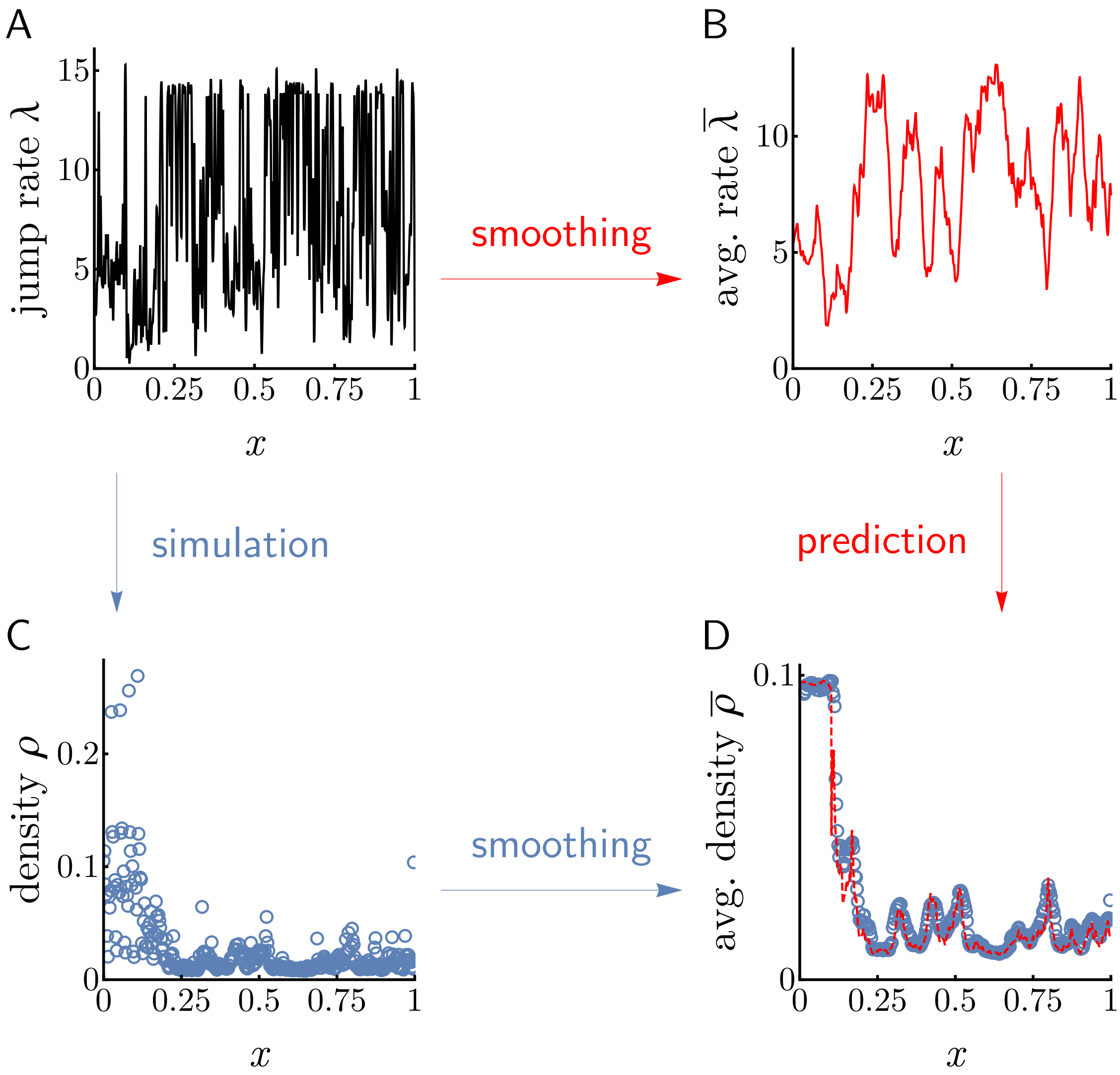


Figure 2.2: **Local averaging reproduces hydrodynamic limit in lattices with discontinuous rate functions.** Applying the hydrodynamic theory to smoothed jump rates correctly predicts smoothed density profiles and currents. **A:** Elongation rates of the yeast gene YHR025W arbitrarily chosen from [DDS18] (see Section 2.3 for further details). **B:** Smoothed elongation rates obtained by applying a ten-codon moving average to the raw profile in **A**. **C:** Density profile resulting from simulation (as in Figure 2.1E except with $\ell = 10, N = 357$) under discontinuous profile in **A**. **D:** The hydrodynamic density profile (dashed red) associated with the smoothed elongation rates of **B** reproduces the smoothed density profile obtained from averaging the raw densities in **C** by a moving ten-codon window. Similarly, simulated and predicted currents are in excellent agreement (0.1072 and 0.1077, respectively).

from being a straight line in the homogeneous ℓ-TASEP [CL04] to a generally nonlinear curve (see Figure 2.1D) determined by solutions (α, β) of

$$\frac{\alpha(\lambda_0 - \alpha)}{\lambda_0 + (\ell - 1)\alpha} = \frac{\beta(\lambda_1 - \beta)}{\lambda_1 + (\ell - 1)\beta}, \tag{2.9}$$

corresponding to the condition $J_L = J_R$. This is a consequence of α and β affecting the system at different scales whenever $\lambda_0 \neq \lambda_1$, resulting in a phase diagram that is no longer symmetric. Lastly, our observation of density profiles performing branch switching in the MC phase was indiscernible in the homogeneous case, as the high density and low density branches merge into a single value (viz. $\rho = \frac{1}{\sqrt{\ell}+\ell}$).

2.3 Application: Design Principles for Translational Systems

We sought to apply our theoretical analysis to understand how the translational system can be regulated and optimized with regard to protein synthesis rate and ribosome usage. The hydrodynamic theory developed above singles out the key parameters that determine the current and particle densities. We illustrate in Figure 2.3 how λ_0, λ_{min}, and x_{min} impact the current capacity, its sensitivity to the initiation rate α, and the global particle density, suggesting the following principles:

1. *The initiation rate α (and not termination rate β) should regulate the production rate J.* As shown by our analysis of the current, any value of the current that lies below the system's production capacity J_{max} can be attained through either HD or LD regime. In order to avoid overuse of resources, however, a transcript should always operate in LD, where the main determinant for currents is the initiation rate α (cf. (2.6)). To guarantee LD profiles, termination rates merely need to exceed the critical value β^*, whereas initiation rates are more tightly controlled, varying between 0 and α^*. Within this interval, the current J increases with α according to (2.6), as illustrated in Figure 2.3A.

2. *The minimum elongation rate λ_{min} determines the production capacity J_{max}.* As α increases in the LD regime, the current J reaches a plateau that is associated with the maximal current (MC) regime (see Figure 2.3A). By (2.8), the maximum possible current is directly proportional to λ_{min}, which therefore sets the range within which production rates may vary. Large values of λ_{min} allow for both constitutively high expression of genes as well as highly variable protein levels, while small values of λ_{min} guarantee constitutively low expression.

3. *In the LD regime, the sensitivity of production rate J to α is moderated by λ_0 and varies across different values of α.* Our theory predicts that for $\beta > \beta^*$ (i.e., provided that the termination rate is sufficiently high), the dynamic range of the initiation rate (i.e., the range of α within which the overall protein production rate J varies with α) is given by $(0, \alpha^*)$, where the critical initiation rate α^* is defined in (2.4). Furthermore, the degree to which J varies with α is fully determined by the elongation rate λ_0, as shown in (2.6). Indeed, λ_0 controls the time spent by particles at the start of the lattice, and can induce significant buffering if α is large enough, thereby modulating the effective rate of entrance associated with J. We illustrate this in Figure 2.3A, where we compare how the current varies as a function of α for different values of λ_0 relative to λ_{min}. Recall that the

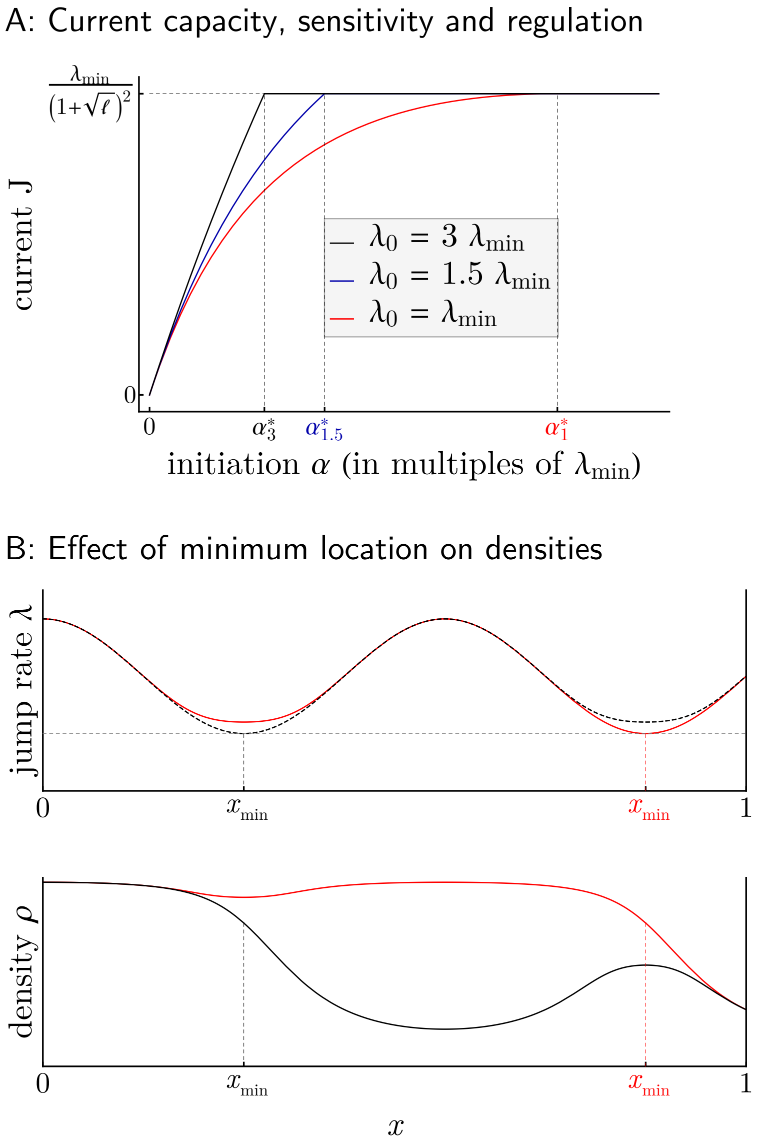


Figure 2.3: **Main determinants of current and particle densities. A:** J in LD and MC as a function of α, for various choices of λ_0. While λ_{min} governs the maximum current at which J plateaus, varying λ_0 results in changes in $\partial_\alpha J$, the α-sensitivity of J. Distinct $(\lambda_{\text{min}}, \lambda_0)$ pairs give rise to different α-dependencies of J, suggesting different responses to global changes in the ribosome pool. α_3^*, $\alpha_{1.5}^*$, and α_1^* correspond to the α^* value (in units of λ_{min}) when $\lambda_0 = 3\lambda_{\text{min}}, \lambda_0 = 1.5\lambda_{\text{min}}$, and $\lambda_0 = \lambda_{\text{min}}$, respectively. **B:** Two λ-profiles that are close in L_p, but with far apart x_{min} are plotted (top panel) together with their associated MC ribosome densities (bottom panel). The branch switching strongly affects equilibrium particle densities and hence ribosomal costs, with λ-profiles achieving x_{min} close to 0 (top, dotted black curve) benefiting from substantial savings (bottom, black curve) compared to otherwise similar profiles (red curves).

critical initiation rate α^* satisfies the inequalities in (2.5), and that α^* increases as λ_0 decreases. Figure 2.3A also shows that for λ_0 fixed, the production rate of a system closer to the MC regime (i.e., with α just below α^*) is less sensitive to changes in α, and that this effect is more pronounced the closer λ_0 is to λ_{min}. More generally, the α-sensitivity of J increases as λ_0 increases. While the dependence of J in α is sublinear for $\lambda_0 = \lambda_{\text{min}}$, it becomes linear as λ_0 gets large (see (2.6)). This suggests in particular that changes in the free ribosome pool (changing the initiation rate globally) can impact the protein production rate differently across different genes.

4. *Positioning λ_{min} close to the start site can reduce the amount of ribosomes used.* At maximum production capacity (MC regime), we have shown that the density profile follows the high density branch from the start of the lattice until the location x_{min} of λ_{min} whereafter it adopts the low density branch. This characteristic branch switching phenomenon makes x_{min} critical for the purpose of resource allocation. In Figure 2.3B, we illustrate how a small local change in the rate function can induce a large increase of average particle density when x_{min} changes substantially. Therefore, a way to limit the excessive usage of ribosomes induced by traffic jams at maximum capacity is to position the minimum rate close to the start. However, as previously shown, positioning it too close to the start (such that $\lambda_0 = \lambda_{\text{min}}$) would also decrease the sensitivity of the system to α.

Empirical Study: Translational Efficiency in Yeast

In light of the aforementioned principles, we explored the extent to which the translational system in yeast is efficient. For this study, we used elongation rates previously inferred from ribosome profiling data for a set of 850 genes in *S. cerevisiae* [DDS18] (see Section A.2 of the appendix). These genes were selected in [DDS18] based on length and footprint coverage, to yield robust estimates of rates. The advantage of using this particular dataset over most others lies in the fact that the inferred rates for this subset of genes faithfully reproduce ribosome profiling data, incorporating several experimental artifacts of ribo-seq such as undetected stacked ribosomes, thereby minimizing confounding from technical biases. Furthermore, primarily analyzing high-coverage (and thus likely highly expressed) genes does not confound our study of design principles, but rather provides us an increased signal-to-noise ratio, as these genes are precisely those on which our design principles are expected to act most strongly.

We analyzed the location of these 850 genes in the phase diagram, and the distribution of the key parameters and variables that determine the ribosomal currents and densities. We found the aforementioned theoretical design principles being reflected as follows:

1. *Translation mainly operates in LD regime.* Upon computing α^* and β^*, we located the position of each gene in the phase diagram (see Figure 2.4A). Over the 850 genes in our dataset, we found 841 in LD and the remaining 9 in the MC region. No genes were found in HD, suggesting no excessive usage of ribosome to achieve any protein level. As a result, the initiation rate is the main determinant and limiting factor of the current (Spearman's rank correlation coefficient $\rho = 0.979$). The strength of this correlation nevertheless decreases as genes get closer to the MC regime, since J becomes less sensitive to α and λ_{min} becomes its rate limiting factor (see Figure 2.4C). To quantify this reduction in correlation, we binned the data by quartiles of J and computed Spearman correlations within each bin, which yielded (in order of quartiles): $0.93, 0.72, 0.64$, and 0.58.

2. *Wide ranges of currents are covered within production capacity.* For each gene in our dataset, we examined the maximal protein production rate, which according to our theory is proportional to λ_{min}. The data exhibit an overall range of λ_{min} between 1.01 and 6.01 codons/second, and for any fixed λ_{min}, currents are well spread out across $[0, J_{\text{max}}]$ (see Figure 2.4D). Given that genes cover almost all of the theoretically possible range of currents, we investigated whether certain configurations of λ_{min} and J are associated with the biological function of specific genes. To do so, we compared ribosomal protein genes (known to be highly expressed) and genes related to stress response (requiring variable expression over time, see Section A.2 of the appendix). We found that, while both sets of genes display comparable λ_{min}, ribosomal genes are more likely to be close to their maximal production capacity ($p < 7 \times 10^{-3}$, see of the appendix) and more consistently so (the coefficient of variation is 0.22 for ribosomal genes and 0.36 for stress response).

3. *λ_0 (associated with sensitivity to α) is higher for genes that are either highly expressed or subject to varying expression demand.* The impact of increasing α-sensitivity is primarily twofold: First, for fixed production capacity, large currents may be attained with smaller initiation rates; and second, more substantial changes in currents may be achieved with small changes in α. To investigate the former we computed α^*, the critical rate necessary for a gene to attain maximum capacity, across all genes whose λ_{min} exceeded the median λ_{min} of the data set (as large currents presuppose large capacities). Further binning this range into quartiles (to isolate the dependence of α^* on λ_0), we found that genes whose currents are at least 90% of the production capacity are significantly more sensitive ($p < 0.008, 0.01, 0.05$, and 0.004, respectively; see Figure 2.4E), requiring smaller initiation rates to reach peak production rate (cf. Figure 2.4C). To inspect the second aspect of λ_0 as facilitator or inhibitor of rapid changes in current, we explored the ratio of λ_0 to λ_{min} again in ribosomal and stress response genes. For constitutively highly expressed genes like ribosomal genes, we expect this ratio to be small to maintain stable current close to MC (cf. Figure 2.3), whereas genes with variable expression demands like the ones associated with stress response should exhibit larger ratios. Confirming this intuition, we found significantly reduced levels of $\lambda_0/\lambda_{\text{min}}$ in ribosomal genes ($p < 2 \times 10^{-6}$), and significantly increased levels in stress response genes ($p < 0.04$).

4. *The position of λ_{min} is preferentially located early in the open reading frame.* Upon analyzing the distribution of x_{min} from our dataset (see Figure 2.4B), we found it preferentially located in the codon positions between 30 to 40, consistent with genes forestalling excessive ribosome usage through enforcing branch switching early on. More specifically, we reasoned that both genes closer to MC and those highly sensitive to α run higher risk of incurring substantial ribosome cost and should thus locate x_{min} early in the coding sequence. Indeed, both the top quartile of genes close to MC (as measured by α/α^*) and stress response associated genes showed significantly smaller x_{min} ($p < 0.03$ and 0.01, respectively). Moreover, genes with unusually large values of x_{min} are significantly less likely to be close to MC (top quartile of x_{min}: $p < 1 \times 10^{-3}$).

To check for systematic biases potentially present in our subsampled gene set and to show replicability of our main biological conclusions, we also analyzed two other independent (and much larger) datasets from [WJW14] (combined with polysome profiling from [MLF+04]) and [PRI+14] (see Section A.2 of the appendix). We inverted the solution of (2.3) to obtain approximate estimates of initiation rates, termination rates, and smoothed elongation rates for these datasets, and repeated our analyses. As shown in Figure A.6, the results are generally in excellent agreement with what is discussed above (Figure 2.4A,B).

2.4 Discussion

While past quantitative studies of the TASEP under general conditions of extended particle size and/or rate heterogeneity have mostly been limited to numerical simulations or mean-field approximations, [LC03, SZL03, SSL04, CL04, DSZ07], we used here a different approach that relies on studying the hydrodynamic limit of the process. In the case of homogeneous rates, previous studies [Sch05, SS04] established this hydrodynamic limit, but without further analyzing the subsequent PDE. After deriving this limit for inhomogeneous rates, we obtained closed-form formulas for the associated current, densities, and phase diagram, generalizing previous theoretical results for the TASEP [DEHP93, BE07] and its variants [SZL03, CL04, SdQ11]. Our approach has the advantage of revealing the key parameters that the current and densities depend on, enabling an immediate quantification of the process and its phase diagram. Such a quantification is difficult to achieve via conventional stochastic simulations or approximations used in the past several years [ZDS11, ZT16, SNCR18].

Our characterization of the current and densities in the phase diagram suggests that, in agreement with earlier experimental studies [KGC^{+}13, SMV09], translation dynamics should be mainly governed by the initiation rate, while the termination rate and most elongation rates have negligible impact. In particular, our results explain why having the initiation rate as the main limiting factor of the current [PK11] minimizes ribosome usage. In addition, we discovered the importance of smoothed rather than raw elongation profiles in predicting translation dynamics, explaining the previously observed mild effect that any individual elongation change has compared to accumulated, neighboring changes [LT18]. This allowed us to identify two key parameters of the system, namely, the smoothed elongation rate λ_0 immediately following initiation and the minimal smoothed elongation rate $\lambda_{\min}$. Previous studies have established some association between the sequence context in the early $5'$ coding region and protein production levels [FSR^{+}17, BLN^{+}16, BYAZ^{+}15]. For example, it has been shown that mRNA secondary structure in the first ~ 16 codons (which locally decreases the elongation rate) negatively affects the translation rate in *E. coli*, while no significant contribution of mRNA folding in other regions was found [FSR^{+}17]. By exposing α and λ_0 as the only parameters that currents in LD depend on, our analysis suggests a direct explanation for such contrast.

We also highlighted the impact of λ_0 on the sensitivity of the current to changes in α. In practice, initiation rates can vary at the individual gene level (e.g., through interactions with specific miRNAs [HWMP05]). According to our theory, the way that these variations impact the protein production rate depends on λ_0; we hence suggest that this may explain why genes associated with stress response present higher values of λ_0, as it facilitates the response to changes in α. At a more global level, our study shows how protein levels can be more or less robust against changes in the ribosomal pool, which can simultaneously affect all initiation rates in a cell [SDN^{+}13]. Since the level of ribosomes present in a cell fluctuates over time [WARF+18], it would be interesting to see if protein levels scale uniformly with these variations across genes, and if not, whether the differences in λ_0 can explain it.

To the best of our knowledge, the role of the minimum elongation rate $\lambda_{\min}$ has so far received attention only indirectly, through the study of what is known as the "$5'$ translational ramp" [TCV+10]. This ramp is a pattern of translational slowdown around codon position 30-50 followed by steadily accelerating elongation rates, which is mirrored by the spatial distribution of minimum elongation rates we found here. This ramp has been hypothesized to prevent crowding of ribosomes on the transcript [TCV+10], for which we provide a theoretical basis, exposing $\lambda_{\min}$ as a separator between crowded and freely elongating ribosomes. More generally, the complex interplay between the maximum current capacity, ribosome usage, and sensitivity to the initiation rate suggests various ways to set the parameters λ_0, $\lambda_{\min}$ and $x_{\min}$, depending on the desired object to optimize. For example, allocating the minimum elongation rate near the beginning of the ramp region provides an optimal trade-off between high sensitivity and minimal traffic jams. On the other hand, it would be optimal for genes with housekeeping function to have a decreased sensitivity, which would push the minimum to earlier positions.

Our analysis can also help to answer the long-debated question regarding the implication of translation on codon usage bias [HP08, FLG+18, SDN+13]. Since highly expressed genes are enriched for synonymous codons translated by more abundant tRNAs [YDZ+15, HC18], it has been hypothesized that codon usage bias increases the overall protein synthesis rate by accelerating elongation [HP08]. However, recent studies have challenged such a hypothesis, suggesting that translational selection for speed is not sufficient to explain the observed variation in codon usage bias [MA18]. Synonymous changes of the coding sequence modify local elongation rates, but, according to our theory, such a modification impact the overall protein production rate only if the smoothed elongation rates λ_0 or $\lambda_{\min}$ are affected. In addition, our work implies that synonymous codon replacements that substantially change the location $x_{\min}$ of $\lambda_{\min}$ affect the efficiency of ribosome usage, and hence are more likely to be under selective pressure.

Aside from these cases, there should be little *direct* impact of synonymous codon usage on translation efficiency; this prediction is consistent with previous studies that tried to explain differences in expression using codon identity [GMG+12], and to characterize the sensitivity of translational output with respect to changes in elongation [LT18]. Codon usage bias could affect the protein production rate *indirectly*, however, by reducing the cost of translation: replacing a codon by a "faster" synonymous codon helps to reduce the local ribosome density on the transcript, and this can in turn increase the availability of free ribosomes and therefore increase the initiation rate α slightly; in the LD regime, increasing α would increase the protein production rate.

We note that other factors such as mRNA decay [HC18], or reduction of nonsense errors or co-translational misfolding [Gil07, FLG+18] might be more important drivers of codon usage bias.

Finally, it would be interesting to experimentally test our theoretical predictions, e.g., using cell-free expression protocols such as lysate-based systems, which have been developed to optimize protein synthesis and more recently refined to study translation dynamics [MMF17, RC14, KF19]. By designing an appropriate mRNA sequence and controlling different components (NTPs, ribosomes, tRNAs, specific amino acids), these systems allow to manipulate the initiation and elongation rates, and hence tune the key parameters identified by our theoretical analysis. For example, one can modify λ_{min} or λ_0 by changing the level of corresponding amino acids, and vary α by modifying the $5'$ UTR sequence or changing the ribosome concentration. The flexible nature of such cell-free expression systems, coupled with precise measurement of protein levels (e.g., via isotope-labeled amino acids or reporter proteins), should help to verify our theoretical results. In particular, it would be interesting to experimentally demonstrate the existence of phase transitions, and by modifying the mRNA sequence, test our predictions on how to effectively control the robustness and sensitivity of the translation system. We are currently pursuing these research directions.

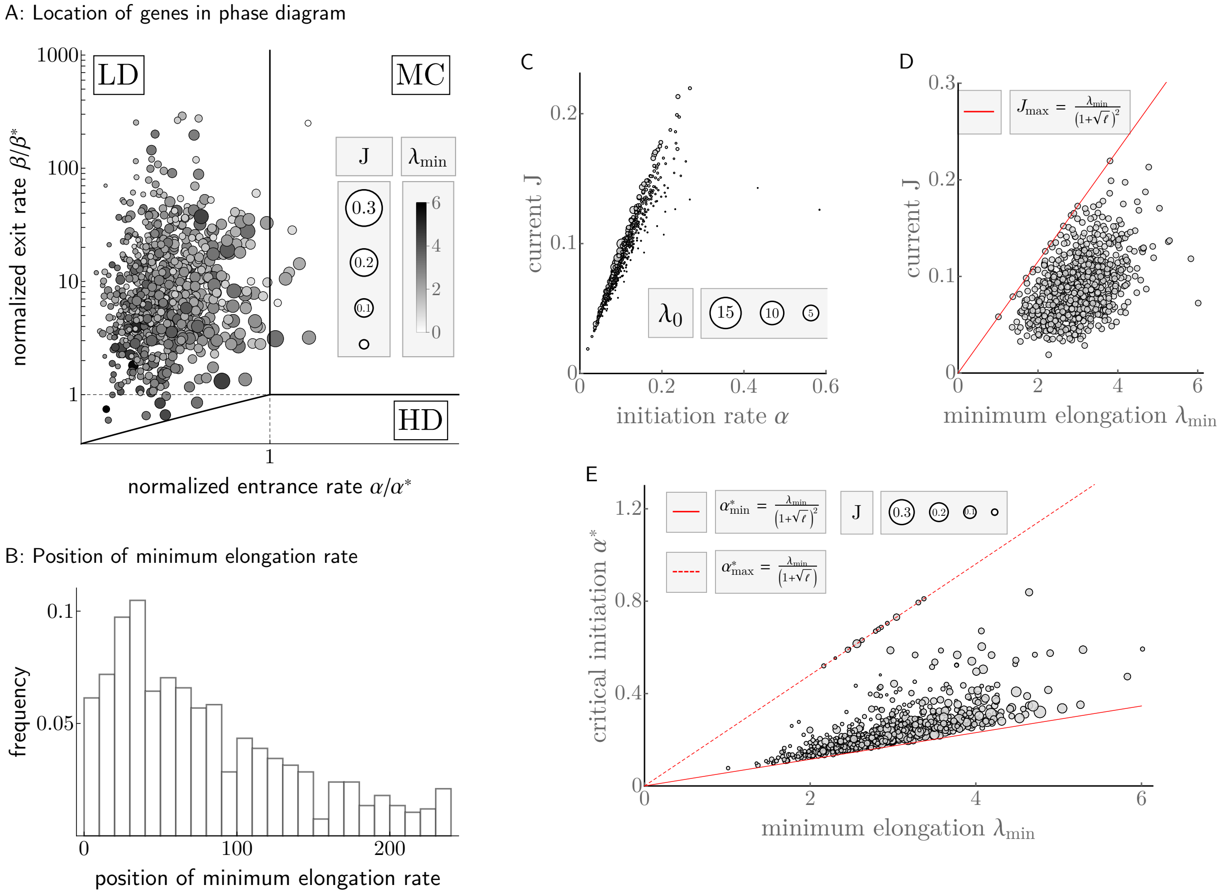


Figure 2.4: **Translation machinery in *S. cerevisiae* optimizes for ribosomal cost, flexible regulation and production capacity.** All rates are in codons per second, while currents are in ribosomes per second. **A:** 850 genes of *S. cerevisiae* located in the phase diagram, with size and hue reflecting current and minimum elongation rate. Systems of comparable production capacities ($\propto \lambda_{\min}$) fully exploit their dynamic range through α, with highly expressed proteins situated inside or close to MC. **B:** $\lambda_{\min}$ is placed early on in codon sequences to minimize ribosome cost. **C:** α is the main determinant of currents for low to average current genes, which the correlation for highly expressed genes decreases due to stronger variation in λ_0 and transitions into MC. **D:** Genes utilize full dynamical ranges of J through varying α and λ_0. Constitutively highly expressed genes tend to be closer to maximum capacity (red line), while genes with variable expression demands are distributed more broadly (see main manuscript). **E:** For fixed production capacity $\propto \lambda_{\min}$, α^* tends to be smaller for genes with larger production rates. That is, larger λ_0 facilitate attainment of large currents. Moreover, within highly expressed genes, those associated with variable expression patterns exhibit higher sensitivities (smaller α^*), whereas genes with constitutive high expression are found closer towards maximal insensitivity (dotted red line) as these configurations favor stable expression.

Chapter 3

Deconvolution of Bulk Transcriptomic Data

This chapter is joint work with Jonathan Fischer, Justin Hong and Yun S. Song, and is published in *Genome Research* [EPFHS21].

3.1 Introduction

Bulk RNA sequencing (RNA-seq) has proven a useful tool to investigate transcriptomic variation across organs/tissues, individuals, and various other biological conditions [MFR+15, SAB15]. Despite many successes, this technology's full potential is inherently limited because each experiment measures the average gene expression among a large group of cells, the composition of which is unknown. Thus, despite the reduction in technical and biological variability attained by averaging, bulk experiments are potentially confounded by cell type proportions when considering heterogeneous cell mixtures [LR14, SHF+16]. Such confounding impedes the direct comparison of samples, possibly leading to the spurious or missed inference of biologically relevant genes when attempting to identify clinically important differences. Moreover, cell type compositions are often independently informative of biological processes including organ function [CBK+06, HSL+18, KRS+13, YH17] and development [HSL+18, HFK+17]. For example, cell type infiltration has been found to correlate with disease progression, disease status, and complex phenomena such as aging [FNK+03, BSPG+17, ZZZ+19, BBK+16, SJNK17]. Unlike bulk experiments, single-cell technologies allow us to query the transcriptome at the resolution of individual cells. Resulting analyses often seek to characterize the heterogeneity within, or the differences between, specified cell types [SWGV14]. By isolating the expression patterns of each measured cell type, single-cell gene expression data can provide a reference to aid the inference of the cell type compositions of bulk samples; this process is known as deconvolution.

Computational rather than experimental estimation of cell type compositions is attractive for several reasons. Single-cell experiments are more expensive than their bulk counterparts and require heightened technical expertise to perform, often rendering the large-scale generation of single-cell gene expression data infeasible [GMA+19]. Furthermore, even when performed correctly, many protocols fail to capture cell types in an unbiased fashion, meaning empirical cell type proportions often are not reliable estimators of true organ/tissue compositions [Tra15]. Finally, deconvolution can be applied to the deep compendium of available bulk RNA-seq data to refine earlier analyses and probe previously unanswerable or heretofore unformulated questions. Consequently, the computational deconvolution problem has become a topic of intense methodological research (as detailed in [ACVMDP18]). The problem may be represented mathematically as

$$M\boldsymbol{\alpha} = \boldsymbol{b}, \tag{3.1}$$

where M is a gene-by-cell type matrix of cell type-specific gene expression averages, $\boldsymbol{\alpha}$ a vector of cell type mixing proportions, and $\boldsymbol{b}$ a vector of gene expression values in a bulk RNA-seq experiment. Depending on which of M, $\boldsymbol{\alpha}$, and $\boldsymbol{b}$ are measured, different approaches are appropriate. We focus on the case in which both M and $\boldsymbol{b}$ have been observed, albeit noisily, and it remains to infer $\boldsymbol{\alpha}$; this is known as supervised deconvolution. Early approaches frequently utilized predefined marker genes for well-studied cell types, restricting their applicability. More recent methods formulate the problem as a regression task to be solved by variants of non-negative least squares (e.g., MuSiC [WPS+19], DWLS [TDC+19], SCDC [DTU+20], and Bisque [JAR+20]) or with more sophisticated machine learning techniques (such as CIBERSORTx [NSL+19] and Scaden [MMO+20]). Though each paradigm presents its own strengths, both fail to replicate the benefits of explicit generative modeling. The result is algorithms which may perform well but lack the flexibility to extend beyond the estimation of cell type proportions.

In this chapter, we propose a new method, RNA-Sieve, which employs asymptotic theory and a novel optimization procedure to solve a probabilistic model of deconvolution via maximum likelihood estimation. We demonstrate its highly capable performance across a diverse array of scenarios, including different organs/tissues, cell types, and practical challenges. We then highlight newly opened avenues for continued development made feasible by our generative framework, including confidence regions and general hypothesis tests.

3.2 Results

Method Overview

Although a single run of bulk RNA-seq produces only a solitary gene expression vector, myriad cells contribute to this measurement. The obtained profile is hence a composite snapshot of the gene expression levels of numerous individual, putatively independent cells. When coupled with an assumption that cells of the same type behave similarly, this large number of cells permits the application of the central limit theorem (CLT) and the wealth of normal distribution approximations it implies. Conveniently, the marginal distribution for gene expression of an arbitrary cell in the bulk sample is a straightforward mixture distribution (see Equation (B.1) in Section B.1). The resulting CLT-derived likelihood consequently depends only on the means and variances of gene expression for each cell type and the respective cell type proportions in the bulk sample, the latter of which we make our goal to infer computationally. To estimate the requisite cell type-specific moments, RNA-Sieve uses gene expression measurements from scRNA-seq experiments. We further model the estimation error of the computed moments by once again invoking the CLT, and the combination of these two approximations yields a full, composite likelihood built using normal distributions. We subsequently infer cell type proportions via a custom-made maximum likelihood estimation procedure. Several features of our algorithm ensure accurate and robust results. Our alternating optimization scheme is split into two components to better avoid sub-optimal local minima, with a final projection step handling flat extrema to avoid slow convergence. We also incorporate a gene filtering procedure explicitly devised to improve cross-protocol stability, a crucial concern given that single-cell and bulk experiments will always be performed with different technological platforms. Our algorithm can also perform joint deconvolutions, leveraging multiple samples to produce more reliable estimates while parallelizing much of the optimization. In this setting, each included bulk sample improves the denoising of the single-cell reference regardless of its mixture proportions, leading to improved statistical performance. Finally, we wish to emphasize that our likelihood-based model allows us to pursue extensions which are infeasible using prior approaches. A notable example includes confidence regions for estimates (see Section 3.2), among others (see Section 3.3). Full mathematical and computational details are presented in Section B.1, and a schematic is displayed in Figure 3.1.

Performance in Pseudobulk Experiments

To establish RNA-Sieve's effectiveness, we performed *in silico* experiments using scRNA-seq data from the *Tabula Muris Senis* Consortium [T^+20]. In these experiments we built "pseudobulks" by aggregating reads from labeled cells in known proportions to use for deconvolution. We considered thirteen organs with between two and eleven cell types per organ. Moreover, counts generated via both the Smart-Seq2 and 10x Chromium protocols are available for each organ, enabling convenient cross-protocol comparisons. These are particularly important given that bulk and single-cell RNA-seq samples are always processed using different techniques. To evaluate RNA-Sieve, we compared its performance to that of six recently published methods as well as non-negative least squares (NNLS). Performance was assessed for each organ by computing the L_1 distance (absolute difference) between inferred and true proportions and dividing by the number of cell types present. Further details are provided in Section 18. We found that RNA-Sieve produced the smallest mean error in both possible reference/bulk configurations (Figure 3.2 and Table 3.1; full results in Table B.2). To better understand performance, we also visualized errors when aggregating by organ (that is, the column-wise distributions of the checkerboard plots in Figure 3.2, see Figure B.1). This demonstrated that RNA-Sieve performed at least as well as all competitors, and often notably better, in nearly all organs. Our strong performance across organs regardless of the number of cell types or similarities among them suggests that RNA-Sieve is versatile over a range of scenarios. Finally, we directly compared each method's errors to those of RNA-Sieve on the same deconvolution tasks (given by the row-wise distributions of the checkerboards in Figure 3.2, see Figure B.2). In each case, RNA-Sieve produced smaller errors than the other methods a majority of the time.

Although RNA-Sieve's nominal improvement in the average per-cell-type L_1 metric may appear minor at first glance, we note that a typical tissue consists of several cell types, and thus the overall error may accumulate rapidly. The constraint that mixture proportions sum to 1 means such

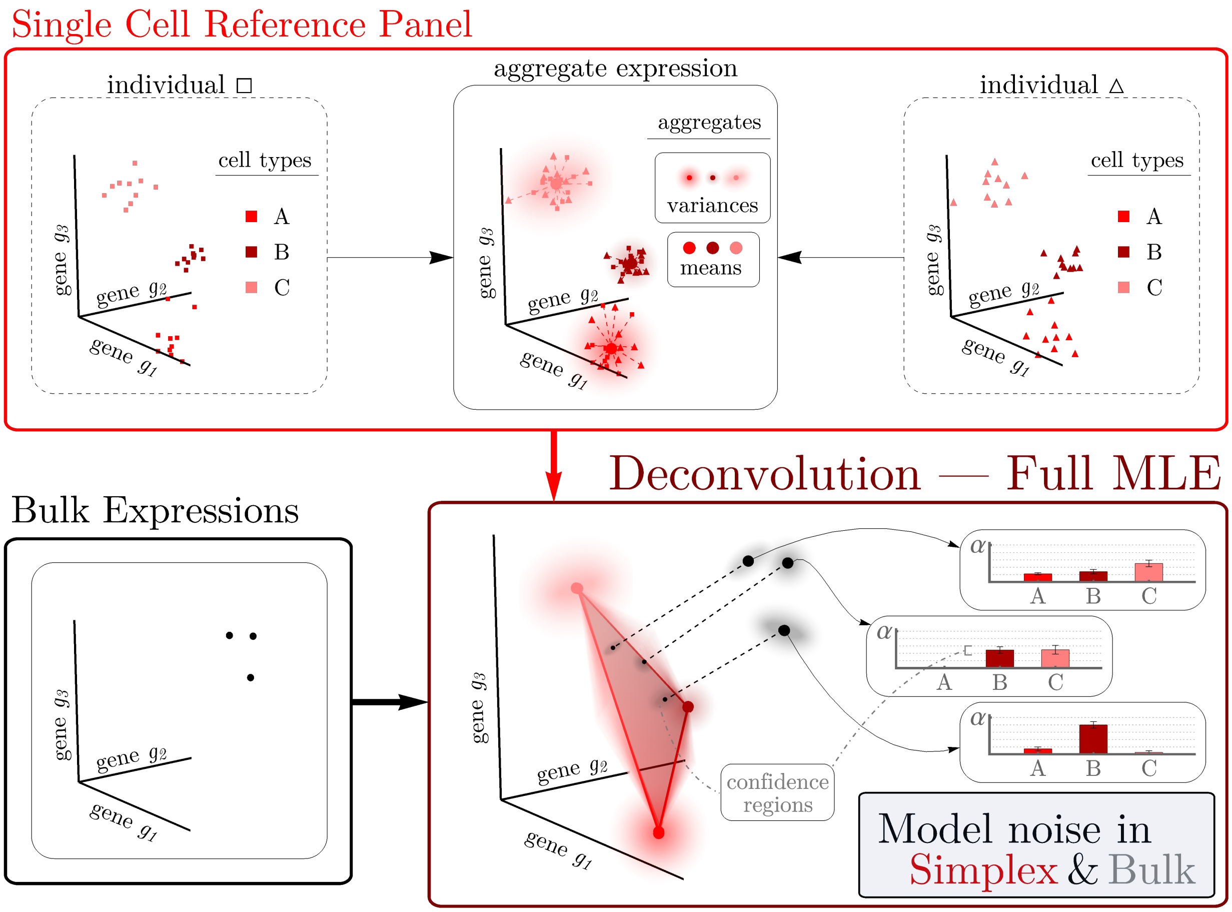


Figure 3.1: **The RNA-Sieve pipeline.** After applying a filtering procedure to scRNA-seq data, RNA-Sieve builds reference matrices for the mean and variance of expression for each gene across cell types. Using these estimates and bulk RNA-seq data, it performs joint deconvolution via maximum likelihood estimation by expressly modeling noise both in the reference and bulk data, yielding cell type proportion estimates and confidence regions for each sample.

Table 3.1: **Summary of deconvolution errors for each considered method in pseduobulk experiments.** Errors were computed as the L_1 distance (in %) between the inferred and true proportions averaged over the number of present cell types per organ. Single-cell RNA-seq data for the references and pseudobulks were taken from the *Tabula Muris Senis* experiment. The mean, median, and interquartile range are displayed for the results in thirteen different organs; see Section 18 for additional details.

(a) Smart-Seq2 reference and 10x Chromium pseudobulk.

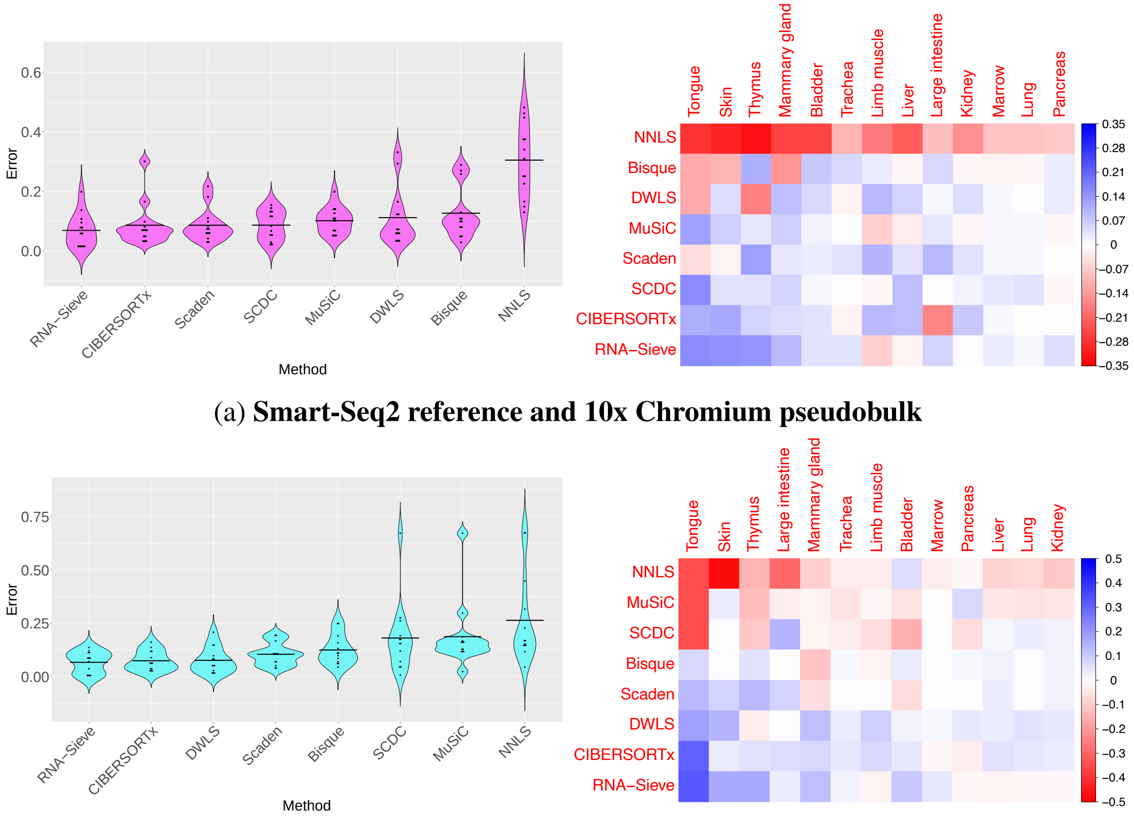
	RNA-Sieve	CIBERSORTx	Scaden	SCDC	MuSiC	DWLS	Bisque	NNLS
Mean	**6.9**	8.6	8.6	8.7	10.1	11.2	12.7	30.5
Median	**6.1**	7.1	7.2	8.3	10.6	7.2	10.1	31.0
IQR	7.7	3.4	4.1	7.9	6.0	6.9	5.2	15.3

(b) 10x Chromium reference and Smart-Seq2 pseudobulk.

	RNA-Sieve	CIBERSORTx	DWLS	Scaden	Bisque	SCDC	MuSiC	NNLS
Mean	**6.7**	7.4	7.6	10.5	12.5	18.1	18.7	26.4
Median	8.2	6.2	**5.4**	10.5	11.0	15.7	15.7	16.4
IQR	9.9	7.8	6.9	4.6	8.9	12.1	4.6	17.0

reductions in error are likely to be meaningful; when errors in inference are of the same order as the proportion of common cell types, it becomes very easy to arrive at incorrect biological conclusions, especially in more complex tissues with many cell types. We show in Figure B.3 a representative example of seemingly minor average *per-cell-type* improvement resulting in comparatively major *individual-cell-type* differences. Other error metrics are more sensitive to different aspects of performance and may detect such improvements more reliably, and are touched upon below. By virtue of having to consider multiple distinct algorithms, tissues, cell types, and experimental protocols, any benchmarking evaluation must necessarily consist of a large number of combinations. Between these many factors and the random nature of the data, it is (even theoretically) nearly impossible for any one algorithm to dominate the others in all situations (as is recognized and discussed in [MMO+20]). We believe that evaluation should therefore focus on aggregate measures of accuracy across many situations. We hence supplement Table 3.1 and Figure 3.2 with Table 3.2, which presents the mean ranks of all eight algorithms aggregated across all (26) cross-protocol experiments using the L_1, L_2, L_∞, and KL error quantifications as these accuracy metrics assess different aspects of model performance (see Section 18). We find RNA-Sieve outperforms its nearest competitor by roughly one-half rank regardless of error metric, a gap which is at least as large as those between other neighboring methods (NNLS excluded).

In practice, complications to the generic deconvolution problem may arise. For example, the scRNA-seq reference data may lack one or more cell types found in the bulk sample, or may even contain extra ones. Such problems are more likely to occur when performing cross-



(a) **Smart-Seq2 reference and 10x Chromium pseudobulk**

(b) **10x Chromium reference and Smart-Seq2 pseudobulk**

Figure 3.2: **Distribution of errors for each method in pseudobulk experiments.** Pseudobulk experiments were performed in 13 different organs using data from the *Tabula Muris Senis* experiment. Errors were computed as the average L_1 error across cell types in each organ. In the violin plots, horizontal black bars correspond to the mean error and methods are ordered left to right from lowest to greatest mean error. In the grid plots, methods and organs were ordered using SVD-induced clustering. Roughly speaking, the methods from top to bottom are characterized by improving performance while the organs from left to right are characterized by decreasing variability in different methods' performances. Color indicates the difference between the average error across methods in that organ; deeper shades of red (blue) indicate poor (good) relative performance. See Table B.1 for context regarding the cell types present in each organ.

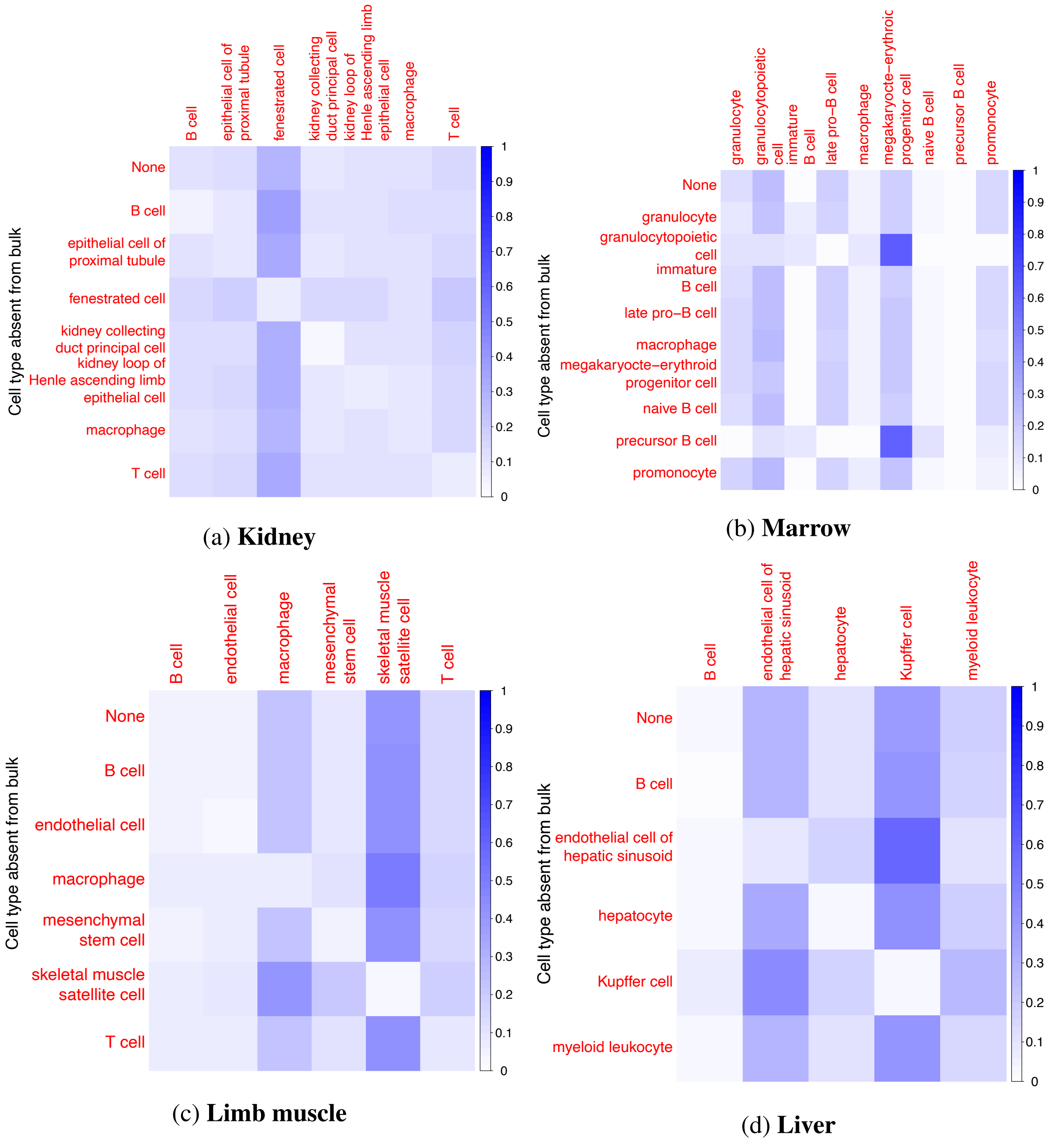
Table 3.2: **Mean ranking of algorithms under various error metrics.** All eight methods were ranked 1 (best) to 8 (worst) on all 26 cross-protocol deconvolutions using the L_1, L_2, L_∞ and KL (KL-divergence) metrics, and their mean ranks computed.

	RNA-Sieve	DWLS	CIBERSORTx	Scaden	Bisque	SCDC	MuSiC	NNLS
L_1	**2.9**	3.4	3.6	3.9	4.5	5.0	5.3	7.4
L_2	**3.0**	3.5	3.5	4.0	4.5	4.8	5.2	7.4
L_∞	**3.0**	3.7	3.5	4.0	4.4	4.8	5.2	7.3
KL	**2.9**	3.3	3.8	3.8	4.5	4.9	5.3	7.5

experimental or cross-subject deconvolutions, as we typically must. It is thus important to examine how algorithms perform in these situations. We further recognize the necessity to demonstrate robustness to misspecification for a model-based approach like RNA-Sieve. To do so, we selected the kidney, limb muscle, liver, and marrow due to their representative ranges of cell type number and dissimilarities, and considered all possible configurations containing one extra or missing cell type in the single-cell references. When the reference contains too many cell types, deconvolution schemes should infer proportions near zero for these extra cell types. We found that to be the case with RNA-Sieve (Figures 3.3 and B.4) as long as the extra reference cell type is sufficiently distinct from the other cell types present in the reference. When cell types are highly similar, inferred proportions may be shared among them and might not change substantially upon removal of one of these cell types from the bulk. Meanwhile, when a cell type present in the bulk is absent from the reference, the more likely of these two scenarios, the deconvolution problem becomes overdetermined. Ideally, deconvolution algorithms would move the weight of the removed cell type to those most similar to it. Our empirical results (Figures 3.4 and B.5) indicate that RNA-Sieve tends to do precisely this. In some cases, this means mass transfer to one single cell type, while in others the weight is shared among multiple. This result suggests that in the case of misspecification, RNA-Sieve will still achieve sensible solutions as long as sufficiently representative cell types are captured in the reference set. We note that given the generative nature of our model, a hypothesis test to detect missing cell types is, as opposed to existing methods, within the capabilities of our framework (see Section 3.3).

Validation with Real Bulk RNA-seq Data

In certain rare instances, bulk RNA-seq data sets with known or experimentally estimated cell type proportions are available. We considered three such data sets in order to evaluate RNA-Sieve under more realistic conditions. The first of these data sets is a bulk RNA-seq mixture of two human breast cancer cell lines and fibroblasts with accompanying scRNA-seq data first published in [DTU$^+$20]. These cells were mixed together in proportions of 60% MDA-MB-468, 30% MCF-7, and 10% fibroblasts. As shown in Table 3.3, RNA-Sieve yields highly accurate results, attaining the lowest average error among all methods. With the exception of SCDC, other methods overestimated the



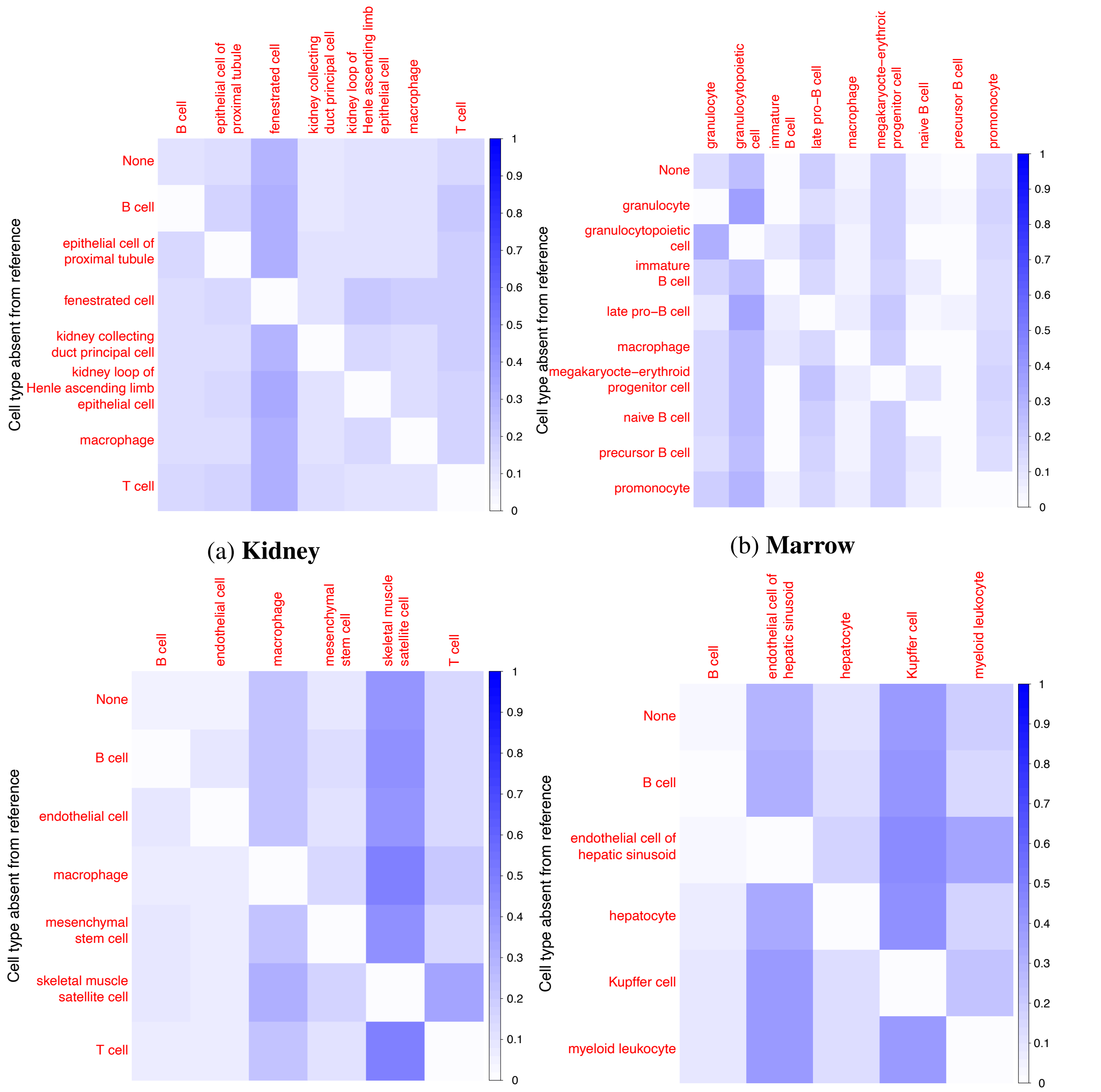
(a) **Kidney**

(b) **Marrow**

(c) **Limb muscle**

(d) **Liver**

Figure 3.3: **Deconvolution with extra cell types as reference.** Pseudobulk experiments were performed on four organs from the *Tabula Muris Senis*, in which one cell type is removed from the pseudobulk at a time through a leave-one-out procedure. The top row shows the inferred proportions with no extra reference cell types. Darker colors indicate a higher estimated proportion value. Here we used Smart-Seq2 data for the references and 10x Chromium for the pseudobulks.



(a) **Kidney**

(b) **Marrow**

(c) **Limb muscle**

(d) **Liver**

Figure 3.4: **Deconvolution with missing cell types in the reference matrix.** Experiments were conducted as in Figure 3.3 but with missing cell types in the reference instead of the bulk.

Table 3.3: **Inferred proportions from different methods in cell line mixture experiment.** Data from [DTU+20] with known cell type proportions was used to evaluate each applicable method (displayed proportions may not sum precisely to 1 due to rounding). Bisque and MuSiC are not intended for use with only one individual in the bulk data and/or single-cell reference and were thus not included. SCDC was run in tree mode for this deconvolution.

	Estimated Proportions (%)			
Method	60% MDA-MB-468	30% MCF-7	10% Fibroblasts	Mean L_1 Error
RNA-Sieve	62	26	13	3
SCDC	60	19	21	4
Scaden	35	44	21	17
CIBERSORTx	32	52	16	19
DWLS	26	48	27	23
NNLS	22	56	21	25

fraction of the MCF-7 cell line in the bulk while underestimating the MDA-MB-468 cell line by large amounts, and most methods substantially overestimated the fibroblast proportions.

Because this experiment contains only three cell types and a single bulk sample all from one experiment, we sought to validate using larger data sets containing expression measurements from peripheral blood mononuclear cells (PBMCs) which promised to be more heterogeneous. The first of these, analyzed in [NSL+19], measures gene expression in twelve bulk samples and a scRNA-seq reference from one individual. Ground-truth cell type proportions in all bulk samples were estimated using flow cytometry and were grouped into six primary categories: B cells, CD4+ T cells, CD8+ T cells, monocytes, natural killer (NK) cells, and neutrophils. The second PBMC bulk data set comes from [MLX+19] and contains a further twelve individuals, with flow cytometry again providing cell type proportion estimates for the same cell types. We obtained two scRNA-seq PBMC reference data sets. The first, which we used with the *Newman et al.* bulk, also comes from *Newman et al.* and assays one individual. To explore the effect of multiple individuals in the reference, we downloaded two reference sets from the public repository managed by 10x Genomics and subsequently merged them; this reference was used with the *Monaco et al.* bulk data samples. As neutrophils are notably difficult to assay accurately at the single-cell level, they were not present in either of the original reference panels. However, given the large fractions of neutrophils estimated by flow cytometry, particularly for the *Newman et al.* data set, we identified a publicly available data set which contains scRNA-seq data for human neutrophils [XSW+20]. These data were then incorporated into both reference sets in order to enable more effective comparisons. Because the *Newman et al.* scRNA-seq reference was relatively small (tens to hundreds of cells per cell type) and only had one individual present, we subsampled the neutrophil data down to 250 cells from one individual to be consistent with the other cell types. Conversely, because the 10x Genomics PBMC reference had more cells (hundreds to thousands of cells per cell type) and multiple individuals, we

Table 3.4: **Average L_1 errors with PBMC data and ground-truth cell proportions from flow cytometry.** The first two columns display average L_1 errors for the two PBMC data sets individually, while the last column aggregates L_1 errors across both data sets. Bisque and MuSiC do not provide proportion estimates for the Newman et al. data because only one individual is present for all reference cell types. CIBERSORTx was run in B-mode per their recommendation with a UMI-based scRNA-seq reference.

	Average L_1 Error (%)		
Method	Aggregate	Newman et al. Data	Monaco et al. Data
RNA-Sieve	**4.8**	4.8	**4.7**
DWLS	7.2	**4.7**	9.7
Scaden	9.4	11.3	7.6
CIBERSORTx	14.4	17.7	11.2
SCDC	19.3	17.2	21.3
NNLS	25.2	27.7	22.7
Bisque	n/a	n/a	17.3
MuSiC	n/a	n/a	22.7

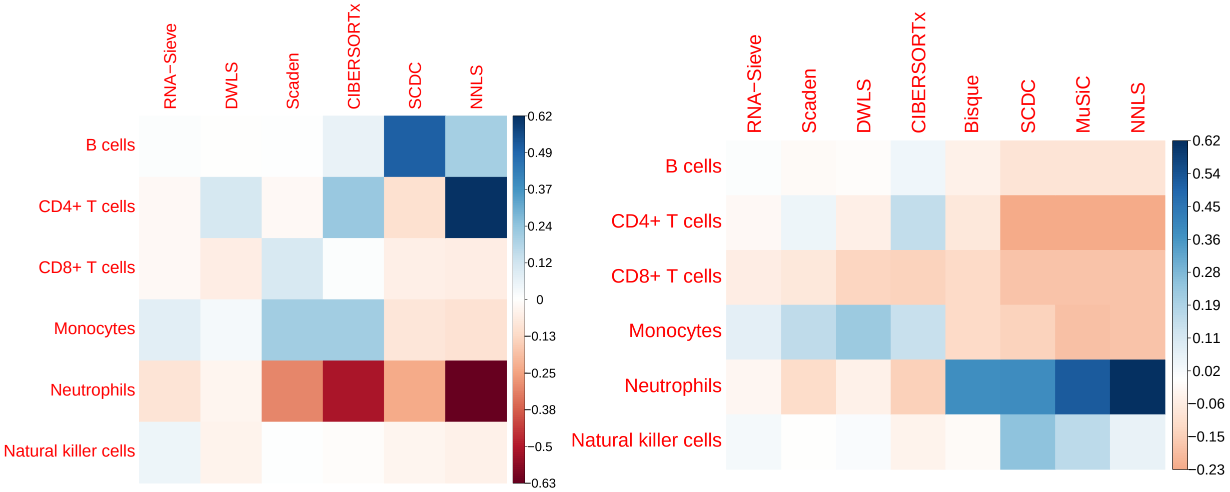
subsampled 1250 neutrophils in total from three individuals for use in the reference (see Section 18).

Subsequent deconvolutions showed that RNA-Sieve performed the best out of all methods as measured by the mean absolute deviation (L_1 error) when aggregating across both analyses (Table 3.4). The results are summarized graphically in Figure 3.5. The presence of neutrophils presented a challenge for several methods, perhaps due to the fact that they came from a different experiment or because of their uniquely low RNA counts. For example, in the bulk data from *Newman et al.*, neutrophils were strongly underestimated by CIBERSORTx, Scaden, and SCDC with most of that mass being allocated to either monocytes, CD4+ T cells, or B cells, respectively. RNA-Sieve and DWLS both performed well on these bulk samples, though RNA-Sieve slightly underestimated neutrophils in favor of monocytes while DWLS had minor difficulty distinguishing between CD4+ and CD8+ T cells. A similar story emerged for the *Monaco et al.* data, with CIBERSORTx, DWLS, Scaden, and, to a lesser extent, RNA-Sieve underestimating neutrophil and CD8+ T cell proportions while overweighting monocytes or CD4+ T cells. In contrast, Bisque, SCDC, and MuSiC strongly overweighted neutrophils (and sometimes natural killer cells) at the expense of other cell types. To produce a more formal and comprehensive comparison, we computed summary statistics in the same manner as Table 3.2 using the 24 bulk samples comprising the two data sets. RNA-Sieve achieves the best performance among all considered methods (Table 3.5) in each metric as it exhibits strong performance for both data sources. DWLS performs well on the *Newman et al.* data but fails to attain that level of accuracy on the *Monaco et al.* data.

Finally, we analyzed samples from the pancreatic islets region of the human pancreas where ground-truth proportions were not available. This region has previously been used for validation in the absence of ground-truth proportions because of prior knowledge of the general ranges of

Table 3.5: **Mean ranking of algorithms under various error metrics combined across the two PBMC deconvolutions.** All six applicable methods were ranked 1 (best) to 6 (worst) across 24 bulk samples from the Newman et al. and Monaco et al. data using the L_1, L_2, L_∞ and KL (KL-divergence) metrics, and their mean ranks were computed.

	RNA-Sieve	DWLS	Scaden	CIBERSORTx	SCDC	NNLS
L_1	**1.3**	2.2	2.7	4.1	4.8	6.0
L_2	**1.3**	2.2	2.7	4.1	4.8	6.0
L_∞	**1.2**	2.7	2.8	3.6	4.9	5.9
KL	**1.3**	2.5	2.4	4.0	5.0	5.9



(a) **Newman et al. Data**

(b) **Monaco et al. Data**

Figure 3.5: **Deconvolution biases for PBMC data with known ground-truth proportions.** Average differences between inferred and true proportions were computed within each cell type across the twelve bulk samples present in each scenario. Consistent overestimation of a cell type's abundance results in darker blue squares, while red corresponds to chronic underestimation. Methods are ordered left-to-right by overall performance.

constituent cell types. Moreover, the well-known negative relationship between beta cell proportions and hemoglobin A1c (Hb1Ac) levels allows us to test whether different deconvolution approaches can recapitulate this relationship. As shown in Figure B.6, RNA-Sieve is among the methods which successfully identify the expected negative correlation. As ground-truth values were not available for these data, it is impossible to ascertain precisely how methods performed, though it appears each method's average inferred beta cell proportions are below the expected $\sim 50\%$. Nevertheless, successful recovery of the expected association between beta cell proportions and Hb1Ac levels serves as a useful benchmark. Given the necessity to demonstrate robust performance across a range of tissues and cell type groups, we feel this result provides important support to RNA-Sieve's strong performance in the cell line and PBMC deconvolution tasks above.

Analysis of Real Bulk Organ Samples

We next applied RNA-Sieve to real bulk RNA-seq data to look for interesting patterns in organ composition. We chose to continue working with the *Tabula Muris Senis* data set as it contains many bulk RNA-seq samples in addition to the scRNA-seq data previously described. Due to its expansive experimental design across organs and ages, this resource is uniquely suited to interrogate changes in cell type compositions associated with the process of aging. By identifying changes in the balance of cell classes, we hope to provide insight into shifts in the mechanisms driving organ functions at different stages of life. In general, aging represents one of the more complicated biological processes, and one which occurs in every person or organism. Due to its ubiquity and significant effects on quality of life, improved understanding of the etiologies underlying age-associated functional deficits holds great potential therapeutic value. Degradation of the musculoskeletal and immune systems are among the most apparent phenotypic changes occurring during mammalian aging. Here we highlight intriguing results from three organs with roles in these bodily systems–the limb muscle in the former, and the spleen and bone marrow in the latter. In the absence of ground truth, we judge the reliability of our estimates by the relative consistency of inferred proportions within and across age groups.

Limb muscle in the arms and legs provides support and locomotion. It primarily consists of skeletal muscle, stromal, immune, and endothelial cells. Differentiated muscle cells contract to produce the aforementioned support and motion, while satellite cells serve in myogenesis and muscular repair [CCAK$^+$07]. Stromal cells comprise the connective organ which binds sarcomeres together and connects muscles to bones in addition to displaying certain regenerative capabilities [KAK$^+$19]. Immune response is often noted in muscle cells after mechanical stress-induced damage to muscle fibers, when the tissue becomes inflamed due to tearing [NWHK14, BKK$^+$13]. Endothelial cells are present due to the often high degree of vascularization needed to support muscle function [CCAK$^+$07]. Upon application of RNA-Sieve to the available bulk muscle samples, we observed a noticeable increase in skeletal muscle satellite cells and a substantial decrease in the mesenchymal stem cell proportion in older mice (Figure 3.6a). These trends are present, albeit fairly gradual, until around 21 months old with more sudden changes apparent thereafter. There was also an apparent increase in macrophage proportions up until 15 months of age, followed by a slow decline for the remainder of life. Each of these three cell types has been demonstrated to function in muscle fiber repair through different mechanisms [SNM$^+$15]. This pattern in cell type composition may thus indicate changes in the relative use of different regenerative pathways as individuals age.

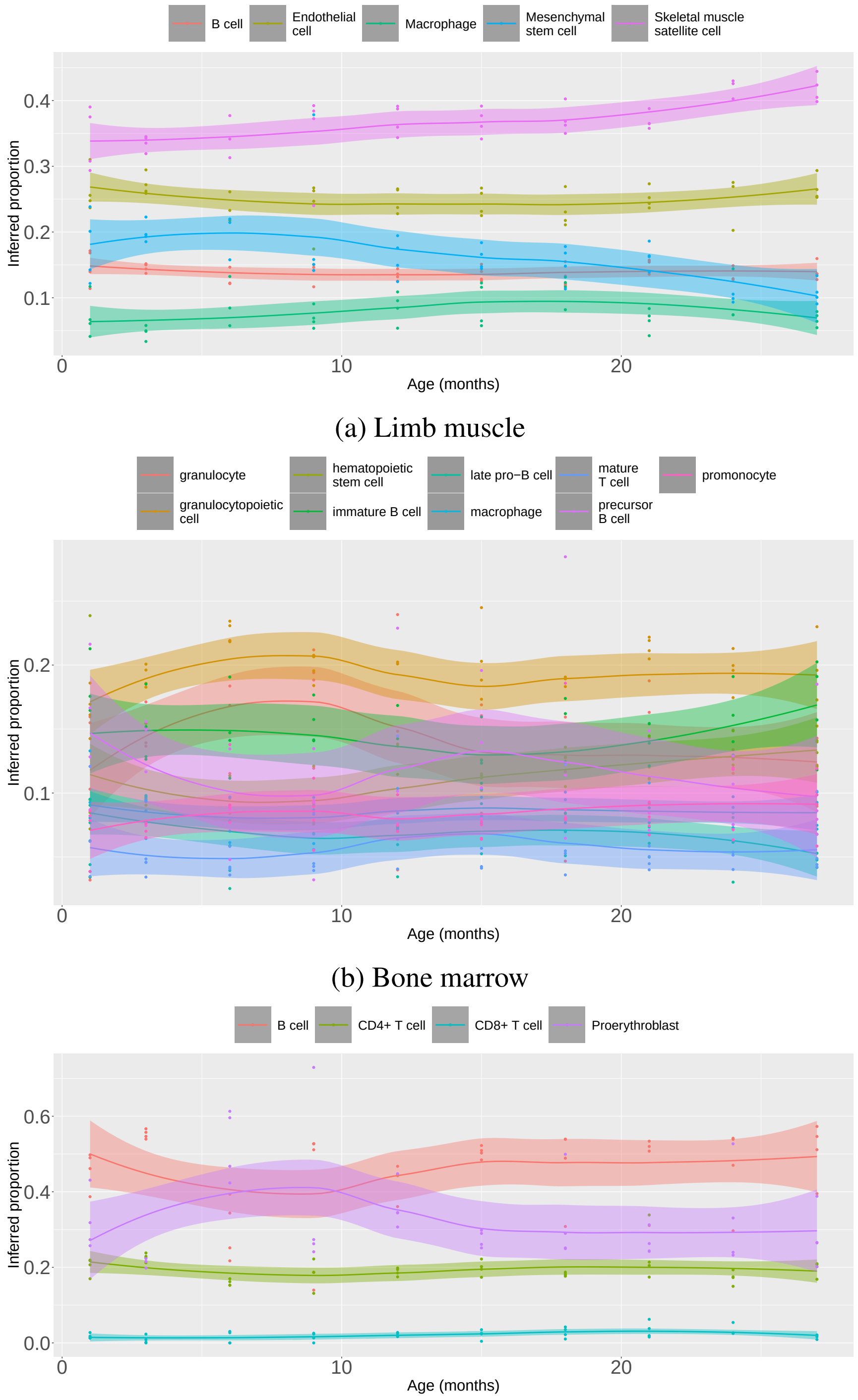
The bone marrow is a vital component of the immune system which executes the bulk of hematopoiesis and contains numerous constituent cell types ranging from various stem cells to more mature cell classes [GA08]. This rich combination yielded several age-associated trends in cell type composition, and we choose to focus on two. First, an effectively linear growth in the number of hematopoietic stem cells was observed with increasing age. Though this may seem surprising given reduced adaptive immunity with age, this exact phenomenon has been previously observed in both mice and humans [PPS+11], and it is accompanied by a decrease in functionality of these cells. Conversely, granulocyte proportion appeared to decrease after roughly 9 months of age. Further examination reveals that the granulocyte fraction tends to mirror that of granulocytopoietic cells, but with an increasing deficit between the two as age increases. Such a pattern is suggestive of the reduced potency of granulocytopoiesis that we would expect with age. Hence, in the marrow we are able to identify known patterns of cell type composition variation despite the presence of many transcriptomically similar cell types.

The spleen occupies a central role in the lymphatic system and is important, though not strictly essential, for proper functioning of the immune system and red blood cell recycling. This organ is split into two pieces—the red pulp, which contains blood cells, and the white pulp, which is primarily lymphatic [MK05]. Typically, the large majority of present immune cells are B and T cells, with smaller quantities of other cell types [HKP19]. Various progenitor cells may be present to spur production of immune and red blood cells, though these processes are primarily performed in the bone marrow [MVL+18]. It is notably difficult to distinguish among these progenitor cells in their early stages, making it possible that several varieties are captured within the single label of proerythroblasts. Upon deconvolution, we found that our inferred proportions for B and T cells matched accepted ranges [HKP19]. More interestingly, we noticed an unexpected and transient spike in the proportion of proerythroblasts peaking at roughly nine months of age (Figure 3.6c). Importantly, this increase is observed in all four of the 9-month-old individuals, and is thus not an artifact of outlying samples. Mice at this age are roughly analogous to humans of between 30-40 years of age, and as hematopoiesis is generally restricted to the marrow at this age except under stress conditions, it is unclear whether a programmed hematopoeitic process is occurring or if we are capturing the behavior of a cell type not enumerated in the reference set.

Extension to Confidence Regions

Within the deconvolution task, the generative framework of RNA-Sieve permits extensions which remain out of reach using prior approaches. One such possibility is the computation of confidence regions for inferred cell type proportions. Despite its clear importance, error quantification in deconvolution is challenging and has received relatively scant attention, leaving users to only guess at the reliability of their results. As deconvolution is sometimes performed upstream of tasks such as differential expression or eQTL detection, it is critical to know whether inferred proportions are precise. Because RNA-Sieve infers these proportions via maximum likelihood estimation, we can directly tap into the wide array of theory on asymptotic confidence bounds. Specifically, we construct confidence regions for inferred proportion values through numerical computation of the inverse empirical Fisher information matrix (see Section 18 in Section B.1). We demonstrate RNA-Sieve's ability to produce well-calibrated confidence regions by constructing them for within- and cross-protocol pseudobulk deconvolutions using *Tabula Muris Senis* data as well as both real PBMC bulk data sets in Section 3.2.

We began with within-protocol comparisons where all modeling assumptions are generically met. As shown in Figures 3.7a and B.7A, we obtain narrow, yet well-calibrated, confidence intervals, indicating the effectiveness of our procedure in this simplified scenario. However, the



(a) Limb muscle

(b) Bone marrow

(c) Spleen

Figure 3.6: **Real bulk deconvolutions from the *Tabula Muris Senis*.** ∼ 40 samples across ten ages were deconvolved using Smart-Seq2 references in the limb muscle, bone marrow, and spleen.

typical deconvolution setting will present complications in the form of protocol differences in the scRNA-seq reference and bulk RNA-seq data. Under mild and plausible assumptions on these distributional shifts, our MLE framework is robust to such model misspecification (see Section 18), and we still achieve good performance in spite of protocol mismatch (Figures 3.7b and B.7B). Aggregating across runs, our 95%-confidence intervals contain the true cell type proportions 96.7% and 91.8% of the time in the within- and across-protocol deconvolutions, respectively.

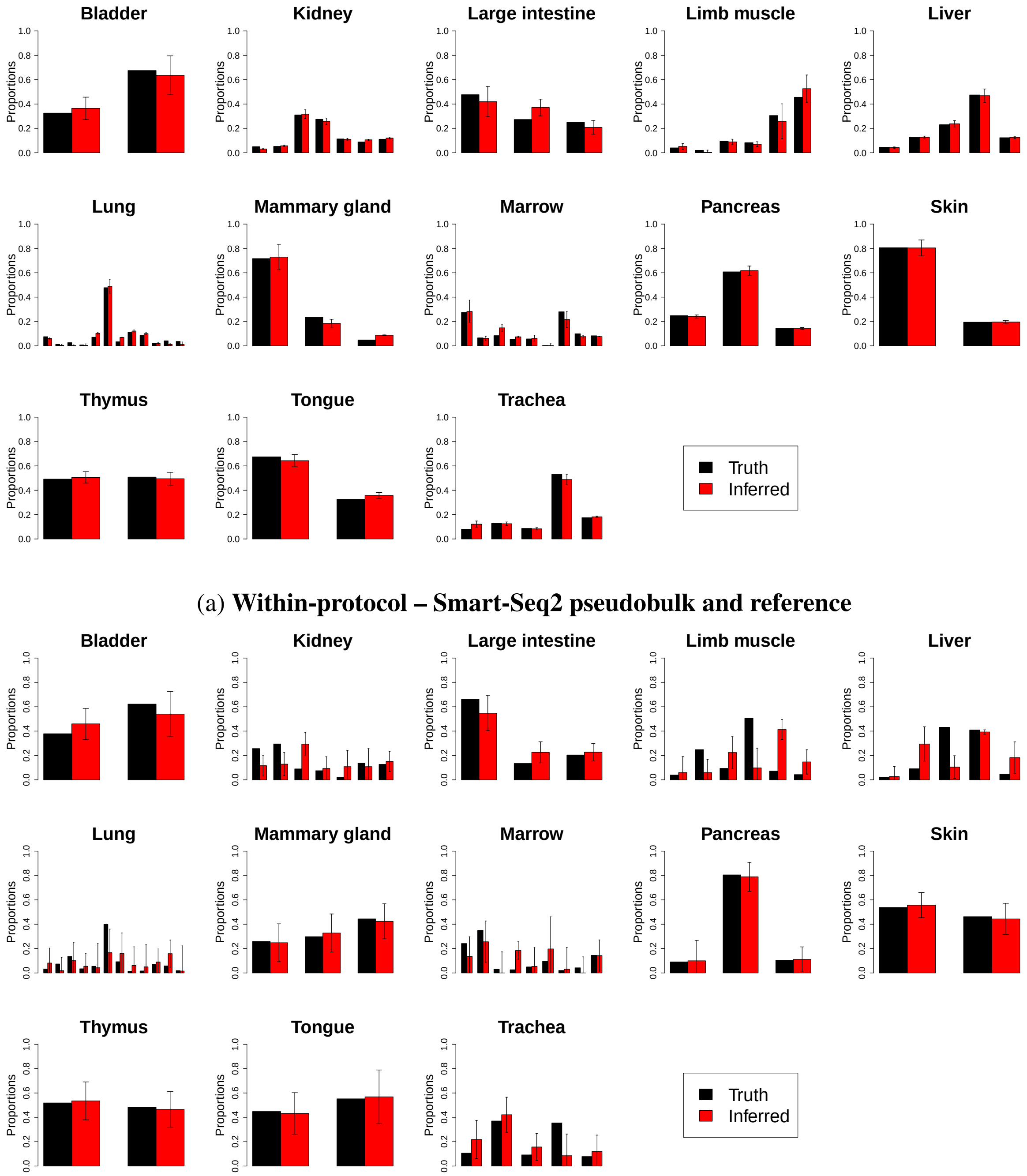
To ensure that we obtain sensible results with real bulk RNA-seq data, we also generated confidence intervals for the whole blood samples analyzed in Section 3.2. We again obtain calibrated and sensible results, with our confidence intervals containing the truth 95.8% of the time in the *Monaco et al.* bulk samples and 90.3% of the time in the *Newman et al.* bulk samples (Figure 3.8). Though assessing their accuracy is impossible absent ground-truth proportions, we also computed confidence intervals for the real bulks deconvolved in Section 3.2 to verify that RNA-Sieve's confidence intervals were reasonable in tissues besides whole blood. We found that these interval widths were similar to those we obtained in our other trials (Figure B.8). The distribution of confidence interval half-widths for cell type proportions were also generally consistent across samples (Figure B.9). We note that MuSiC presents a quantity which seemingly corresponds to the variance in proportion estimates, though it was not emphasized in their manuscript, and we generally found the produced values to be overly small in practice.

In principle, the widths of confidence intervals should depend on the number of cells and genes in the reference, similarity among cell types in the reference, and agreement between the reference and bulk measurements. Our empirical results suggest that these eminently logical factors do, in fact, drive the widths of our intervals. For example, the confidence intervals in cross-protocol deconvolutions are a bit wider than their within-protocol counterparts, due to our adaptive procedure's conservative nature when it detects differences between the reference and bulk. This arises in part because we deem fewer genes reliable when compared to within-protocol experiments. Moreover, evidence of the contribution of reference sample size is present in a few organs, most notably the lung with its many low-frequency cell types.

3.3 Discussion

Here we have introduced our method for supervised bulk gene expression deconvolution, RNA-Sieve, and illustrated its robust performance in a variety of settings. Unlike methods which rely on variants of least squares or the application of complex machine learning algorithms, we place the deconvolution problem into a generative probabilistic framework that models random noise in both the reference panel and bulk samples by relying on asymptotic theory. Through simulations and applications to real data, we demonstrated the broad applicability of our method and its utility to investigate biological questions of interest.

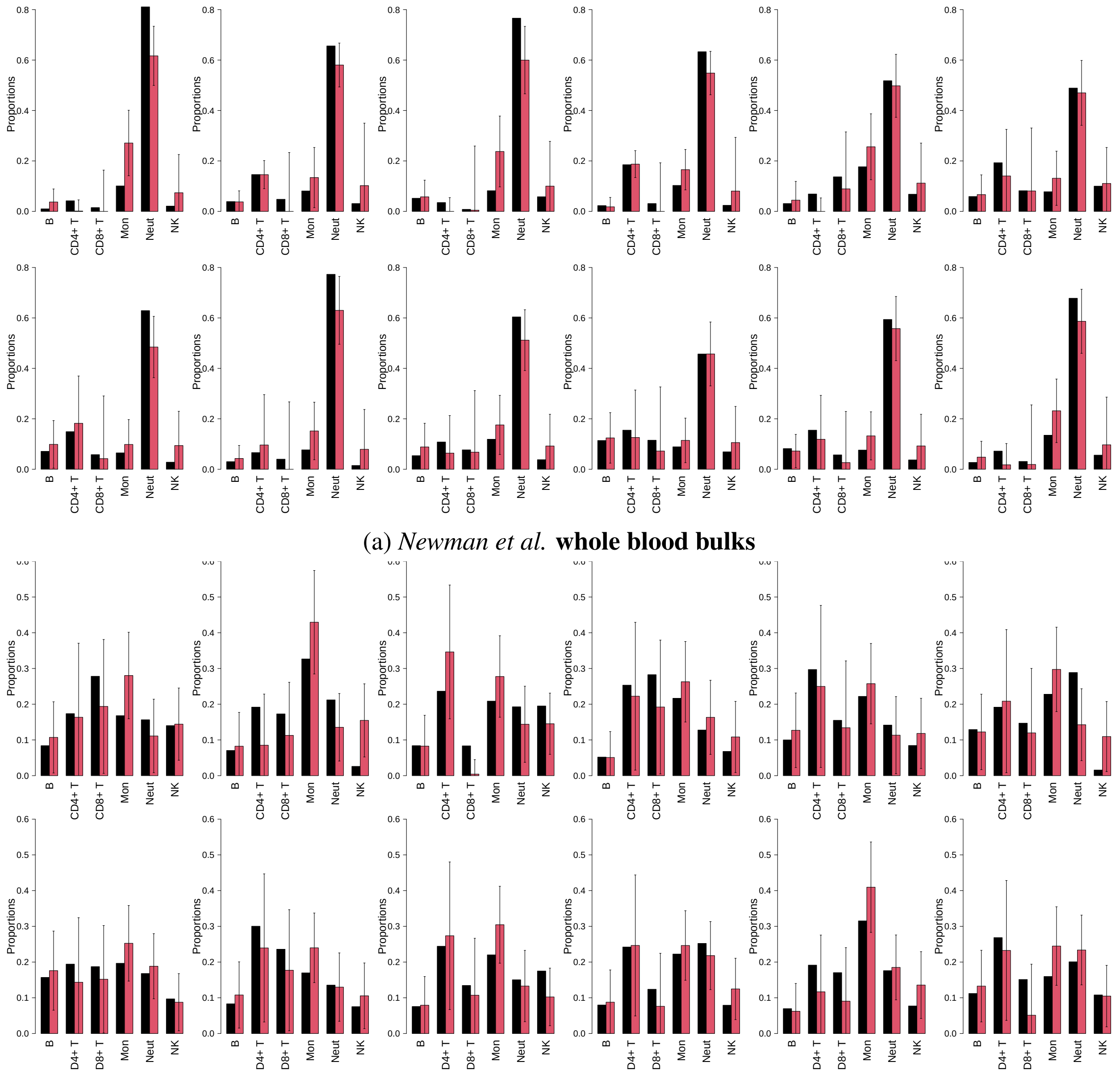
It is valuable to understand how RNA-Sieve differs from other approaches and to consider the consequences of these divergent design choices. Least squares-based solutions such as MuSiC [WPS+19], SCDC [DTU+20], and DWLS [TDC+19] devise their own implementations of weighted non-negative least squares (W-NNLS). These methods aim to handle heteroskedasticity



(a) **Within-protocol – Smart-Seq2 pseudobulk and reference**

(b) **Cross-protocol – Smart-Seq2 reference and 10x Chromium pseudobulk**

Figure 3.7: **Deconvolution results for pseudobulk experiments.** *Tabula Muris Senis* were used as described in Section 3.2. Error bars represent marginal 95% confidence intervals as described in Section 18 For cell type specification, see Table B.1



(a) *Newman et al.* **whole blood bulks**

(b) *Monaco et al.* **whole blood bulks**

Figure 3.8: **Deconvolution results for real bulk samples with known ground truth.** PBMC references and whole blood bulks were used as described in Section 3.2. True proportions estimated by flow cytometry are in black while RNA-Sieve estimates are in red, with 95% confidences indicated.

across genes by re-weighting them according to their variability and specificity, allowing genes which are putatively more informative to carry increased importance in the regression task. Alternatively, Bisque [JAR+20] uses NNLS after applying a transformation to bring the reference and bulk data into better distributional agreement. From a modeling perspective, least squares-based solutions generally address uncertainty in the bulk only, leaving stochasticity in the single-cell reference unaccounted for. With RNA-Sieve, rather than devising a specialized gene-weighting scheme, we naturally emphasize some genes more than others via variances resulting from an explicit generative model incorporating noise in both single cells and the bulk. We also do not explicitly attempt to bring reference and bulk data into better agreement a là Bisque, instead relying on a filtering protocol to remove genes which display signs of significant deviation from our assumptions. Integrating an explicit transformation is an interesting possibility for RNA-Sieve, and should only boost its performance by further aligning data to our model assumptions. Other methods employ machine learning techniques, such as CIBERSORTx [NSL+19], which uses ν-support vector regression, and Scaden [MMO+20], which utilizes deep neural networks. These approaches can be opaque to the user due to their reliance on high-complexity algorithms which often lack theoretical guarantees of optimality and provably accurate inference despite continuing advances in explainability techniques. Comparatively, our formulation of RNA-Sieve as the MLE of an explicit generative model is transparent in both parameter interpretation and performance guarantees. The parameters updated during optimization have explicit biological meanings and tracing their values allows for a deeper interrogation of the predictions RNA-Sieve generates. This is a useful feature when providing context to inferred cell type proportions as well as exploring the theoretical limits of deconvolution as a function of cell type properties. Like MuSiC, SCDC, Bisque, and Scaden, we do not select marker genes in RNA-Sieve. This helps us maintain computational efficiency, while simultaneously providing robustness with respect to outlier fluctuations in gene expression. We also parallelize our optimization steps and jointly update parameters when deconvolving multiple bulk samples. This yields significant speedups relative to serial runs and allows us to share statistical strength across all bulks.

RNA-Sieve is embedded in a flexible generative framework, which can be adapted to a variety of situations to make deconvolution performance more effective. One of these is the modeling of further sources of variation. For instance, if gene expression distributions are expected to differ drastically across individuals from which samples are taken, this knowledge can be explicitly incorporated into our likelihood. Without such modification, RNA-Sieve implicitly follows the paradigm of MuSiC, SCDC, and Bisque in penalizing genes of large inter-individual variance via the marginal variances resulting from estimation of the reference panel. A similar notion applies to mitigating potential batch effects or effectively combining disparate references. Currently, different reference matrices which are believed to have the same expression distributions can be averaged together to increase statistical power without further modification of our present implementation.

A principal motivation of this work was to expand the scope of accessible questions in the deconvolution setting. Our likelihood-based approach facilitates extensions which are intractable with current algorithms. As a first step, we have chosen to demonstrate our ability to explicitly construct confidence regions for inferred proportions, producing a mathematically rigorous quantification of the uncertainty in our estimates. The necessity of these bounds is plainly substantiated by the use of deconvolution upstream of tasks ranging from cell type-specific differential expression to eQTL detection using heterogeneous RNA-seq organ samples. The credibility of any such analyses is predicated on the accuracy of deconvolution, because any errors in this initial step will propagate through to the final result. Consequently, we anticipate that our confidence regions will encourage improved assessment of the reliability of results obtained through these types of analyses. Our confidence intervals are also of obvious inherent value when using deconvolution results to infer differences in cell type composition between samples, whether due to disease status or other factors. Beyond error quantification via confidence intervals, potent possibilities lie in hypothesis testing. Currently, CIBERSORTx does propose one type of test, though our understanding is that it tests whether *any* of the bulk cell types were found in the reference. This is rather restrictive, so we hope to develop procedures with broader utility. One example with clinical impact is a test to determine whether the reference panel is missing cell types present in the bulk sample. Even though we have demonstrated that RNA-Sieve is robust with respect to such misspecification (see Section 3.2), it is nonetheless beneficial to know whether the deconvolution performed was sufficiently valid using a principled approach. Such a test can be directly developed in our framework by examining the residuals produced by our maximum likelihood estimate, and work in this direction is underway.

Despite the flurry of recently developed methods, the question of statistical deconvolution of gene expression data remains far from solved. RNA-Sieve illustrates the efficacy, adaptability as well as promise of generative modeling in this setting, and we hope it spurs continued development within other methodological paradigms. In particular, notions of error quantification and hypothesis testing merit further attention.

Chapter 4

Nonparametric Goodness-of-Fit Testing

This chapter is joint work with Jonathan Terhorst and Yun S. Song. It is available as an *arXiv* preprint [EPTS20].

4.1 Introduction

In an address to the Royal Statistical Society, epidemiologist and statistician [Gre46] introduced what became known as Greenwood's statistic in order to study the randomness of infection times $T_1, \dots, T_k$ of a given number k of patients. More precisely, he was interested in testing the hypothesis of whether these infection times were generated by a homogeneous Poisson process; that is, whether $(T_1, T_2, \dots, T_k)/(\sum_{j=1}^k T_j)$ is distributed uniformly on the $(k-1)$-dimensional simplex. To do so, he proposed a test based on the null distribution of the statistic $(\sum_{j=1}^k T_j^2)/(\sum_{j=1}^k T_j)^2$, for which he was able to provide a complete description for $k = 2$, but none for $k > 2$. This sparked a flurry of studies attempting both to better understand Greenwood's statistic in higher dimensions, as well as to clarify its power and efficiency as a hypothesis test. In a series of papers, [Mor47, Mor51, Mor53] computed the test statistic's first four moments for an arbitrary k, proved a central limit theorem as $k \to \infty$ (albeit with very slow convergence), and estimated the test's efficiency against a specific class of alternatives. Around the same time, [Gar52] fully described the case $k = 3$ by fruitfully interpreting the distribution function of the test statistic as the volume of intersection between a sphere and the simplex. Subsequently, [Dar53] found a closed-form, but difficult to invert, characteristic function, while [Wei56] and [SR70] investigated the role of Greenwood's statistic in the context of the general goodness-of-fit test of whether a sample $X_1, \dots, X_n$ follows a given arbitrary distribution F. There they proved that a test based on Greenwood's statistic is both most powerful against symmetric linear alternatives, and enjoys the greatest asymptotic relative efficiency against a wide class of tests and alternative hypotheses. Given these favorable properties, but analytically intractable nature of Greenwood's statistic, approximations and numerical schemes were devised by [Bur79], [Cur81] and [Ste81] to tabulate scores for certain significance levels up to $k = 20$. Most recently, [SZ00] exploited the geometric interpretation of [Gar52] to derive an explicit rate function to characterize the large deviations of Greenwood's statistic as $k \to \infty$.

Closely connected to Greenwood's statistic, and applicable in an equal variety of settings, is a discretized version of it: instead of sampling uniformly from the simplex Δ^{k-1}, the null distribution may be uniform over all integral points in the scaled simplex $n \cdot \Delta^{k-1}$ for some $n \in \mathbb{Z}^+$. Tests and computations based on such measure occur in various contexts in computational biology (e.g., [PTSP18, RCK07]), physics (where it is known as the Bose-Einstein distribution) and theoretical statistics, where it emerges naturally as an urn model [Hol79]. Greenwood's statistic has a natural analogue in this discretized scenario, where it is known as Dixon's statistic after its proposed use by [Dix40] for performing non-parametric two-sample testing. However, although its asymptotic behavior has been well studied by [HR80], a description in the non-asymptotic regime as well as convergence rates have, to the best of our knowledge, remained elusive.

Here we fill this gap by studying a generalized family of Greenwood's statistics (including Dixon's statistic) for finite n and k, for which we are able to exactly and efficiently compute its moments. Using this knowledge, we examine various scaling limits, proving CLT results as well as identifying novel limiting distributions. We then quantify the connection between Dixon's and Greenwood's statistic through precise convergence rates and monotonicity results, while using our understanding from the discrete setting to offer new insights into the moment sequence, smoothness and monotonicity of Greenwood's statistic. Finally, we propose a simple and efficient algorithm that recovers an underlying continuous distribution from its first m moments up to $O(m^{-1})$ accuracy, and use it to devise a powerful hypothesis test of whether two data sets $\{X_1,\ldots,X_{k-1}\}$ and $\{Y_1,\ldots,Y_n\}$ were sampled from the same distribution. We demonstrate the test's suitability for a large class of alternatives through extensive power studies, and compare it with the classical Kolmogorov-Smirnov, Cramér-von Mises and Mann-Whitney tests. We illustrate how the same principles employed in designing our two-sample test can be applied equally successfully to the settings of one-sample tests, and tests using paired data.

4.2 Preliminaries and notation

For a positive integer n, we use $[n]$ to denote the set $\{1,\ldots,n\}$. The two probability spaces underlying most of our discussion will consist of the $(k-1)$-dimensional probability simplex

$$\Delta^{k-1} = \left\{ (x_1, x_2, \ldots, x_k) \in [0,1]^k \ : \ \sum_{i=1}^{k} x_i = 1 \right\}$$

together with the uniform measure $\mu_{\Delta^{k-1}} = \sigma / \sigma(\Delta^{k-1}) = (k-1)!\sigma/\sqrt{k}$, where σ is surface measure in $\mathbb{R}^k$, and its discretized version

$$D_{n,k} = (n\Delta^{k-1}) \cap \mathbb{Z}^k = \left\{ (z_1, z_2, \ldots, z_k) \in \{0,\ldots,n\}^k \ : \ \sum_{i=1}^{k} z_i = n \right\}$$

with its uniform measure $\mu_{D_{n,k}} = |D_{n,k}|^{-1} = \binom{n+k-1}{k-1}^{-1}$. In other words, $\mu_{\Delta^{k-1}}$ is the law of a Dirichlet$(1,\ldots,1)$ variable, while a k-part weak composition of n chosen uniformly at random is distributed according to $\mu_{D_{n,k}}$. Occasionally we will refer to this latter distribution as a uniform configuration of n balls distributed over k bins. To test the hypothesis of whether a sampled random variable $\boldsymbol{X} = (X_1, \ldots, X_k)$ in $\mathbb{Z}^k$ or $\mathbb{R}^k$ has distribution $\mu_{\Delta^{k-1}}$ or $\mu_{D_{n,k}}$, respectively, we are interested in comparing the distribution of its weighted ℓ_p-norms

$$\|\boldsymbol{X}\|_{p,\boldsymbol{w}}^p = \sum_{i=1}^{k} w_i X_i^p ,$$

for some fixed weight vector $\boldsymbol{w} = (w_1, \ldots, w_k) \in \mathbb{R}^k$, against its null distribution. That is, if $\boldsymbol{S}_k \sim \mu_{\Delta^{k-1}}$ and $\boldsymbol{S}_{n,k} \sim \mu_{D_{n,k}}$, we are interested in studying the distributions of $\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^p$ and $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ (to prevent confusion of powers and vector indices, we will denote entries of $\boldsymbol{S}_k$ and $\boldsymbol{S}_{n,k}$ by $S_k[\![j]\!]$ and $S_{n,k}[\![j]\!]$ for $j \in [k]$). For $p = 2$ and $\boldsymbol{w} = \boldsymbol{1}_k$, with $\boldsymbol{1}_k := (1, \ldots, 1)$ being the all-ones vector of length k, these are precisely Greenwood's and Dixon's statistics, respectively.

A primary reason why understanding these statistics is important is their application to nonparametric testing. If $Z_1, \ldots, Z_N$ are independently sampled from the same continuous distribution F, then the spacings (that is, the differences between consecutive order statistics) of $0, F(Z_{(1)}), \ldots, F(Z_{(N)}), 1$ are distributed according to μ_{Δ^N}, so characterizing the distribution of Greenwood's statistic allows to test if the sample is distributed according to F. Similarly, if $X_1, \ldots, X_{k-1}, Y_1, \ldots, Y_n$ are i.i.d. samples from the same continuous distribution G, and we define the order statistics

$$-\infty =: X_{(0)} < X_{(1)} < \cdots < X_{(k-1)} < X_{(k)} := \infty,$$

and the number of Y_i sandwiched between every two consecutive order statistics

$$S_{n,k}[\![j]\!] := \#\left\{i \,:\, X_{(j-1)} \leq Y_i < X_{(j)}\right\}, \tag{4.1}$$

for $j \in [k]$, then $(S_{n,k}[\![1]\!], \ldots, S_{n,k}[\![k]\!])$ is distributed according to $\mu_{D_{n,k}}$.

These two hypothesis tests will be our main application considered in Section 4.5. Before coming to those, in Section 4.3 we present our main theoretical results about $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ and $\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^p$. In particular, we will detail a simple and efficient way to compute their moment sequences exactly. Recovering the distribution of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ and $\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^p$ from these moments seems analytically intractable, but can be done algorithmically in an efficient and exact manner. Section 4.4 is dedicated to describing such an algorithm.

4.3 Moments and scaling limits of generalized spacing-statistics

We start by investigating the discrete generalized spacing-statistics $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$, from which their continuous analogues will follow. Our point of departure is the well-known Wilcoxon-Mann-Whitney U statistic [MW47, Wil45]. It is easy to see that in the special case of $p = 1$ and $\boldsymbol{w} = (k-1){\downarrow} := (k-1, k-2, \ldots, 1, 0)$, we recover, up to an explicit constant depending only on k, the U statistic. Indeed, for two samples $X_1, \ldots, X_{k-1}$ and $Y_1, \ldots, Y_n$, with $R_1, \ldots, R_{k-1}$ the ranks of

$X_1,\ldots,X_{k-1}$ computed in the joint ensemble $\{X_1,\ldots,X_{k-1},Y_1,\ldots,Y_n\}$, their U statistic is given by $U=\sum_{j=1}^{k-1}R_j$. In our notation introduced in (4.1), we then have $R_j=j+\sum_{i=1}^{j}S_{n,k}[\![i]\!]$, and consequently

$$\begin{aligned}U=\sum_{j=1}^{k-1}R_j=\sum_{j=1}^{k-1}\left[j+\sum_{i=1}^{j}S_{n,k}[\![i]\!]\right]&=\sum_{j=1}^{k-1}j+\sum_{i=1}^{k-1}\sum_{j=i}^{k-1}S_{n,k}[\![i]\!]\\&=\binom{k-1}{2}+\sum_{i=1}^{k-1}(k-i)S_{n,k}[\![i]\!]\\&=\binom{k-1}{2}+\|\boldsymbol{S}_{n,k}\|^1_{1,(k-1)\downarrow},\end{aligned}$$

as desired. In order to both compute the exact distribution of U under the null hypothesis of all X_i,Y_i being generated i.i.d, as well as its asymptotic normality, [MW47] exploited a recurrence relation which, in our language, consists of conditioning on the occupancy of the very last entry in $\boldsymbol{S}_{n,k}$. Conditional on the event $\{S_{n,k}[\![k]\!]>0\}$—that is, conditional on the last bin containing at least one ball—we may remove one such ball to find that $\left(\|\boldsymbol{S}_{n,k}\|^1_{1,(k-1)\downarrow}\mid\{S_{n,k}[\![k]\!]>0\}\right)\overset{d}{=}\|\boldsymbol{S}_{n-1,k}\|^1_{1,(k-1)\downarrow}$. Similarly, on $\{S_{n,k}[\![k]\!]=0\}$ we may omit the very last bin to arrive at $\left(\|\boldsymbol{S}_{n,k}\|^1_{1,(k-1)\downarrow}\mid\{S_{n,k}[\![k]\!]=0\}\right)\overset{d}{=}\|\boldsymbol{S}_{n,k-1}\|^1_{1,(k-2)\downarrow}+n$. Combining these two observations, and writing $q_{n,k}(x)=\mathbb{P}\left(\|\boldsymbol{S}_{n,k}\|^1_{1,(k-1)\downarrow}=x\right)$, yields the two-term recursion

$$q_{n,k}(x)=\frac{n}{n+k-1}q_{n-1,k}(x)+\frac{k-1}{n+k-1}q_{n,k-1}(x-n),\tag{4.2}$$

from which the whole probability mass function of $\|\boldsymbol{S}_{n,k}\|^1_{1,(k-1)\downarrow}$ can be computed in $O(n^2k^2)$ time. Moreover, this recursion directly translates into a recurrence of the moments of $\|\boldsymbol{S}_{n,k}\|^1_{1,(k-1)\downarrow}$, which allowed [MW47] to prove a central limit theorem as $n,k\to\infty$, thus rendering U a versatile and quickly computable two-sample test statistic.

Unfortunately, the recurrence relation (4.2) lacks robustness with respect to varying either p or $\boldsymbol{w}=(w_1,\ldots,w_k)$: for $p>1$ and generic $\boldsymbol{w}$, removing a ball from any bin changes $\|\boldsymbol{S}_{n,k}\|^p_{p,\boldsymbol{w}}$ by an amount depending on the total number of balls in that bin, and removing a bin may result in weights that are in no way related to the original weights. This can be fixed by conditioning not only on the vacancy of $\boldsymbol{S}_{n,k}[\![k]\!]$, but its precise occupancy. Defining $q^{p,\boldsymbol{w}}_{n,k}(x):=\mathbb{P}\left(\|\boldsymbol{S}_{n,k}\|^p_{p,\boldsymbol{w}}=x\right)$ and $\boldsymbol{w}_{-k}:=(w_1,\ldots,w_{k-1})$, we have

$$\begin{aligned}q^{p,\boldsymbol{w}}_{n,k}(x)&=\sum_{j=0}^{n}\mathbb{P}\left(S_{n,k}[\![k]\!]=j\right)q^{p,\boldsymbol{w}_{-k}}_{n-j,k-1}\left(x-w_kj^p\right)\\&=\binom{n+k-1}{k-1}^{-1}\sum_{j=0}^{n}\binom{n-j+k-2}{k-2}q^{p,\boldsymbol{w}_{-k}}_{n-j,k-1}\left(x-w_kj^p\right),\end{aligned}\tag{4.3}$$

with initial conditions

$$q_{n,k}^{p,\boldsymbol{w}}(x) = \begin{cases} 0, & \text{if } x \notin \{x_{\min}, x_{\max}\} \text{ or } n < 0, \\ \mathbb{1}_{x=w_1 n^p}, & \text{if } k = 1, \\ \mathbb{1}_{x=0}, & \text{if } n = 0, \end{cases} \tag{4.4}$$

where $x_{\min} = \min_{\boldsymbol{S}_{n,k} \in D_{n,k}} \|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ and $x_{\max} = \|\boldsymbol{w}\|_\infty n^p$ bound the range of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$. However, solving (4.3) and (4.4) generically requires $O\left(n^{p+1} k \|\boldsymbol{w}\|_\infty\right)$ time, thus rendering it unfeasible to compute for even modestly large n and p. To circumvent this problem, in what follows we will instead use a moment-based approach.

Exact moments of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ and $\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^p$

The following theorem, together with Proposition 6 in Section 4.4, demonstrates how the aforementioned computational intractability can be circumvented by lifting (4.3) to the moments of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$:

Theorem 1. *Let $G(x,y) = \sum_{m=0}^\infty \mathrm{Li}_{-pm}(x) y^m / m!$, where $\mathrm{Li}_s(x) = \sum_{j=1}^\infty j^{-s} x^j$ is the polylogarithm function. Denoting by $[x^n y^m] P(x,y)$ the $(n,m)^{th}$ coefficient of a power series P in x and y, we have*

$$\mathbb{E}\left(\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p\right)^m = \frac{m!}{\binom{n+k-1}{k-1}} [x^n y^m] \prod_{i=1}^k G(x, w_i y). \tag{4.5}$$

In particular, the first m moments of $\|S_{i,j}\|_{p,\boldsymbol{w}}^p$ for $(i,j) \in \{0,\dots,n\} \times \{1,\dots,k\}$ can be computed in $O(nm \cdot (\log nm) \cdot (\log k))$ time.

Proof. We first expand the left-hand side of (4.5) to find

$$\begin{aligned} \mathbb{E}\left(\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p\right)^m &= \sum_{\boldsymbol{\sigma} \in D_{n,k}} \mathbb{P}(\boldsymbol{S}_{n,k} = \boldsymbol{\sigma}) \left(\sum_{j=1}^k w_j \sigma_j^p\right)^m \\ &= \binom{n+k-1}{k-1}^{-1} \sum_{\boldsymbol{\sigma} \in D_{n,k}} \sum_{\boldsymbol{\eta} \in D_{m,k}} \binom{m}{\eta_1, \dots, \eta_k} \prod_{j=1}^k w_j^{\eta_j} \sigma_j^{\eta_j p} \\ &= \frac{m!}{\binom{n+k-1}{k-1}} \underbrace{\sum_{\boldsymbol{\eta} \in D_{m,k}} \left(\sum_{\boldsymbol{\sigma} \in D_{n,k}} \prod_{j=1}^k \frac{(w_j \sigma_j^p)^{\eta_j}}{\eta_j!}\right)}_{A_{n,k,m,w}}, \end{aligned} \tag{4.6}$$

so it remains to show that $A_{n,k,m,w} = [x^n y^m] \prod_{j=1}^k G(x, w_j y)$. By definition of $\mathrm{Li}_x(x)$, we have for every fixed $\boldsymbol{\eta} \in D_{m,k}$

$$\sum_{\boldsymbol{\sigma} \in D_{n,k}} \prod_{j=1}^k \frac{w_j^{\eta_j} \sigma_j^{p\eta_j}}{\eta_j!} = [x^n] \prod_{j=1}^k \frac{\mathrm{Li}_{-p\eta_j}(x)}{\eta_j!} w_j^{\eta_j}, \tag{4.7}$$

and so

$$\begin{aligned} A_{n,k,m,w} &= [x^n] \sum_{\boldsymbol{\eta} \in D_{m,k}} \prod_{j=1}^{k} \frac{\mathrm{Li}_{-p\eta_j}(x)}{\eta_j!} w_j^{\eta_j} \\ &= [x^n] \left\{ [y^m] \prod_{j=1}^{k} \left(\sum_{i=0}^{\infty} \frac{\mathrm{Li}_{-pi}(x)}{i!} (w_j y)^i \right) \right\} \\ &= [x^n y^m] \prod_{j=1}^{k} G(x, w_j y), \end{aligned} \tag{4.8}$$

as desired. The $O(nm \cdot (\log nm) \cdot (\log k))$ runtime is now a direct consequence of computing the Cauchy product of k bivariate degree-(n,m) polynomials using the Fast Fourier Transform. □

The above theorem will be our main tool for devising an efficient, powerful, and general two-sample test in Section 4.5, if the two samples are modestly sized. In the case where one sample is significantly larger than the other, the following result, paired with Proposition 6 below, yields an even more efficient testing procedure.

Proposition 1. *For fixed k, $\boldsymbol{S}_{n,k}/n$ converges in distribution to $\boldsymbol{S}_k \sim \mu_{\Delta^{k-1}}$. In particular, we have*

$$\left\| \frac{\boldsymbol{S}_{n,k}}{n} \right\|_{p,\boldsymbol{w}}^{p} \xrightarrow{d} \|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^{p} \quad \text{as } n \to \infty. \tag{4.9}$$

Proof. It suffices to show that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\boldsymbol{S}_{n,k}}{n} \in E \right) = \mathbb{P}\left(\boldsymbol{S}_k \in E\right) \tag{4.10}$$

for any Lebesgue-measurable set $E \subset \Delta^{k-1}$. Moreover, by Dynkin's π-λ theorem, we may without loss of generality assume E to be a box with rational vertex points inside Δ^{k-1} (see e.g. Theorem 1.1 in [Kal06]), which will allow us to count the number of lattice points in E as $n \to \infty$. To wit, let $L(E,n)$ be the cardinality of the set $nE \cap \mathbb{Z}^k$, then Ehrhart theory informs us that (cf. [Ehr67])

$$L(E,n) = \mathrm{Vol}_\Lambda(E) \cdot n^{k-1} + O(n^{k-2}), \tag{4.11}$$

where $\mathrm{Vol}_\Lambda(E) = \lambda_{k-1}(E)/\mathrm{d}(\Lambda)$ is the $(k-1)$-dimensional Lebesgue volume of E, normalized by the co-volume of the lattice $\Lambda = \mathbb{Z}^k \cap H$ induced by $\mathbb{Z}^k$ on the hyperplane $H = \{\boldsymbol{x} \in \mathbb{R}^k \,:\, \sum_{j=1}^{k} x_j = 0\}$. But since the fundamental region of Λ is the parallelepiped formed by $\{\boldsymbol{e}_1 - \boldsymbol{e}_j\}_{j\in\{2,\ldots,k\}}$, where $\boldsymbol{e}_j \in \mathbb{R}^k$ is the j^{th} standard basis vector, the co-volume can be computed to be

$$\mathrm{d}(\Lambda)^2 = \det V^t V = \det\left(I_{k-1} + \mathbf{1}_{k-1}\mathbf{1}_{k-1}^T\right), \tag{4.12}$$

with the columns of V being given by $\{\boldsymbol{e}_1 - \boldsymbol{e}_j\}_{j\in\{2,\ldots,k\}}$. It is straightforward to check that $V^t V$ has eigenvalue 1 of multiplicity $k-2$ (with the associated eigenspace spanned by $\{\boldsymbol{e}_1 - \boldsymbol{e}_j\}_{j\in\{2,\ldots,k-1\}}$), and eigenvalue k of multiplicity 1 (with eigenvector $\mathbf{1}_{k-1} = (1,\ldots,1)$), and therefore

$$\mathrm{d}(\Lambda) = \sqrt{k}. \tag{4.13}$$

Since $\sqrt{k}/(k-1)!$ is precisely the $(k-1)$-dimensional volume of Δ^{k-1}, we finally arrive at

$$\begin{aligned}\mathbb{P}\left(\frac{\boldsymbol{S}_{n,k}}{n}\in E\right) &= \frac{L(E,n)}{\binom{n+k-1}{k-1}} = \frac{L(E,n)}{\frac{n^{k-1}}{(k-1)!}+O(n^{k-2})} = (k-1)!\mathrm{Vol}_{\Lambda}(E)+O\left(n^{-1}\right)\\ &\xrightarrow{n\to\infty} \frac{(k-1)!}{\sqrt{k}}\lambda_{k-1}(E) = \mathbb{P}\left(\boldsymbol{S}_k\in E\right), \qquad (4.14)\end{aligned}$$

which proves the first part of our proposition. The result in (4.9) is now a direct consequence of the continuous mapping theorem. □

We note in particular that for $p=2$ and $\boldsymbol{w}=\boldsymbol{1}_k=(1,\dots,1)$, the limiting random variable in (4.9) is precisely Greenwood's original test statistic. It is in this sense that we consider $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ generalized Greenwood statistics. In order for this connection between two-sample testing and tests of uniformity to be of any use, a clear understanding of both the limiting distributions, as well as the convergence rates is necessary. We begin with the former, for which reasoning akin to Theorem 1 is available:

Theorem 2. *Let $Q_p(x)=\sum_{m=0}^{\infty}(pm)!x^m/m!$. Then,*

$$\mathbb{E}\left(\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^p\right)^m = \frac{(k-1)!m!}{(pm+k-1)!}[x^m]\prod_{j=1}^{k}Q_p(w_jx). \qquad (4.15)$$

In particular, the first m moments of $\|S_j\|_{p,\boldsymbol{w}}^p$ for $j\in\{1,\dots,k\}$ can be computed in $O(m\cdot(\log m)\cdot(\log k))$ time.

Proof. As in (4.6), we expand the left-hand side of (4.15) to obtain

$$\begin{aligned}\mathbb{E}\left(\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^p\right)^m &= \int_{\Delta^{k-1}}\left(\|\boldsymbol{x}\|_{p,\boldsymbol{w}}^p\right)^m \,\mathrm{d}\mu_{\Delta^{k-1}}(\boldsymbol{x})\\ &= \sum_{\boldsymbol{\eta}\in D_{m,k}}\binom{m}{\eta_1,\dots,\eta_k}\int_{\Delta^{k-1}}\prod_{j=1}^{k}\left(w_j^{\eta_j}x_j^{p\eta_j}\right)\,\mathrm{d}\mu_{\Delta^{k-1}}(x)\\ &= \frac{(k-1)!m!}{\sqrt{k}}\sum_{\boldsymbol{\eta}\in D_{m,k}}\left(\prod_{j=1}^{k}\frac{w_j^{\eta_j}}{\eta_j!}\right)\int_{\Delta^{k-1}}\prod_{i=1}^{k}x_i^{p\eta_i}\,\mathrm{d}\sigma(x)\\ &= \frac{(k-1)!m!}{\sqrt{k}}\sum_{\boldsymbol{\eta}\in D_{m,k}}\left(\prod_{j=1}^{k}\frac{w_j^{\eta_j}}{\eta_j!}\right)\times\\ &\quad\int_{\Pi\Delta^{k-1}}\left(\prod_{i=1}^{k-1}x_i^{p\eta_i}\right)(1-x_1-\dots-x_{k-1})^{p\eta_k}\sqrt{k}\,\mathrm{d}\lambda_{k-1}(x) \qquad (4.16)\\ &= \frac{(k-1)!m!}{(pm+k-1)!}\sum_{\boldsymbol{\eta}\in D_{m,k}}\prod_{j=1}^{k}\frac{(p\eta_j)!}{\eta_j!}w_j^{\eta_j} \qquad (4.17)\end{aligned}$$

$$= \frac{(k-1)!m!}{(pm+k-1)!}[x^m]\prod_{j=1}^{k}\left(\sum_{i=0}^{\infty}\frac{(pi)!}{i!}(w_jx)^i\right) \tag{4.18}$$

where $\sigma(\mathrm{d}x)$ is (unnormalized) surface measure on Δ^{k-1}, $\Pi\Delta^{k-1}$ the projection of Δ^{k-1} on the hyperplane spanned by the first $k-1$ coordinate axes, and (4.17) follows from recognizing the integral in (4.16) as the partition function of a Dirichlet variable with parameters $(p\eta_1,\dots,p\eta_k)$. We identify (4.18) as (4.15), and thus complete the first part of the proof. The second part now follows as in Theorem 1 from computing (4.18) using the Fast Fourier Transform. □

Remark 1. *The generating function $Q_p(x)$ in Theorem 2 belongs to a class of well-known special function, which is not overly surprising given the occurrence of $\mu_{\Delta^{k-1}}$ in various applications in physics (where it is known as the Bose-Einstein distribution). More precisely, $Q_p(x)$ can be expressed as the generalized hypergeometric series*

$$Q_p(x) = {}_pF_0\left[1,\frac{1}{p},\frac{2}{p},\dots,\frac{p-1}{p}\right]\left(p^2x\right).$$

In particular, for $p=2$ (i.e., including the Greenwood statistic $\|\boldsymbol{S}_k\|^2_{2,\mathbf{1}_k}$), we have $Q_2(x) = {}_2F_0\left[1,\frac{1}{2}\right](4x) = \frac{1}{\sqrt{x}}D\left(\frac{1}{2\sqrt{x}}\right)$, where Dawson's integral

$$D(x) = e^{-x^2}\int_0^x e^{t^2}\,\mathrm{d}t \tag{4.19}$$

is interpreted through its asymptotic expansion (cf., formula 7.1.23 in [AS65]).

Together with Proposition 6 to be discussed later, Theorem 2 clarifies the distributional properties of the continuous approximations of $\|\boldsymbol{S}_{n,k}\|^p_{p,\boldsymbol{w}}$ for large n (while also satisfactorily answering Greenwood's question of describing the distributional properties of $\|\boldsymbol{S}_k\|^2_{2,\mathbf{1}_k}$, cf. Section 4.3). The following proposition guarantees the quality of these approximations:

Proposition 2. *Let $F^{p,\boldsymbol{w}}_{n,k}, F^{p,\boldsymbol{w}}_k : [0,1] \to [0,1]$ be the cumulative distribution functions of $\|\frac{\boldsymbol{S}_{n,k}}{n}\|^p_{p,\boldsymbol{w}}$ and $\|\boldsymbol{S}_k\|^p_{p,\boldsymbol{w}}$, respectively. Then we have*

$$\|F^{p,\boldsymbol{w}}_{n,k} - F^{p,\boldsymbol{w}}_k\|_\infty = O(n^{-1}), \tag{4.20}$$

for every fixed $k \geq 2$.

Proof. As in the proof of Proposition 1, let $\Lambda = \mathbb{Z}^k \cap H$ where $H = \{\boldsymbol{x} \in \mathbb{R}^k : \sum_{j=1}^k x_j = 0\}$. Denoting by

$$E^t = \{x \in \Delta^{k-1} : \|x\|^p_{p,\boldsymbol{w}} \leq t\} = \{\|\boldsymbol{S}_k\|^p_{p,\boldsymbol{w}} \leq t\} \tag{4.21}$$

the t-level set of $F^{p,\boldsymbol{w}}_k$, we observe that since the fundamental domain of Λ has diameter $\|(k-1,-1,\dots,-1)\|_2 = \sqrt{k(k-1)}$, the number $L(E^t,n)$ of lattice points in nE^t is bounded by

$$\left(n-\sqrt{k(k-1)}\right)^{k-1}\mathrm{Vol}_\Lambda\left(E^t\right) \leq \mathrm{d}\left(\Lambda\right)L(E^t,n)$$

$$\leq \left(n+\sqrt{k(k-1)}\right)^{k-1} \mathrm{Vol}_\Lambda\left(E^t\right). \tag{4.22}$$

Thus, in particular,

$$\begin{aligned}
\mu_{D_{n,k}}\left(nE^t\right) - \mu_{\Delta^{k-1}}\left(E^t\right) &= \frac{L(E^t,n)}{\binom{n+k-1}{k-1}} - (k-1)!\mathrm{Vol}_\Lambda\left(E^t\right) \\
&\leq (k-1)!\mathrm{Vol}_\Lambda\left(E^t\right)\left[\left(1+\frac{k}{n}\right)^{k-1} - 1\right] \\
&\leq \sqrt{k}\sum_{j=1}^{k-1}\binom{k-1}{j}\left(\frac{k}{n}\right)^j, \qquad (4.23)
\end{aligned}$$

where using $\mathrm{Vol}_\Lambda\left(E^1\right) = \sqrt{k}/(k-1)!$ as an upper bound for $\mathrm{Vol}_\Lambda\left(E^t\right)$ turns (4.23) independent of t. Similarly, a uniform lower bound is given by

$$\begin{aligned}
\mu_{D_{n,k}}\left(nE^t\right) - \mu_{\Delta^{k-1}}\left(E^t\right) &\geq (k-1)!\mathrm{Vol}_\Lambda\left(E^t\right)\left[\left(1-\frac{2k}{n+k-1}\right)^{k-1} - 1\right] \\
&\geq \sqrt{k}\sum_{j=1}^{k-1}\binom{k-1}{j}\left(\frac{-2k}{n+k-1}\right)^j. \qquad (4.24)
\end{aligned}$$

Combining (4.23) and (4.24) gives (4.20) as desired. □

The results presented so far cover a wide range of scenarios encountered in practice when performing two-sample tests. Before considering other limiting situations, we first examine the insight that Theorem 2 provides into the distributional properties of the generalized Greenwood statistics $||\boldsymbol{S}_k||_{p,\boldsymbol{w}}^p$, with particular emphasis on the Greenwood statistic $||\boldsymbol{S}_k||_{2,\mathbf{1}_k}^2$ itself.

Analysis of $||\boldsymbol{S}_k||_{p,\boldsymbol{w}}^p$ and the right tail probability

While providing an efficient means of computing large n limits of any generalized Greenwood statistics $||\boldsymbol{S}_k||_{p,\boldsymbol{w}}^p$, Theorem 2 (together with Proposition 6) also satisfactorily answers Greenwood's question of describing the distribution of $||\boldsymbol{S}_k||_{2,\mathbf{1}_k}^2$, which had remained open since [Gre46]. In particular, using the formulas given in (4.15) and Proposition 6, the approximate z-score tabulations of [Bur79, Cur81, Ste81] can be extended to arbitrary k and arbitrary accuracy at reasonable runtime. In practice, this should be useful both for moderate and large k, for even though a central limit theorem (in k) exists for $||\boldsymbol{S}_k||_{2,\mathbf{1}_k}^2$ [Mor47], its convergence rate is unfeasibly slow. Apart from computational improvements, Theorem 2 also sheds light on analytic properties of $||\boldsymbol{S}_k||_{p,\boldsymbol{w}}^p$, and $||\boldsymbol{S}_k||_{2,\mathbf{1}_k}^2$ more specifically.

We begin by controlling the decay of moments, which in turn will inform us about tail behavior near $x_0 = 1$.

Proposition 3. *For $p \geq 2$ and $k \geq 2$, and fixed weights $w_i \in [0,1]$, for all $i \in [k]$, we have*

$$\lim_{m\to\infty} \frac{\left(\mathbb{E}\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^{p}\right)^m}{m^{k-1}} = \frac{(k-1)!}{p^{k-1}} \cdot W_{\boldsymbol{w}}, \tag{4.25}$$

where $W_{\boldsymbol{w}} = |\{1 \leq i \leq k \ : \ w_i = 1\}|$ is the number of weights taking value 1. In particular, the Greenwood statistic satisfies

$$\lim_{m\to\infty} \frac{\left(\mathbb{E}\|\boldsymbol{S}_k\|_{2,\boldsymbol{1}_k}^{2}\right)^m}{m^{k-1}} = \frac{k!}{2^{k-1}}. \tag{4.26}$$

Proof. We first rewrite (4.17) as

$$\mathbb{E}\left(\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^{p}\right)^m = \frac{1}{\binom{pm+k-1}{k-1}} \sum_{\boldsymbol{\eta}\in D_{m,k}} \frac{\binom{m}{\eta_1,\dots,\eta_k}}{\binom{pm}{p\eta_1,\dots,p\eta_k}} \prod_{j=1}^{k} w_j^{\eta_j} = \frac{1}{\binom{pm+k-1}{k-1}} s_m^{\boldsymbol{w}}, \tag{4.27}$$

which has leading order $O\left(m^{-(k-1)}\right)$, if we can show that $s_m^{\boldsymbol{w}}$ is $\Omega(1)$. To do so, we proceed by induction on k, the length of w, proving that in fact $\lim_{m\to\infty} s_m^{\boldsymbol{w}} = W_{\boldsymbol{w}}$. It is straightforward to check that for $\eta \in \{2,\dots,m-1\}$, $\binom{m}{\eta}/\binom{pm}{p\eta}$ is bounded above by $\binom{m}{2}/\binom{2m}{2\eta}$, and thus for the base case $k=2$ we have

$$\begin{aligned} s_m^{(w_1,w_2)} = \sum_{\eta=0}^{m} \frac{\binom{m}{\eta}}{\binom{pm}{p\eta}} w_1^{\eta} w_2^{m-\eta} &\leq w_1^m + w_2^m + \frac{\binom{m}{1}}{\binom{pm}{p}} + (m-2)\frac{\binom{m}{2}}{\binom{pm}{2p}} \\ &\xrightarrow{m\to\infty} \mathbb{1}_{w_1=1} + \mathbb{1}_{w_2=1} = W_{(w_1,w_2)}, \end{aligned} \tag{4.28}$$

as desired. For the inductive step, we condition on the first entry of η to obtain

$$\begin{aligned} s_m^{(w_1,\dots,w_k)} &= \sum_{\ell=0}^{m} \frac{\binom{m}{\ell}}{\binom{pm}{p\ell}} w_1^{\ell} \sum_{\boldsymbol{\eta}\in D_{m-\ell,k-1}} \frac{\binom{m-\ell}{\eta_1,\dots,\eta_{k-1}}}{\binom{p(m-\ell)}{p\eta_1,\dots,p\eta_{k-1}}} \prod_{j=1}^{k-1} w_{j+1}^{\eta_j} \\ &= s_m^{(w_2,\dots,w_k)} + w_1^m + O\left(m^{-1}\right) \\ &\xrightarrow{m\to\infty} W_{(w_2,\dots,w_k)} + \mathbb{1}_{w_1=1} = W_{(w_1,\dots,w_k)}, \end{aligned} \tag{4.29}$$

where we used the inductive hypothesis on $s_m^{(w_1,\dots,w_k)}$, and as in (4.28), bounded summands corresponding to $\ell \in \{2,\dots,m-1\}$ by $\binom{m}{2}/\binom{pm}{2p}$. (4.25) and (4.26) now follow from taking the limit as $m \to \infty$ in (4.27). □

The above result is useful primarily for describing the right tail of $\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^{p}$. Despite the obvious utility of such a result to hypothesis testing, such a description has not been available so far, even for $\|\boldsymbol{S}_k\|_{2,\boldsymbol{1}_k}^{2}$.

Corollary 1. *For $\mathbf{w} = (w_1, \ldots, w_k)$ such that $W_{\mathbf{w}} \geq 1$, the density $f_k^{p,\mathbf{w}}$ of $\|\boldsymbol{S}_k\|_{p,\mathbf{w}}^p$ is analytic at $x_0 = 1$, and its first non-zero term in the Taylor expansion is $(k-1)W_w/2^{k-2}(1-x)^{k-2}$. That is, for x close to $x_0 = 1$, we have*

$$f_k^{p,\mathbf{w}}(x) = \frac{(k-1)W_{\mathbf{w}}}{2^{k-1}}(1-x)^{k-2} + O\left((1-x)^{k-1}\right). \tag{4.30}$$

In particular, Greenwood's statistic satisfies

$$f_k^{2,\mathbf{1}_k}(x) = \frac{\binom{k}{2}}{2^{k-2}}(1-x)^{k-2} + O\left((1-x)^{k-1}\right). \tag{4.31}$$

Proof. Let $f_k^{p,\mathbf{w}}(x) = \sum_{j=0}^{\infty} c_j(1-x)^j$ be the Taylor expansion of $f_k^{p,\mathbf{w}}$ around $x_0 = 1$. We first notice that for any $r \geq 0$,

$$\int_0^1 x^m(1-x)^r \, \mathrm{d}x = \frac{1}{m+r+1} \cdot \frac{1}{\binom{m+r}{r}}, \tag{4.32}$$

and hence, using the fact that $f_k^{p,\mathbf{w}}$ is bounded,

$$\begin{aligned} \mathbb{E}\left(\|\boldsymbol{S}_{n,k}\|_{p,\mathbf{w}}^p\right)^m &= \int_0^1 x^m f_k^{p,\mathbf{w}}(x) \, \mathrm{d}x + O\left(e^{-m}\right) \\ &= \sum_{j=0}^{\infty} c_j \int_0^1 x^m(1-x)^j \, \mathrm{d}x + O\left(e^{-m}\right) \\ &= \sum_{j=0}^{\infty} c_j \frac{1}{m+j+1} \frac{1}{\binom{m+j}{j}} + O\left(e^{-m}\right). \end{aligned} \tag{4.33}$$

Identifying the $(k-2)^{\text{nd}}$ term with (4.25) immediately yields (4.30). □

While Corollary 1 is phrased so as to clarify the decay properties of the right tail, its proof readily allows characterization of all the coefficients beyond the $(k-2)^{\text{nd}}$ one in the Taylor expansion of $f_k^{p,\mathbf{w}}$ around $x_0 = 1$. For instance, it is straightforward to compute $c_{k-1} = \binom{k}{2}(k+2)/2^k$ and $c_k = (k+1)(k+6)(k^2+7k+16)/2^{k+4}$ by hand, and more generally, c_r for arbitrary $r \in \mathbb{N}$ can be efficiently computed in $O\left(\frac{r}{p}\log\frac{r}{p}\log k + [r\log r]^2\right)$ time. See Section C.1 for the detailed algorithm. This distributional understanding of the right tail complements an explicit description of $f_k^{p,\mathbf{w}}$ on the far left worked out by [Mor53] for the Greenwood statistic, valid for $x \in [0, 1/(k-1)]$. The reason for this rather narrow understanding near 0, and a guarantee that the right tail fares much better, is provided by the following lemma.

Lemma 1. *For $k, p > 1$, the density $f_k^{p,\mathbf{1}_k}(x)$ of $\|\boldsymbol{S}_k\|_{p,\mathbf{1}_k}^p$ is analytic on the intervals $\left\{\left(\frac{1}{j^{p-1}}, \frac{1}{(j-1)^{p-1}}\right)\right\}_{j\in\{2,\ldots,k\}}$. In particular, the Taylor expansion of $f_k^{p,\mathbf{1}_k}(x)$ around $x_0 = 1$ has radius of convergence $1/2^{p-1}$.*

Proof. As we increase the radius r of an ℓ_p ball centered at the origin, the ball will intersect the $j < k$ dimensional faces of Δ^{k-1} for the first time at $r_j^p = \|\frac{1}{j}\mathbf{1}_j\|_p^p = \frac{1}{j^{p-1}}$, which can be seen from projecting the ℓ_p ball onto the j-dimensional coordinate-hyperplanes. These are the only points where $f_k^{p,\mathbf{1}_k}$ is not smooth, and therefore $f_k^{p,\mathbf{1}_k}$ must be analytic on $\{(\frac{1}{j^{p-1}}, \frac{1}{(j-1)^{p-1}})\}$ for $j \in \{2, \ldots, k\}$. □

In light of Lemma 1, it is clear that the narrow applicability of the left-tail formulas in [Mor53] is due to the quickly decreasing volume $\lambda_{k-1}\left(\Delta^{k-1}\right)$ in k: the only regime in which the intersection of an ℓ_2 and an ℓ_1 ball is straightforward to compute is when this intersection is empty (i.e., $\|\boldsymbol{S}_k\|_{2,\mathbf{1}_k}^2 \leq 1/k$) or restricted to the $k-1$-dimensional face of Δ^{k-1} (in which case this computation reduces to a calculation of the volume of spherical caps). However, the volumes of these regimes are exhausted quickly, highlighting the importance of radius-independent descriptions like Theorem 2 (in combination with Proposition 6). For the particular case of (one-sided) hypothesis testing, the following monotonicity result guarantees that small k approximations provide conservative estimates to large k instances.

Proposition 4. *For $p > 1$, the c.d.f. $F_k^{p,\mathbf{1}_k}$ of $\|\boldsymbol{S}_k\|_{p,\mathbf{1}_k}^p$ is increasing in k. That is, $F_{k'}^{p,\mathbf{1}_{k'}}(x) \geq F_k^{p,\mathbf{1}_k}(x)$ for all $x \in [0,1]$ and $k' > k$.*

Proof. We let $B_k^p(r)$ be the ℓ_p ball of radius r in $\mathbb{R}^k$, and recall that

$$\mu_{\Delta^{k-1}}\left\{\|\boldsymbol{S}_k\|_{p,\mathbf{1}_k}^p \leq r^p\right\} = \frac{\lambda_{k-1}\left(B_k^p(r) \cap \Delta^{k-1}\right)}{\lambda_{k-1}\left(\Delta^{k-1}\right)}. \tag{4.34}$$

From Lemma 1 it is clear that the proposition is true for $x \leq \frac{1}{(k-1)^{p-1}}$. In order to relate $\lambda_{k-1}\left(B_k^p(r) \cap \Delta^{k-1}\right)$ to $\lambda_{k-2}\left(B_{k-1}^p(r) \cap \Delta^{k-2}\right)$ for $x > \frac{1}{(k-1)^{p-1}}$, we define $K_i(r)$ for $i \in \{1, \ldots, k\}$ to be the cone of apex $\frac{1}{k}\mathbf{1}_k$ and base formed by the intersection of $B_k^p(r)$ with the i^{th} $(k-2)$-dimensional face of Δ^{k-1} (for some fixed enumeration of the k $(k-2)$-dimensional faces). Since $\bigcup_{i=1}^k K_i(r) \subset B_k^p(r) \cap \Delta^{k-1}$, it follows that

$$\begin{aligned}\lambda_{k-1}\left(B_k^p(r) \cap \Delta^{k-1}\right) &> \sum_{i=1}^k \lambda_{k-1}(K_i) && (4.35)\\ &= \frac{k}{k-1} \cdot \frac{1}{\sqrt{k(k-1)}} \lambda_{k-2}\left(B_k^p(r) \cap \Delta^{k-1}\right) \\ &= \frac{\sqrt{k}}{(k-1)^{\frac{3}{2}}} \lambda_{k-2}\left(B_{k-1}^p(r) \cap \Delta^{k-2}\right), && (4.36)\end{aligned}$$

where by slight abuse of notation we used $\lambda_{k-2}\left(B_k^p(r) \cap \Delta^{k-1}\right)$ for the $(k-2)$-dimensional volume of $B_k^p(r)$ intersected with the $(k-2)$-dimensional faces of Δ^{k-1}. The fact that $k \cdot \lambda_{k-2}\left(B_k^p(r) \cap \Delta^{k-1}\right) = \lambda_{k-2}\left(B_{k-1}^p(r) \cap \Delta k - 1\right)$ follows from the same projection argument as used in Lemma 1. □

Alternative scaling limits

Our treatment up to now has primarily focused on the large n limit of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ while keeping k fixed, for which we saw a rich limiting distribution with fruitful connections to the previously studied Greenwood statistic emerge. However, in the setting of two-sample testing, it may very well happen that k and n are of comparable order, in which case similar behavior cannot be expected to govern the distribution of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$. What happens in these cases is much simpler, as the following proposition demonstrates.

Proposition 5. *Assume $n,k \to \infty$, and define $\mu_{n,k,p} = k^{-1}\mathbb{E}\|\boldsymbol{S}_{n,k}\|_{p,\mathbf{1}_k}^p$, $\sigma_{n,k,p}^2 = k^{-1}\mathrm{Var}\|\boldsymbol{S}_{n,k}\|_{p,\mathbf{1}_k}^p$. If $0 < W_{\min} \le w_i \le W_{\max}$ for all $i \in \{1,\dots,k\}$, then*

$$Z_{n,k,p,\boldsymbol{w}} = \frac{\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p - \mu_{n,k,p}\left(\sum_{j=1}^k w_j\right)}{\sigma_{n,k,p}\left(\sum_{j=1}^k w_j^2\right)^{1/2}} \xrightarrow{d} \begin{cases} \mathcal{N}(0,1), & \text{if } \frac{k}{n} \to \alpha \le 1, \\ 0, & \text{if } k = o(n). \end{cases} \tag{4.37}$$

Moreover, in the case of $k \to \infty$ while n remains fixed, if for every k, $\{w_i\}_{i\in[k]}$ is the discretization $w_i = w(i/k)$ of some function $w : [0,1] \to [W_{\min}, W_{\max}]$ continuous (Lebesgue) almost everywhere, then

$$\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p \xrightarrow{d} \sum_{j=1}^n w\left(U_j\right), \tag{4.38}$$

where $\{U_j\}_{j\in[n]}$ are i.i.d. $\mathrm{Uniform}\left([0,1]\right)$.

The proof of this central limit theorem is by the method of moments, where explicit combinatorial expressions like (4.27) and known large deviations allow for precise quantification of the decorrelation in $\boldsymbol{S}_{n,k}$. The full details are presented in Section C.2. We point out here that in the two-sample setting, constraining k to grow at most linearly in n is no restriction, as the roles of balls and bins turn out to be easily exchanged. Before elaborating on the application to two-sample testing and making this statement precise, we give an efficient algorithm to reconstruct a distribution from its moments.

4.4 Reconstruction of a distribution from its moment sequence

Reconstructing a probability measure from its moments is a task that has received attention both in theoretical settings (see e.g. [AK65]), where existence and uniqueness questions are addressed, and applied statistical problems, where existence and uniqueness are typically taken for granted, and efficient algorithms for computing the distribution in question are sought.

For a discrete distribution of n atoms, the latter can be done by solving an $n \times n$ Vandermonde system, which [BP70] showed is solvable in $O\left(n^2\right)$ time. However, in our setting $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ generically has $O\left(n^{k-1}\right)$ atoms, whose precise location within $\{x_{\min},\dots,\|\boldsymbol{w}\|_\infty n^p\}$ is typically unknown. That is, solving the moment problem for $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ exactly via its associated Vandermonde system requires $O\left(\min\left\{\|\boldsymbol{w}\|_\infty^2 n^{2p}, n^{2(k-1)}\right\}\right)$ operations, which already for small values of p or k becomes prohibitively large. Moreover, aside from this computational intractability in our discrete setting, it is clear that such direct approach is unfeasible to conduct in the corresponding infinite-dimensional scenario required for $\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^p$, and hence new algorithms are needed.

A commonly proposed alternative consists of forfeiting the demand for an exact recovery and focus on approximate reconstruction instead, attempting to trade off accuracy for accelerated run-times. Examples of such approximation schemes include maximum entropy based algorithms (see e.g. [MP84]) as well as various applications of the method of moments. While the latter relies to a large extent on strong parametric assumptions, which are not available for our generalized Greenwood statistics, the former is primarily useful for density estimations with only a few explicit (or estimated) moments. Theorem 1 and Theorem 2, however, allow us access to a vast number of moments quickly, suggesting that a maximum entropy ansatz could waste valuable information. To remedy this situation, we recall a fact that is mentioned in [Fel08] (p. 227, Theorem 2), but that to the best of our knowledge has not found widespread use in applied statistics.

Fact 1. *Let $X \in [0,1]$ be a (not necessarily continuous) random variable with cumulative distribution function F and moments $\mu_m = \mathbb{E}X^m$, then at every continuity point x of F, we have $\lim_{n\to\infty} \hat{F}_n(x) = F(x)$, where*

$$\hat{F}_n := \sum_{j=0}^{n-1} \mathbb{1}_{\frac{j}{n}\le x} \binom{n}{j} (-1)^{n-j} \left(\delta^{n-j}\mu\right)_j, \tag{4.39}$$

with $\delta : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ being the difference operator $\delta : (a_j)_{j\in\mathbb{N}} \mapsto (a_{j+1} - a_j)_{j\in\mathbb{N}}$.

Part of the reason for the modest popularity of Fact 1 in statistical estimation problems may be the presence of generally large alternating summands, causing possibly uncontrollable instabilities if the moments μ_m are not known exactly. In our situation, this instability does not present any limitations, since Theorem 1 and Theorem 2 allow computation of $\mu_m = \mathbb{E}\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ (or $\mu_m = \mathbb{E}\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^p$, respectively) to arbitrary precision, thus rendering (4.39) a promising candidate for reconstructing the distributions in question. It remains to clarify its convergence speed:

Proposition 6. *Let $X \in [0,1]$ be a random variable which is either (i) absolutely continuous with respect to λ_1, with density $f \in C^1([0,1])$, or (ii) discrete with support $\mathrm{supp}_X = \{x_0, \dots, x_N\}$. Then for any resolution $\varepsilon_n \to 0, \varepsilon_n > \frac{1}{n^{0.51}}$, there exists $n_0(f,\varepsilon) \in \mathbb{N}$, so that for all $n \ge n_0$,*

$$\sup_{x\in[0,1]} \left|\hat{F}_n(x) - F(x)\right| \le \frac{\|f\|_\infty + 2\|f'\|_\infty + 2}{n+1}, \tag{i}$$

$$\sup_{x\in[0,1]\setminus \mathrm{supp}_X^{\varepsilon_n}} \left|\hat{F}_n(x) - F(x)\right| \le 2e^{-2n\varepsilon_n^2} + \left(|\mathrm{supp}_X| - 2\right) e^{-2nh^2}, \tag{ii}$$

where $\hat{F}_n(x)$ is defined in (4.39), $\mathrm{supp}_X^\varepsilon = \{x \in [0,1] \,:\, d(x, \mathrm{supp}_X) < \varepsilon\}$ is the ε-fattening, and $h = \min_{i,j} |x_i - x_j|$ is the mesh of supp_X.

Proof. We first tackle (i) by recalling from [Fel08] that the summation in (4.39) is nothing but

$$\mathbb{E}B_{n,x}(X) = \mathbb{E}\sum_{j=0}^{n-1} \mathbb{1}_{\frac{j}{n}\leq x}\binom{n}{j}X^k(1-X)^{n-k}, \tag{4.40}$$

where $B_{n,x}$ is the degree n approximation of $\mathbb{1}_{[0,x]}$ by Bernstein polynomials (see [Ber12]). To compute its approximation error, we choose a threshold $\varepsilon_n \to 0$ and investigate

$$\begin{aligned} F(x) - \mathbb{E}B_{n,x}(X) &= \mathbb{E}\left(\mathbb{1}_{[0,x]}(X) - B_{n,x}(X)\right) \\ &= \underbrace{\int_{[0,1]\setminus\{x\}^{\varepsilon_n}} \left(\mathbb{1}_{[0,x]}(y) - B_{n,x}(y)\right) f(y)\,\mathrm{d}y}_{A_{n,x}} \\ &\quad + \underbrace{\int_{\{x\}^{\varepsilon_n}} \left(\mathbb{1}_{[0,x]}(y) - B_{n,x}(y)\right) f(y)\,\mathrm{d}y}_{A'_{n,x}}, \end{aligned} \tag{4.41}$$

in which we treat the term $A_{n,x}$ first: Interpreting $B_{n,x}(y)$ as $\mathbb{P}(S_{n,y} \leq nx)$, where $S_{n,y} \sim \text{Binomial}(n,y)$, we see that by standard large deviation estimates and Pinsker's inequality

$$\begin{aligned} |A_{n,x}| \leq (x-\varepsilon_n)\|f\|_\infty e^{-nD_{\mathrm{KL}}(x|x-\varepsilon_n)} \\ + \|f\|_\infty(1-x+\varepsilon_n)e^{-nD_{\mathrm{KL}}(x|x+\varepsilon_n)} \leq \|f\|_\infty e^{-2n\varepsilon_n^2}, \end{aligned} \tag{4.42}$$

where $D_{\mathrm{KL}}(p \mid q)$ is the Kullback-Leibler divergence (or the relative entropy) between a Bernoulli(p) and Bernoulli(q) distribution. To control $A'_{n,x}$ then, we Taylor expand f to rewrite the integral in (4.41) as

$$\begin{aligned} A'_{n,x} &= \int_{\{x\}^{\varepsilon_n}} \left(\mathbb{1}_{[0,x]}(y) - B_{n,x}(y)\right)\left(f(x) + f'(\xi_{y,x})(y-x)\right)\,\mathrm{d}y \\ &= (f(x) - M_n\cdot x)\underbrace{\int_{\{x\}^{\varepsilon_n}} \left(\mathbb{1}_{[0,x]}(y) - B_{n,x}(y)\right)\mathrm{d}y}_{A''_{n,x}} \\ &\quad + M_n\underbrace{\int_{\{x\}^{\varepsilon_n}} \left(\mathbb{1}_{[0,x]}(y) - B_{n,x}(y)\right) y\,\mathrm{d}y}_{A'''_{n,x}}, \end{aligned} \tag{4.43}$$

where $\min_{y\in\{x\}^{\varepsilon_n}} f'(y) \leq M_n \leq \max_{y\in\{x\}^{\varepsilon_n}} f'(y)$. In particular, since we assumed $f \in C^1([0,1])$ and $\varepsilon_n \to 0$, there must exist a n'_0 so that $f'(x) - 1 \leq M_n \leq f'(x) + 1$ for all $n \geq n'_0$. So it remains to control $A''_{n,x}$ and $A'''_{n,x}$, which can be done in a manner similar to (4.42):

$$\left|A''_{n,x}\right| \leq \int_{[0,1]} \left(\mathbb{1}_{[0,x]}(y) - B_{n,x}(y)\right)\,\mathrm{d}y + e^{-2n\varepsilon_n^2} = \frac{x-1}{n+1} + e^{-2n\varepsilon_n^2} \tag{4.44}$$

$$\begin{aligned}
&\leq \frac{1}{n+1} + e^{-2n\varepsilon_n^2} \\
\left|A'''_{n,x}\right| &\leq \int_{[0,1]} \left(\mathbb{1}_{[0,x]}(y) - B_{n,x}(y)\right) y \, \mathrm{d}y + e^{-2n\varepsilon_n^2} \qquad (4.45)\\
&= \frac{3nt(x-1)+2(x^2-1)}{2(n+1)(n+2)} + e^{-2n\varepsilon_n^2} \leq \frac{1}{n+1} + e^{-2n\varepsilon_n^2},
\end{aligned}$$

provided $n \geq 4$. Finally, combining (4.41)-(4.45), we obtain

$$\left|\hat{F}_n(x) - F(x)\right| \leq \frac{\|f\|_\infty + 2\|f'\|_\infty + 2}{n+1} + 2\left(\|f\|_\infty + \|f'\|_\infty\right) e^{-2n\varepsilon_n^2}, \qquad (4.46)$$

independently of x. Choosing $\varepsilon_n \geq n^{-\frac{1}{2}+\delta}$ and n_0 so large that the first term dominates the second yields (i). (ii) follows in a very similar manner by observing that for n such that $\varepsilon_n < h$, any $x \in [0,1] \setminus \operatorname{supp}_X^{\varepsilon_n}$ satisfies

$$\left|\mathbb{1}_{[0,x]}(y) - B_{n,x}(y)\right| \leq \begin{cases} e^{-2n\varepsilon_n^2}, & \text{if } y \in \{y_{\min}(x), y_{\max}(x)\}, \\ e^{-2nh^2}, & \text{for all other } y \in \operatorname{supp}_X, \end{cases} \qquad (4.47)$$

where $y_{\min}(x) = \min\{y' \in \operatorname{supp}_X : y' > x\}$ and $y_{\max}(x) = \max\{y' \in \operatorname{supp}_X : y' < x\}$ are the two atoms of X left and right of x. Therefore,

$$\begin{aligned}
|F(x) - \mathbb{E}B_{n,x}(X)| &\leq \sum_{y \in \operatorname{supp}_X} \mathbb{P}(X=y) \left|\mathbb{1}_{[0,x]}(y) - B_{n,x}(y)\right| \\
&\leq 2e^{-2n\varepsilon_n^2} + (|\operatorname{supp}_X| - 2)\, e^{-2nh^2}, \qquad (4.48)
\end{aligned}$$

which is (ii). □

We remark that the proof works equally well for distributions that have both an absolutely continuous and a singular part (with respect to λ_1), in which case the continuous component presents the bottleneck, resulting in an $O\left(n^{-1}\right)$ bound like in (i). For purely discrete measures however, we notice that by setting $\varepsilon_n = \varepsilon < h/2$ in (ii), we can reconstruct $F(x)$ for $x \in \operatorname{supp}_X$ up to exponentially decreasing error (in the number of moments) by computing $\hat{F}(x+2\varepsilon)$. Moreover, the bounds (i)-(ii) present worst case errors that are achieved at x for which $f(x), |f'(x)|$ are large or atoms of X densely packed, respectively. Away from these bottlenecks, and in particular in the tails of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}$ and $\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}$, these guarantees should improve significantly. Lastly, we may replace each use of Bernstein polynomials throughout the entire analysis with any other expedient polynomial approximation of $\mathbb{1}_{[0,x]}$ in order to impose desired properties on the reconstructed density. If, e.g., one-sided reconstructions are preferable (for instance, in order to give rise to conservative hypothesis tests in Section 4.5), then resorting to appropriate one-sided polynomial approximations (the optimal of which is worked out in [BQMC12]) will enforce this preference. To summarize our situation then:

1. We can approximate the distribution of $\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^p$ on $[a_{p,\boldsymbol{w}}, 1]$ (where $a_{2,\boldsymbol{1}_k} = \frac{1}{2}$ and $a_{p,\boldsymbol{w}} < 1$ generally) by computing its exact Taylor expansion of order r around $x_0 = 1$ in $O\left(\frac{r}{p}\log\frac{r}{p}\log k + [r\log r]^2\right)$ time (see Proposition 3 and Section C.1).

2. On $[0, a_{p,\boldsymbol{w}}]$, we can achieve a uniform approximation error of ε in $O\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\log k\right)$ (see Proposition 3 and Proposition 6 (ii)). Moreover, in the special case that $p = 2$ and $\boldsymbol{w} = \boldsymbol{1}_k$, exact formulas for any $x \in [0, \frac{1}{k-1}]$ are available by Lemma 1 and its preceding remarks.

3. For n large, we can approximate the distribution of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ by that of $\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^p$ and using the two bullet points above. The additional error incurred is of order $O\left(n^{-1}\right)$ by Proposition 2.

4. For any n and k, Proposition 6 and the remark following its proof allow an ε-approximation of the distribution of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ in $O\left(\log\frac{1}{\varepsilon}\right)$ time.

Observations 1, 3 and 4 in particular render the generalized Greenwood statistics as promising statistics for hypothesis testing.

4.5 Application to non-parametric hypothesis tests

Having developed a thorough distributional understanding of both Greenwood's and Dixon's statistics as well as their various generalizations, we are now in a position to illuminate their role in the hypothesis tests that motivated them. We begin by carrying out the original test of uniformity proposed by [Gre46]. We demonstrate its power in comparison to other commonly used test statistics, and describe its implications for the wider class of one-sample tests. Our second application is then devoted to clarifying completely the two-sample test described in [Dix40] by replacing low-order approximations and lifting limiting sample size constraints that were assumed therein; in addition to illustrating how the flexibility that comes with our family of generalized test statistics can substantially improve power.

Tests of uniformity and one-sample tests

Recall from Section 4.1 that the null hypothesis to be queried in [Gre46] concerned determining the distributional family of k sample times $T_1, \ldots, T_k$. Namely,

$$\mathscr{H}_0 : \{T_j\}_{j\in[k]} \overset{i.i.d.}{\sim} \mathscr{E}(\lambda), \quad \text{for some } \lambda \in \mathbb{R}_+, \tag{4.49}$$

where $\mathscr{E}(\lambda)$ denotes the exponential distribution of rate λ. It is of particular interest to be able to detect alternatives consisting of point-processes whose inter-arrival times show either under- or overdispersion with respect to this homogeneous Poisson(λ)-process; that is, it is desirable to maximize power against

$$\mathscr{H}_1 : \{T_j\}_{j\in[k]} \overset{i.i.d.}{\sim} X, \quad \text{where } c_V^2 = \frac{\operatorname{Var}X}{(\mathbb{E}X)^2} \neq 1. \tag{4.50}$$

Through normalizing by $\sum_j T_j$, Greenwood noticed that this decision problem is tantamount to the task of distinguishing the null hypothesis

$$\mathscr{H}_0 : \left(T_1, T_1 + T_2, \ldots, \sum_{j=1}^{k-1} T_j\right) \Big/ \left(\sum_{j=1}^{k} T_j\right) \sim \left(U_{(1)}, \ldots, U_{(k-1)}\right), \tag{4.51}$$

where $\{U_j\}_{j\in[k-1]} \overset{i.i.d.}{\sim} \text{Uniform}\left([0,1]\right)$, and $U_{(j)}$ is the j^{th} order statistic, from a class of alternatives where points in $[0,1]$ tend to, intuitively, be overly equi-spaced (corresponding to $c_V^2 < 1$) or overly clustered (resulting from $c_V^2 > 1$). This, in turn, is translated to the level of spacings as

$$\mathscr{H}_0 : \left(T_1, \ldots, T_k\right) \Big/ \left(\sum_{j=1}^{k} T_j\right) \sim \mu_{\Delta^{k-1}}, \tag{4.52}$$

with spacings in the alternative class exhibiting either smaller ($c_V^2 < 1$) or larger ($c_V^2 > 1$) variances than under the null. It is this last formulation (4.52) that motivated Greenwood to introduce his eponymous statistic $\left(\sum_{j=1}^{k} T_j^2\right) / \left(\sum_{j=1}^{k} T_j\right)^2$, whose law under the null is simply that of $\|\boldsymbol{S}_k\|_{2,\mathbf{1}_k}^2$ in our notation above. Greenwood successfully treated the case $k = 2$, but was unable to extend his results to larger sample sizes. Theorem 2 and Proposition 6 fill this gap by allowing us to compute p-values efficiently and accurately. Indeed, our algorithm proceeds fast enough to run large scale power studies for an extensive range of k (computing the p-value of a sample of $10,000$ points takes roughly 5 seconds on an ordinary laptop), all of which return results that qualitatively resemble those depicted in Figure 4.1: both the absolute power of Greenwood's test, as well as its performance relative to three other popular test of uniformity (Pearson's χ^2, Kolmogorov-Smirnov, Cramer-von-Mises), are uniformly high (and particularly pronounced in the case of underdispersed data), rendering it a suitable hypothesis test to decide (4.49) against (4.50).

This performance is especially encouraging in light of the role that tests of uniformity play in the larger context of one-sample testing, where one is given a sample $Z_1, \ldots, Z_k$ of size k, and wants to ascertain whether these k samples all arose in an *i.i.d.* fashion from the same continuous distribution F. To ask whether $\{Z_i\}_{i\in[k]}$ are *i.i.d.* F however, is the same as to ask whether $\{F(Z_i)\}_{i\in[k]}$ are distributed *i.i.d.* as $U \sim \text{Uniform}\left([0,1]\right)$, which is nothing but the test of uniformity we just conducted. Naturally, the classes of alternatives likely encountered in this new setting will most often differ from those in our previous considerations. However, part of the benefit of having substantial analytical control on the entire *family* of generalized Greenwood statistics $\left\{\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p\right\}_{p\in\mathbb{N},\boldsymbol{w}\in\mathbb{R}^k}$ is the ability to accommodate various, even strongly disparate, alternatives by means of adjusting p and $\boldsymbol{w}$. As this is best illustrated in the framework of two-sample tests, and readily reduced to one-sample tests from there, we will not delineate the details here, but rather develop them in the following subsection on two-sample tests, while being careful to point out any particular adjustments that may be necessary for application to one-sample tests.

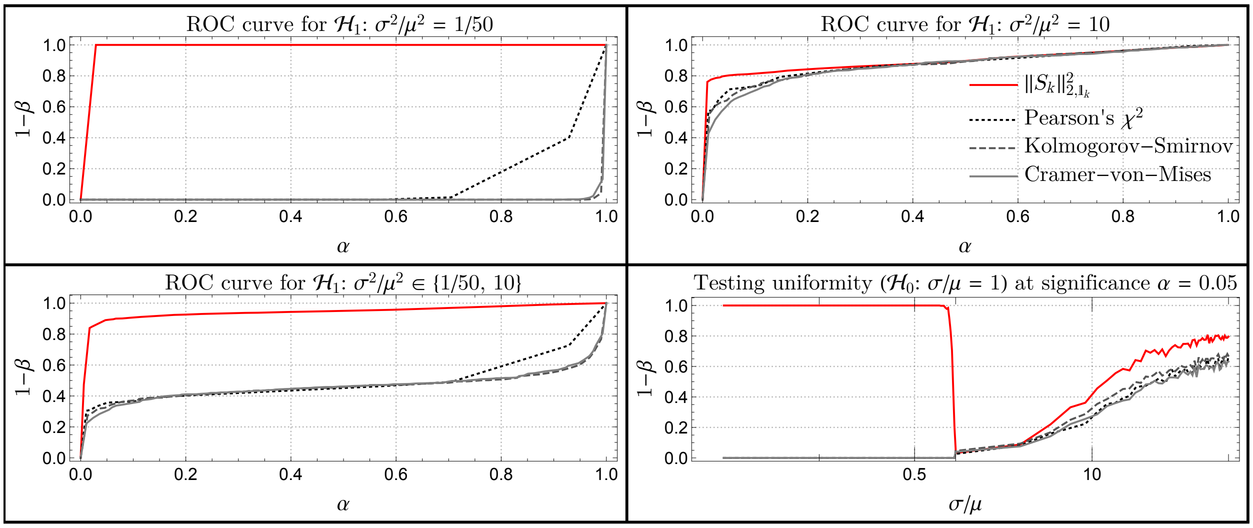


Figure 4.1: One-sample test results. ROC and power curves of tests based on the Greenwood statistic $\|\boldsymbol{S}_k\|^2_{2,\mathbf{1}_k}$ on under- and overdispersed data for $k = 10$, compared to three other commonly used tests of uniformity (Pearson's χ^2, Kolmogorov-Smirnov, Cramer-von-Mises): each experiment consists of 1000 independently drawn samples from the null ($\mu_{\Delta^{k-1}}$) and alternative (Erlang and Hyperexponential, either as individual classes as in the top two panels, or mixed into one class and presented at equal probability as in the bottom left panel) distributions matching the stated coefficients of variation; α denotes the type I error, while β is the type II error (and $1-\beta$ is power).

Two-sample tests

One-sample tests are in a concrete sense (namely, that of (4.38) in Proposition 5) large sample-size limits of two-sample tests: Instead of judging whether the generating distribution of one given sample $\{Z_i\}_{i\in[k]}$ matches a suspected given continuous F, the task is to decide whether two drawn samples $\{X_i\}_{i\in[k-1]}$ and $\{Y_j\}_{j\in[n]}$ have identical generating mechanisms. That is, assuming that $\{X_i\}_{i\in[k-1]}$ and $\{Y_j\}_{j\in[n]}$ are generated *i.i.d.* from F and G, respectively, the null hypothesis to be tested is

$$\mathscr{H}_0 : F = G, \tag{4.53}$$

which indeed in the $n \to \infty$ limit (where G becomes fully known) reduces to the one-sample setting. The equivalent of the uniformizing transformation $\{Z_i\}_{i\in[k]} \to \{F(Z_i)\}_{i\in[k]}$ in the one-sample setting is now given by the discrete uniformization $\{X_i\}_{i\in[k-1]}, \{Y_j\}_{j\in[n]} \to \{S_{n,k}[\![j]\!]\}_{j\in[k]}$, where $S_{n,k}[\![j]\!]$ is as defined in (4.1). It is straightforward to verify that $\boldsymbol{S}_{n,k} = (S_{n,k}[\![1]\!], S_{n,k}[\![2]\!], \ldots, S_{n,k}[\![k]\!])$ is distributed as Multinomial$(n, \boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \sim$ Dirichlet$(1, \ldots, 1)$, and that this law is precisely $\mu_{D_{n,k}}$. Consequently, to probe $\mathscr{H}_0$ is really to probe whether $\boldsymbol{S}_{n,k}$ is distributed according to $\mu_{D_{n,k}}$ or not. An array of test statistics devised to be sensitive against either arbitrary (e.g., the Kolmogorov-Smirnov test [Kol33, Smi48] or the Cramer-von-Mises test [Cra28, VM13]) or specific (e.g., the Mann-Whitney test [MW47]) families of alternatives have surfaced over the last century, yet, to the best of our knowledge, a clear understanding of their relative power against each such family of alternatives has remained elusive. In other words, given various classes of alternatives $\mathscr{A}_1,\dots,\mathscr{A}_d$ that a practitioner might deem likely to present themselves, it is often unclear how an appropriate test statistic is to be chosen. Having access to fast numerical evaluations of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ and their laws for arbitrary p and w offers one attractive solution to this problem: it allows the user to optimize any quantity of interest (like power) over this family of generalized Greenwood statistics quickly. To wit, assuming for now a sufficiently well-behaved class of alternatives $\mathscr{A}$, i.e. $\mathscr{H}_1 : G \in \mathscr{A}$, and denoting by $\mathbb{H}_{n,k}^{p,\boldsymbol{w}}$ the law of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ on $\mathbb{R}$ induced by discretely uniformizing $\{Y_j\}_{j\in[n]} \overset{i.i.d.}{\sim} H \in \mathscr{A}$ through $\{X_i\}_{i\in[n]}$, the (two-sided) power at significance threshold α is computed as

$$\min_{H\in\mathscr{A}}\left\{1-\beta_{p,\boldsymbol{w}}^{\alpha}(H)\right\} = \min_{H\in\mathscr{A}}\left\{1-\mathbb{H}_{n,k}^{p,\boldsymbol{w}}\left([z_{p,\boldsymbol{w}}^{-}(\alpha), z_{p,\boldsymbol{w}}^{+}(\alpha)]\right)\right\}, \tag{4.54}$$

where $z_{p,\boldsymbol{w}}^{\pm}(\alpha)$ are the $\frac{\alpha}{2}$- and $\frac{(1-\alpha)}{2}$-quantiles of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ under $\mu_{D_{n,k}}$ (i.e. $\mu_{D_{n,k}}\left([z_{p,\boldsymbol{w}}^{-}, z_{p,\boldsymbol{w}}^{+}(\alpha)]\right) = 1-\alpha$). $z_{p,\boldsymbol{w}}^{\pm}(\alpha)$ are efficiently computed from the moments of $\mu_{D_{n,k}}$, so in principle, if $\mathscr{A}$ is tractable enough (relative to F) to allow for an explicit characterization of $\mathbb{H}_{n,k}^{p,\boldsymbol{w}}$ for every $H \in \mathscr{A}$, (4.54) is amenable to fast numerical optimization. Alas, in practice we can hardly expect $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ under H to be as accessible as under the null, so computing (4.54) to arbitrarily high accuracy will likely prove unfeasible. Nevertheless, reasonable approximations to (4.54) are often available under mild assumptions on $\mathscr{A}$. The following two examples illustrate the process of performing such approximate optimization, its impact on statistical power, as well as how to extend this selection process to composite hypotheses.

Example 1 (Detecting heteroskedasticity)**.** *Assume without loss of generality that* $\mathbb{E}X = 0, \operatorname{Var}X = 1$, *and that we would like to test against alternatives of the form* $G = F \circ (y \mapsto y/\sigma)$, *i.e.* $Y = \sigma X$, *for any constant* $\sigma \in \mathbb{R}_+$. *It is straightforward to see that as* $\sigma \to \infty$, $\boldsymbol{S}_{n,k}$ *will be concentrated mostly on its far ends* $S_{n,k}[\![1]\!]$ *and* $S_{n,k}[\![k]\!]$ *weighted by* $F(0)$, *i.e.,* $\boldsymbol{S}_{n,k} \xrightarrow{d} \boldsymbol{S}_{n,k}^{\infty} := N_\infty \mathbb{1}_{\{1\}} + (n-N_\infty)\mathbb{1}_{\{k\}} = (N_\infty, 0, \dots, 0, n-N_\infty)$, *where* $N_\infty \sim \operatorname{Binomial}(n, F(0))$ *and* $\mathbb{P}\left(\boldsymbol{S}_{n,k} \neq \boldsymbol{S}_{n,k}^{\infty}\right) = O\left(\sigma^{-1}\right)$. *Likewise, the limiting law of* $\boldsymbol{S}_{n,k}$ *as* $\sigma \to 0$ *is quickly verified to be that of a* $\boldsymbol{S}_{n,k}^{0} := n\mathbb{1}_{\{N_0\}}$ *variable, where* $N_0 \sim \operatorname{Binomial}(n, F(0))$ *as before, again with* $\mathbb{P}\left(\boldsymbol{S}_{n,k} \neq \boldsymbol{S}_{n,k}^{0}\right) = O(\sigma)$. *It is therefore plausible to assume that most mass of* $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ *is tightly concentrated around* $\|\boldsymbol{S}_{n,k}^{\infty}\|_{p,\boldsymbol{w}}^p = (w_1 N_\infty^p + w_k (n-N_\infty)^p)$ *and* $\|\boldsymbol{S}_{n,k}^{0}\|_{p,\boldsymbol{w}}^p = n^p \|\mathbb{1}_{\{N_0\}}\|_{p,\boldsymbol{w}}^p$, *respectively, and that thus* (4.54) *is appreciably large whenever*

$$\begin{aligned}
&\mu_{D_{n,k}}\left(\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p \notin \left[\|\boldsymbol{S}_{n,k}^{0}\|_{p,\boldsymbol{w}}^p, \|\boldsymbol{S}_{n,k}^{\infty}\|_{p,\boldsymbol{w}}^p\right]\right) \\
&= \sum_{i=0}^{n}\sum_{j=0}^{n} \mathbb{P}(N_\infty = i)\mathbb{P}(N_0 = j) \\
&\qquad \times \mu_{D_{n,k}}\left(\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p \notin \left[\|\boldsymbol{S}_{n,k}^{0}\|_{p,\boldsymbol{w}}^p, \|\boldsymbol{S}_{n,k}^{\infty}\|_{p,\boldsymbol{w}}^p\right] \mid N_\infty = i, N_0 = j\right) \lesssim \alpha,
\end{aligned} \tag{4.55}$$

where by slight abuse of notation we use $[a,b]$ *to denote the interval* $[\min\{a,b\}, \max\{a,b\}]$. *This motivates solving the approximate, yet computationally tractable,*

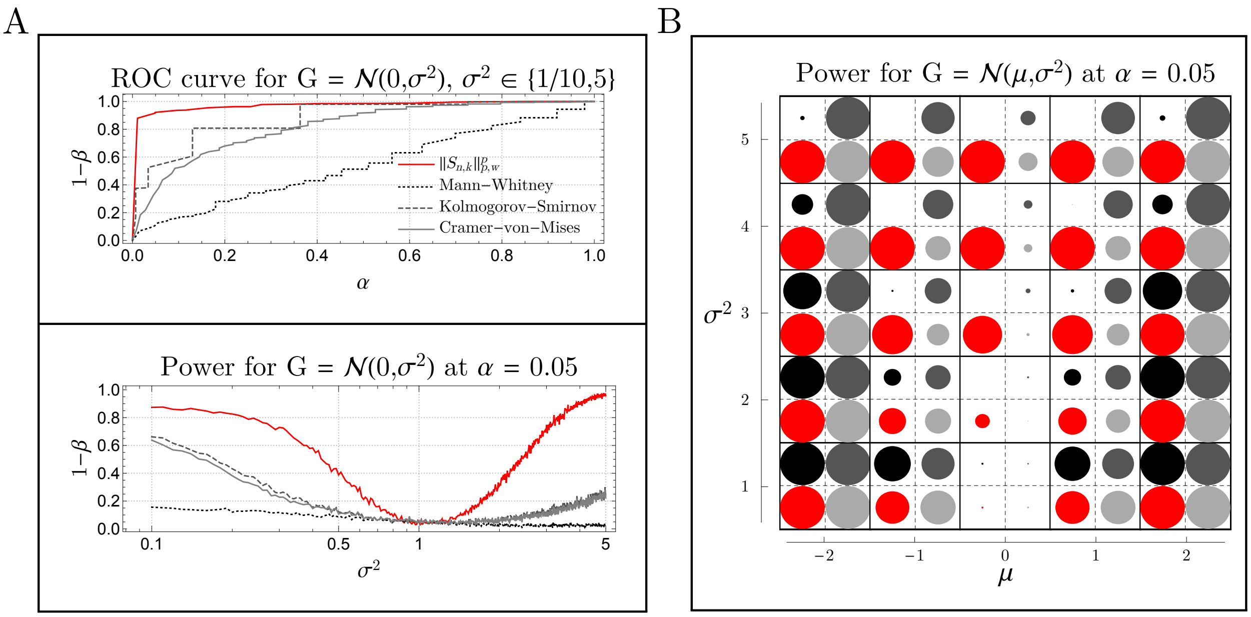


Figure 4.2: Two-sample test results. The null hypothesis tested is $F = G$ using samples $X_1, \ldots, X_{k-1} \overset{iid}{\sim} F = \text{Normal}\,(0,1)$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} G = \text{Normal}\left(\mu, \sigma^2\right)$, for $k = 10$ and $n = 30$ (different parameter choices led to minor qualitative changes only). The experimental setup is similar to that of Figure 4.1. **A**. ROC and power curves for detecting heteroskedasticity. Our new test based on the generalized spacing-statistics $\|\boldsymbol{S}_{n,k}\|^p_{p,\boldsymbol{w}}$ exhibits substantially improved performance over the other non-parametric two-sample tests. **B.** Power against joint variations in location and scale. $G = \text{Normal}\left(\mu, \sigma^2\right)$, with $\mu \in \{-2,-1,0,1,2\}$ and $\sigma^2 \in \{1,2,3,4,5\}$. Colors of bubble indicate the test statistic used—generalized spacing-statistics $\|\boldsymbol{S}_{n,k}\|^p_{p,\boldsymbol{w}}$ (red), Mann-Whitney (black), Kolmogorov-Smirnov (dark gray), Cramer-von-Mises (light gray)—while its radius indicates power $1-\beta$. The column for $\mu = 0$ corresponds to the results illustrated in the bottom panel of **A**.

optimization problem of finding

$$\arg\min_{p,\boldsymbol{w}} \mu_{D_{n,k}} \left(\|\boldsymbol{S}_{n,k}\|^p_{p,\boldsymbol{w}} \notin \left[\|\boldsymbol{S}^0_{n,k}\|^p_{p,\boldsymbol{w}}, \|\boldsymbol{S}^\infty_{n,k}\|^p_{p,\boldsymbol{w}} \right] \right), \tag{4.56}$$

for any given $F(0)$, instead of optimizing (4.54) directly. To verify empirically that any such minimizers do indeed give rise to a powerful test of heteroskedasticity, we ran large scale simulations for various F and G in the assumed family of distributions. As Figure 4.2A illustrates, where performing the optimization in (4.56) yielded parameter choices of $p = 1$ and $\boldsymbol{w} = \frac{1}{10}(10,2,1,0,0,0,0,1,2,10)$, our new two-sample test based on the generalized spacing-statistics $\|\boldsymbol{S}_{n,k}\|^p_{p,\boldsymbol{w}}$ compares favorably with other non-parametric tests (Mann-Whitney, Kolmogorov-Smirnov, and Cramer-von-Mises) commonly used for such tasks.

Although the alternatives in Example 1 are composite, the laws they induce on $D_{n,k}$ are all tightly clustered around two universal ones, thereby effectively reducing the decision task to a semi-simple hypothesis test. The extension to truly composite settings is standard:

Example 2 (Sensing location and scale)**.** *We enrich the class of alternatives in Example 1 by location shifts, i.e. we consider G of the form $G = F \circ (x \mapsto (x-\mu)/\sigma)$ for $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+$. The laws $\mathbb{G}^{\mu,\sigma}$ induced on $D_{n,k}$ now exhibit infinitely many accumulation points, barring any simple optimization of the kind we performed before. Indeed, even if we had the capacity to identify maximizers of* (4.54) *explicitly, they likely would not deliver satisfactory power, for we have*

$$\max_{p,\boldsymbol{w}} \min_{H\in\mathscr{A}} \left(1-\beta^{\alpha}_{p,\boldsymbol{w}}(H)\right) \leq \max_{p,\boldsymbol{w}} \min_{H\in\mathscr{A}_\mu\cup\mathscr{A}_\sigma} \left(1-\beta^{\alpha}_{p,\boldsymbol{w}}(H)\right), \tag{4.57}$$

where $\mathscr{A}_\sigma = \{G : G = F\circ(x\mapsto x/\sigma)$ for some $\sigma\in\mathbb{R}_+\} \subset \mathscr{A}$ and $\mathscr{A}_\mu = \{G : G = F\circ(x\mapsto x-\mu)$ for some $\mu\in\mathbb{R}\} \subset \mathscr{A}$ are the pure location shifts and pure scales, respectively. Computations as in Example 1 successfully yield powerful parameter choices against $\mathscr{A}_\mu$ and $\mathscr{A}_\sigma$ individually, yet fail to do so for $\mathscr{A}_\mu$ and $\mathscr{A}_\sigma$ jointly, producing values of $1-\beta$ not exceeding ≈ 0.6. To mend this shortcoming, we can capitalize on the successful optimizers $(p_\mu, \boldsymbol{w}_\mu)$ and $(p_\sigma, \boldsymbol{w}_\sigma)$ against $\mathscr{A}_\mu$ and $\mathscr{A}_\sigma$ individually by considering their ensemble; that is, we optimize

$$\operatorname*{arg\,max}_{p_\mu,\boldsymbol{w}_\mu,p_\sigma,\boldsymbol{w}_\sigma} \max\left\{\min_{H_\mu\in\mathscr{A}_\mu}\left(1-\beta^{\frac{\alpha}{2}}_{p_\mu,\boldsymbol{w}_\mu}(H_\mu)\right), \min_{H_\sigma\in\mathscr{A}_\sigma}\left(1-\beta^{\frac{\alpha}{2}}_{p_\sigma,\boldsymbol{w}_\sigma}(H_\sigma)\right)\right\}, \tag{4.58}$$

which, again employing strategies as in Example 1, is done efficiently. To verify the utility of $(p_\mu, \boldsymbol{w}_\mu)$ and $(p_\sigma, \boldsymbol{w}_\sigma)$ we, as before, resorted to extensive numerical simulations of which a typical outcome is depicted in the Figure 4.2B. Since the choices of n and k in this particular instance match those of Figure 4.2A, the resulting $(p_\sigma, \boldsymbol{w}_\sigma)$ are identical to $(p, \boldsymbol{w})$ of Example 1. $p_\mu = 1$ and $\boldsymbol{w}_\mu = \frac{1}{10}(10,8,7,5,1,1,0,0,0,0)$, on the other hand, turned out to closely resemble the parameter configurations that gave rise to Mann-Whitney's U statistic, which does not surprise since the latter was designed with locations shifts in mind. As a consequence, our test based on $\|\boldsymbol{S}_{n,k}\|^{p_\mu}_{p_\mu,\boldsymbol{w}_\mu}$ and $\|\boldsymbol{S}_{n,k}\|^{p_\sigma}_{p_\sigma,\boldsymbol{w}_\sigma}$ boasts power comparable to the Mann-Whitney test when sensing location shifts, while extending sensitivity to heteroskedastic alternatives as well. Notably, this comparative advantage in detecting scale changes persists even against tests like Kolmogorov-Smirnov and Cramér-von-Mises whose design is not centered around mean shifts.

Examples 1 and 2 provide manifestations of (4.54) in two concrete two-sample instances, for which the optimal choice of p happened to be $p = 1$. As mentioned before, these optimization tools are easily adapted to the one-sample setting, which often requires mere replacement of $\boldsymbol{S}_{n,k}$ with $\boldsymbol{S}_k$. The following example illustrates such adaptation in practice, while also supplying a family of circumstances in which choices of p greater than 2 lead to greater power.

Example 3 (Spiked spacing model and higher p-norms)**.** *We revisit the one-sample test in* (4.49)*, where we sought to distinguish Markovian arrival times from over- or underdispersed alternatives. In our present case, however, we investigate an alternative hypothesis $\mathscr{H}_1$ consisting of a distribution*

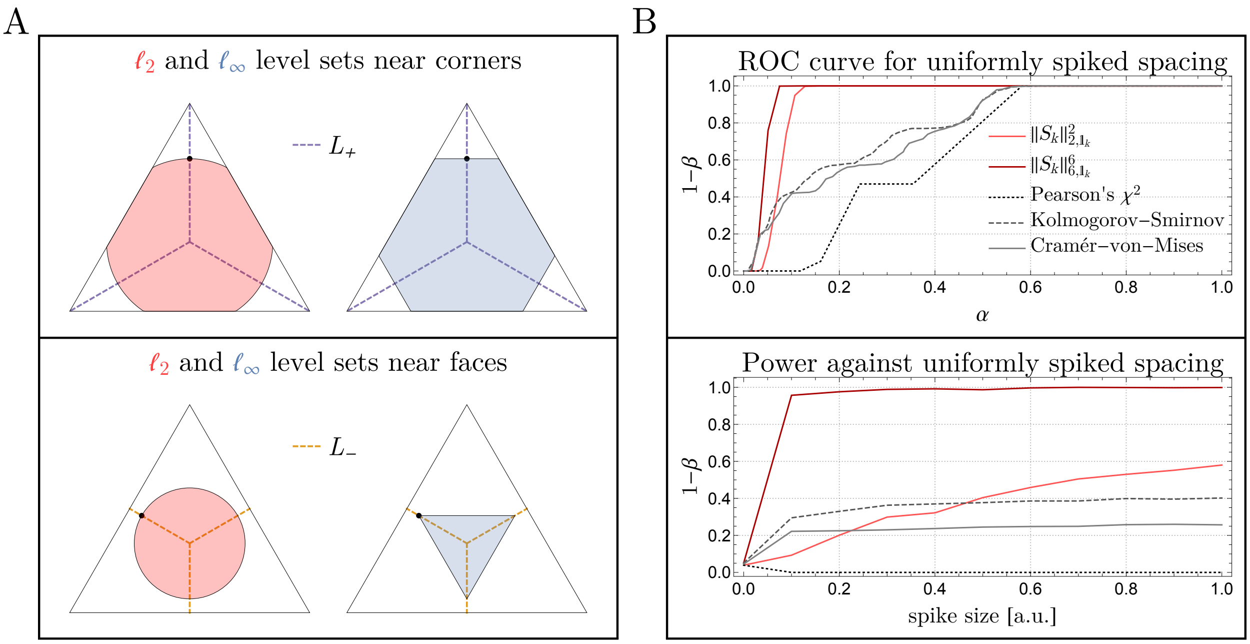


Figure 4.3: Analysis of spiked spacing model (described in detail in Example 3). **A.** Illustration of tail probabilities on Δ^2 in the cases of $p=2$ and $p=\infty$, and the samples (denoted by solid dots) giving rise to them. While a fixed sample near line segments in L_+ (purple, dashed lines in top panel) produces smaller sub-level sets in ℓ_2 than ℓ_∞ (thereby increasing the p-value of said sample, which corresponds to 1 minus the shaded area), this trend reverses for observations near line segments in L_- (orange, dashed lines in bottom panel). **B.** ROC and power curves. The spiked spacing model largely concentrates around L_+ in Δ^{k-1}, with the degree of this concentration increasing with spike size. As a consequence, p-norms of samples generated under such alternative tend to separate more markedly for larger p, which in turn affords increases in power of $\|\boldsymbol{S}_k\|^p_{p,\mathbf{1}_k}$ when $p>2$. The experimental design and choice of under- and overdispersed distributions match those of Figure 4.1; in particular, $k=10$.

G_k of $T_1,\ldots,T_k$ that is both over- and underdispersed in the following sense: Under G_k, arrival times are again drawn iid from an underdispersed distribution H_- (that is, $c_V^2(H_-)<1$), with the exception of a single randomly chosen T_K (i.e., $K\sim \mathrm{Uniform}([k])$) whose law H_+ now exhibits overdispersion ($c_V^2(H_+)>1$). In other words, G_k mixes $k-1$ underdispersed arrivals with 1 uniformly chosen overdispersed spacing. We will call this overdispersed T_K the spiked or outlier arrival time, and refer to the just described model of G_k as the spiked spacing model. Though the subsequent analysis is phrased in terms of this spiked spacing model, much of its reasoning pertains to similar outlier or correlation models of this kind as well.

To design a test capable of reliably detecting this spiked spacing model, we first observe that the symmetry in $T_1,\ldots,T_k$ (induced by the uniform choice of K) suggests little benefit of choices for $\boldsymbol{w}$

other than $\mathbf{1}_k$, leaving p as the sole parameter to optimize in (4.54). To choose among the candidates for p then, it is useful to clarify and compare the geometry that various ℓ_p balls give rise to when intersected with Δ^{k-1}: as the 2-dimensional illustrations of Figure 4.3A demonstrate, the growth of the (normalized) intersection volume $V_k^p(\boldsymbol{s}) = \mu_{\Delta^{k-1}}\left(\|\boldsymbol{S}_k\|_{p,\mathbf{1}_k}^p \leq \|\boldsymbol{s}\|_{p,\mathbf{1}_k}^p\right)$ depends noticeably on the precise location of our observation $\boldsymbol{s}$. If $\boldsymbol{s}$ localizes exactly along any of the line segments $L_+ = \left\{\overleftrightarrow{\tfrac{1}{k}\mathbf{1}_k, \boldsymbol{e}_i}\right\}_{i\in[k]}$, where $\boldsymbol{e}_i$ is the i^{th} standard basis vector, then $V_k^p(\boldsymbol{s}) \subset V_k^q(\boldsymbol{s})$ whenever $p < q$, while $V_k^p(\boldsymbol{s}) \supset V_k^q(\boldsymbol{s})$ in case $\boldsymbol{s}$ falls precisely on any of the line segments $L_- = \left\{\overleftrightarrow{\tfrac{1}{k}\mathbf{1}_k, \boldsymbol{m}_i}\right\}_{i\in[k]}$, where $\boldsymbol{m}_i = \frac{1}{k-1}\left(\mathbf{1}_k - \boldsymbol{e}_i\right)$ is the midpoint of the $(k-2)$-dimensional face opposite of vertex $\boldsymbol{e}_i$. Since p-values are nothing but $1 - V_k^p(\boldsymbol{s})$, it follows that tests based on $\|\boldsymbol{S}_k\|_{\infty,\mathbf{1}_k}^\infty$ should be most powerful in the former scenario, while $\|\boldsymbol{S}_k\|_{2,\mathbf{1}_k}^2$-based tests shine in the latter scenario, with intermediate localizations giving rise to optimal p^* between 2 and ∞. In our spiked spacing model at hand, the support of G_k gravitates towards the line segments L_+, and so we expect choices of p larger than 2 to be profitable. Indeed, carrying out simulations as in Figure 4.3B reveals this to be true, with precise values of p^* depending on the distributional details H_+ and H_-. Generally, p^* is attained around 4 or 5 for modest amplitudes of the spiked T_K and/or moderate degrees of underdispersion in the remaining arrival times, and stabilizes at 6 for more pronounced levels of spiking and/or underdispersion. Past $p = 6$, ROC and power curves tend to change only slightly.

We close this section with a few remarks on the scope and availability of our proposed hypothesis tests:

1. Even though our entire discussion is phrased around continuous null and alternative distributions F and G, the extension to discrete variables is straightforward: it merely requires recourse to a source of independent noise to randomly break ties when forming $\boldsymbol{S}_{n,k}$.

2. Due to their widespread use, our primary focus lies on applications of generalized Greenwood statistics $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$ to unpaired one- and two-sample test. However, they can naturally be deployed in any other goodness-of-fit context in which null distributions effectively reduce to $\mu_{D_{n,k}}$ or $\mu_{\Delta^{k-1}}$, e.g. paired two-sample tests.

3. (4.54) and its derived optimization problems are stated so as to incorporate rare events (under $\mathscr{H}_0$) in both the left and right tail of $\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p$. Of course, a one-sided hypothesis test can be enforced by only considering one such tail.

4. The significance threshold adjustment $\alpha/2$ in (4.58) when considering the ensemble of two generalized Greenwood statistics is exact only if their individual rejection regions are disjoint; in all other circumstances it is conservative. To extract additional power, it is possible to apply the same tools we developed throughout this chapter to compute the joint moments $\mathbb{E}\left(\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p \cdot \|\boldsymbol{S}_{n,k}\|_{q,v}^q\right)^m$, and recover from those the joint distribution $\mathbb{P}\left(\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^p \leq s, \|\boldsymbol{S}_{n,k}\|_{q,v}^q \leq t\right)$, which would allow for more refined adjustment of significance thresholds.

5. An implementation of both the one- and two-sample test in Mathematica together with precomputed parameter configurations optimal against shifts in location, scale, skewness and kurtosis (as well as combinations thereof) is available at `https://github.com/songlab-cal/mochis`.

4.6 Conclusions

Since early on, Greenwood's statistic and its relatives were theorized to be powerful candidates for a variety of goodness-of-fit tasks, yet proving them to be such, either rigorously or empirically, has, due to a lack of distributional understanding, largely remained open. Here we contribute to such distributional understanding by embedding Greenwood's statistic into a larger family of laws, the generalized Greenwood statistics $||\boldsymbol{S}_{n,k}||_{p,\boldsymbol{w}}^{p}$, whose distributional properties are more amenable to analysis. In particular, we were able to obtain explicit, efficiently computable, expressions for their associated moment sequences, and glean both qualitative (e.g., convergence, regularity and monotonicity results) as well as quantitative (convergence rates, tail behaviour, CLT) insights from them. By providing an algorithmic procedure to recover a given distribution to arbitrary accuracy from its truncated moment sequence, we are able to quickly compute quantiles and p-values, which in turn enables accurate and adaptive hypothesis tests based on said generalized Greenwood statistics. As a consequence, we were in a position to empirically verify the gains in power in two such goodness-of-fit settings, namely one- and two-sample tests, compared to conventional non-parametric test statistics widely used for these tasks.

Bibliography

[ACVMDP18] Francisco Avila Cobos, Jo Vandesompele, Pieter Mestdagh, and Katleen De Preter. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, 34(11):1969–1979, 2018.

[AK65] Naum Il'ich Akhiezer and N Kemmer. *The Classical Moment Problem: and Some Related Questions in Analysis*, volume 5. Oliver & Boyd Edinburgh, 1965.

[AS65] Milton Abramowitz and Irene A Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, volume 55. Courier Corporation, 1965.

[Bah12] C Bahadoran. Hydrodynamics and hydrostatics for a class of asymmetric particle systems with open boundaries. *Communications in Mathematical Physics*, 310(1):1–24, 2012.

[BBK$^+$16] Roy M Bremnes, Lill-Tove Busund, Thomas L Kilvær, Sigve Andersen, Elin Richardsen, Erna Elise Paulsen, Sigurd Hald, Mehrdad Rakaee Khanehkenari, Wendy A Cooper, Steven C Kao, et al. The role of tumor-infiltrating lymphocytes in development, progression, and prognosis of non–small cell lung cancer. *Journal of Thoracic Oncology*, 11(6):789–800, 2016.

[BE07] Richard A Blythe and Martin R Evans. Nonequilibrium steady states of matrix-product form: a solver's guide. *Journal of Physics A: Mathematical and Theoretical*, 40(46):R333–R441, 2007.

[Ber12] Serge Bernstein. Démonstration du théorème de weierstrass fondée sur le calcul des probabilités. *Communications de la Société Mathématique*, 13(1):1–2, 1912.

[Bil95] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 1995.

[BKK$^+$13] Dalia Burzyn, Wilson Kuswanto, Dmitriy Kolodin, Jennifer L Shadrach, Massimiliano Cerletti, Young Jang, Esen Sefik, Tze Guan Tan, Amy J Wagers, Christophe Benoist, et al. A special population of regulatory t cells potentiates muscle repair. *Cell*, 155(6):1282–1295, 2013.

[BLN$^+$16] Grégory Boël, Reka Letso, Helen Neely, W Nicholson Price, Kam-Ho Wong, Min Su, Jon D Luff, Mayank Valecha, John K Everett, Thomas B Acton, et al. Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature*, 529(7586):358–363, 2016.

[BP70] Ake Björck and Victor Pereyra. Solution of Vandermonde systems of equations. *Mathematics of computation*, 24(112):893–903, 1970.

[BQMC12] Jorge Bustamante, José M Quesada, and Reinaldo Martínez-Cruz. Best one-sided L_1 approximation to the Heaviside and sign functions. *Journal of Approximation Theory*, 164(6):791–802, 2012.

[BRH$^+$15] Mamatha Bhat, Nathaniel Robichaud, Laura Hulea, Nahum Sonenberg, Jerry Pelletier, and Ivan Topisirovic. Targeting the translation machinery in cancer. *Nature reviews Drug discovery*, 14(4):261–278, 2015.

[BSPG$^+$17] Rico D Bense, Christos Sotiriou, Martine J Piccart-Gebhart, John BAG Haanen, Marcel ATM van Vugt, Elisabeth GE de Vries, Carolien P Schröder, and Rudolf SN Fehrmann. Relevance of tumor-infiltrating immune cell composition and functionality for disease outcome in breast cancer. *JNCI: Journal of the National Cancer Institute*, 109(1), 2017.

[Bur79] Peter M Burrows. Selected percentage points of Greenwood's statistics. *Journal of the Royal Statistical Society. Series A (General)*, 142(2):256–258, 1979.

[BYAZ$^+$15] Tuval Ben-Yehezkel, Shimshi Atar, Hadas Zur, Alon Diament, Eli Goz, Tzipy Marx, Rafael Cohen, Alexandra Dana, Anna Feldman, Ehud Shapiro, et al. Rationally designed, heterologous s. cerevisiae transcripts expose novel expression determinants. *RNA biology*, 12(9):972–984, 2015.

[CBK$^+$06] Over Cabrera, Dora M Berman, Norma S Kenyon, Camillo Ricordi, Per-Olof Berggren, and Alejandro Caicedo. The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. *Proceedings of the National Academy of Sciences*, 103(7):2334–2339, 2006.

[CCAK$^+$07] Christo Christov, Fabrice Chrétien, Rana Abou-Khalil, Guillaume Bassez, Grégoire Vallet, François-Jérôme Authier, Yann Bassaglia, Vasily Shinin, Shahragim Tajbakhsh, Bénédicte Chazaud, et al. Muscle satellite cells and endothelial cells: close neighbors and privileged partners. *Molecular Biology of the Cell*, 18(4):1397–1409, 2007.

[CHA$^+$11] J Michael Cherry, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, Maria C Costanzo, Selina S Dwight, Stacia R Engel, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1):D700–D705, 2011.

[CL04] Tom Chou and Greg Lakatos. Clustered bottlenecks in mRNA translation and protein synthesis. *Phys. Rev. Lett.*, 93:198101, Nov 2004.

[Cor16] Ivan Corwin. Kardar–parisi–zhang universality. *EMS Newsletter*, (101):19–27, 2016.

[CR97] Paul Covert and Fraydoun Rezakhanlou. Hydrodynamic limit for particle systems with nonconstant speed parameter. *Journal of statistical physics*, 88(1):383–426, 1997.

[Cra28] Harald Cramér. On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74, 1928.

[Cur81] Iain D Currie. Further percentage points of Greenwood's statistic. *Journal of the Royal Statistical Society. Series A (General)*, 144(3):360–363, 1981.

[Dar53] DA Darling. On a class of problems related to the random division of an interval. *The Annals of Mathematical Statistics*, 24(2):239–253, 1953.

[DDS18] Khanh Dao Duc and Yun S. Song. The impact of ribosomal interference, codon usage, and exit tunnel interactions on translation elongation rate variation. *PLoS Genetics*, 14(e1007166):1–32, 01 2018.

[DEHP93] Bernard Derrida, Martin R Evans, Vincent Hakim, and Vincent Pasquier. Exact solution of a 1D asymmetric exclusion model using a matrix formulation. *Journal of Physics A: Mathematical and General*, 26(7):1493–1517, 1993.

[Dix40] Wilfrid J Dixon. A criterion for testing the hypothesis that two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):199–204, 1940.

[DKP16] Thomas E Dever, Terri Goss Kinzy, and Graham D Pavitt. Mechanism and regulation of protein synthesis in *Saccharomyces cerevisiae*. *Genetics*, 203(1):65–107, 2016.

[DLS97] B Derrida, JL Lebowitz, and ER Speer. Shock profiles for the asymmetric simple exclusion process in one dimension. *Journal of Statistical Physics*, 89(1-2):135–167, 1997.

[DSZ07] J. J. Dong, B. Schmittmann, and R. K. P. Zia. Inhomogeneous exclusion processes with extended objects: The effect of defect locations. *Phys. Rev. E*, 76:051113, Nov 2007.

[DTU$^+$20] Meichen Dong, Aatish Thennavan, Eugene Urrutia, Yun Li, Charles M Perou, Fei Zou, and Yuchao Jiang. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Briefings in Bioinformatics*, 2020.

[Ehr67] Eugene Ehrhart. Sur un probleme de géométrie diophantienne linéaire ii. *J. reine angew. Math*, 227(25):C49, 1967.

[EPDS20] Dan D Erdmann-Pham, Khanh Dao Duc, and Yun S Song. The key parameters that govern translation efficiency. *Cell systems*, 10(2):183–192, 2020.

[EPFHS21] Dan D Erdmann-Pham, Jonathan Fischer, Justin Hong, and Yun S Song. A likelihood-based deconvolution of bulk gene expression data using single-cell references. *Genome Research*, pages gr–272344, 2021.

[EPTS20] Dan D Erdmann-Pham, Jonathan Terhorst, and Yun S Song. Generalized spacing-statistics and a new family of non-parametric tests. *arXiv preprint arXiv:2008.06664*, 2020.

[Eva10] Lawrence C Evans. *Partial Differential Equations, Vol. 19 of Graduate Studies in Mathematics American Mathematical Society*. American Mathematical Society, Providence, Rhode Island, 2010.

[Fel08] Willliam Feller. *An Introduction to Probability Theory and its Applications*, volume 2. John Wiley & Sons, 2008.

[FLG$^+$18] Idan Frumkin, Marc J Lajoie, Christopher J Gregg, Gil Hornung, George M Church, and Yitzhak Pilpel. Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proceedings of the National Academy of Sciences*, 115(21):E4940–E4949, 2018.

[FNK$^+$03] Yukihiro Funada, Tsuyoshi Noguchi, Ryuichi Kikuchi, Shinsuke Takeno, Yuzo Uchida, and Helmut E Gabbert. Prognostic significance of cd8+ t cell and macrophage peritumoral infiltration in colorectal cancer. *Oncology Reports*, 10(2):309–313, 2003.

[FSR$^+$17] Idan Frumkin, Dvir Schirman, Aviv Rotman, Fangfei Li, Liron Zahavi, Ernest Mordret, Omer Asraf, Song Wu, Sasha F Levy, and Yitzhak Pilpel. Gene architectures that minimize cost of gene expression. *Molecular Cell*, 65(1):142–153, 2017.

[FVL$^+$14] João Fadista, Petter Vikman, Emilia Ottosson Laakso, Inês Guerra Mollet, Jonathan Lou Esguerra, Jalal Taneera, Petter Storm, Peter Osmark, Claes Ladenvall, Rashmi B Prasad, et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proceedings of the National Academy of Sciences*, 111(38):13924–13929, 2014.

[GA08] Umut Atakan Gurkan and Ozan Akkus. The mechanical environment of bone marrow: a review. *Annals of Biomedical Engineering*, 36(12):1978–1991, 2008.

[Gar52] A Gardner. Greenwood's "problem of intervals": An exact solution for $n = 3$. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):135–139, 1952.

[Gil07] Michael A Gilchrist. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Molecular Biology and Evolution*, 24(11):2362–2372, 2007.

[GMA$^+$19] Samantha L Goldman, Matthew MacKay, Ebrahim Afshinnekoo, Ari M Melnick, Shuxiu Wu, and Christopher E Mason. The impact of heterogeneity on single-cell sequencing. *Frontiers in Genetics*, 10:8, 2019.

[GMG$^+$12] Claes Gustafsson, Jeremy Minshull, Sridhar Govindarajan, Jon Ness, Alan Villalobos, and Mark Welch. Engineering genes for predictable protein expression. *Protein Expression and Purification*, 83(1):37–46, 2012.

[God60] Vidyadhar P Godambe. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211, 1960.

[Gre46] Major Greenwood. The statistical study of infectious diseases. *Journal of the Royal Statistical Society*, 109(2):85–110, 1946.

[GVL80] Gene H Golub and Charles F Van Loan. An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, 17(6):883–893, 1980.

[HC18] Gavin Hanson and Jeff Coller. Codon optimality, bias and usage in translation and mRNA decay. *Nature Reviews Molecular Cell Biology*, 19(1):20–30, 2018.

[HFK$^+$17] Peng Hu, Emily Fabyanic, Deborah Y Kwon, Sheng Tang, Zhaolan Zhou, and Hao Wu. Dissecting cell-type composition and activity-dependent transcriptional state in mammalian brains by massively parallel single-nucleus RNA-seq. *Molecular Cell*, 68(5):1006–1015, 2017.

[HKP19] Jonathan A Hensel, Vinayak Khattar, and Selvarangan Ponnazhagan. Characterization of immune cell subtypes in three commonly used mouse strains reveals gender and strain-specific variations. *Laboratory Investigation*, 99(1):93–106, 2019.

[Hol79] Lars Holst. A unified approach to limit theorems for urn models. *Journal of Applied Probability*, 16(1):154–162, 1979.

[HP08] Ruth Hershberg and Dmitri A Petrov. Selection on codon bias. *Annual Review of Genetics*, 42:287–299, 2008.

[HR80] Lars Holst and JS Rao. Asymptotic theory for some families of two-sample nonparametric statistics. *Sankhyā: The Indian Journal of Statistics, Series A*, 42:19–52, 1980.

[HSL+18] Megan Hastings Hagenauer, Anton Schulmann, Jun Z Li, Marquis P Vawter, David M Walsh, Robert C Thompson, Cortney A Turner, William E Bunney, Richard M Myers, Jack D Barchas, et al. Inference of cell type content from human brain transcriptomic datasets illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis. *PLOS One*, 13(7):e0200003, 2018.

[HWMP05] David T Humphreys, Belinda J Westman, David IK Martin, and Thomas Preiss. MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. *Proceedings of the National Academy of Sciences*, 102(47):16961–16966, 2005.

[JAR+20] Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M Garske, Jae Hoon Sul, Kirsi H Pietiläinen, Päivi Pajukanta, and Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications*, 11(1):1–11, 2020.

[KAK+19] Paulina Kasprzycka, Karolina Archacka, Kamil Kowalski, Bartosz Mierzejewski, Małgorzata Zimowska, Iwona Grabowska, Mariusz Piotrowski, Milena Rafałko, Agata Ryżko, Aliksandra Irhashava, et al. The factors present in regenerating muscles impact bone marrow-derived mesenchymal stromal/stem cell fusion with myoblasts. *Stem Cell Research & Therapy*, 10(1):1–17, 2019.

[Kal06] Olav Kallenberg. *Foundations of Modern Probability*. Springer Science & Business Media, 2006.

[Kee11] Robert W Keener. *Theoretical Statistics: Topics for a Core Course*. Springer, 2011.

[KF19] Alexandros Katranidis and Jörg Fitter. Single-molecule techniques and cell-free protein synthesis: A perfect marriage. *Analytical Chemistry*, 91(4):2570–2576, 02 2019.

[KGC+13] Sriram Kosuri, Daniel B Goodman, Guillaume Cambray, Vivek K Mutalik, Yuan Gao, Adam P Arkin, Drew Endy, and George M Church. Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proceedings of the National Academy of Sciences*, 110(34):14024–14029, 2013.

[KGF13] Anders R Kristensen, Joerg Gsponer, and Leonard J Foster. Protein synthesis rate is the predominant regulator of protein expression during differentiation. *Molecular Systems Biology*, 9(689):1–12, 2013.

[KL13] Claude Kipnis and Claudio Landim. *Scaling limits of interacting particle systems*, volume 320. Springer Science & Business Media, 2013.

[Kol33] Andrey Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

[KRS$^+$13] Tomer Kalisky, Pradeep S Rajendran, Debashis Sahoo, Sopheak Sim, Jennifer Okamoto, Stephen P Miranda, Darius M Johnston, Michael F Clarke, Stephen R Quake, and Piero Dalerba. Analysis of human colon tissue cell composition using single-cell gene-expression PCR. *Journal of Biomolecular Techniques*, 24(Suppl):S11, 2013.

[LC03] Greg Lakatos and Tom Chou. Totally asymmetric exclusion processes with particles of arbitrary size. *Journal of Physics A: Mathematical and General*, 36(8):2027–2041, 2003.

[LR14] Robert Lowe and Vardhman K Rakyan. Correcting for cell-type composition bias in epigenome-wide association studies. *Genome Medicine*, 6(3):23, 2014.

[LT18] Doron Levin and Tamir Tuller. Genome-scale analysis of perturbations in translation elongation based on a computational model. *Scientific reports*, 8(1):16191, 2018.

[LVW$^+$07] Peng Lu, Christine Vogel, Rong Wang, Xin Yao, and Edward M Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnology*, 25(1):117–124, 2007.

[MA18] Saurabh Mahajan and Deepa Agashe. Translational selection for speed is not sufficient to explain variation in bacterial codon usage bias. *Genome Biology and Evolution*, 10(2):562–576, 2018.

[MFR$^+$15] Marta Melé, Pedro G Ferreira, Ferran Reverter, David S DeLuca, Jean Monlong, Michael Sammeth, Taylor R Young, Jakob M Goldmann, Dmitri D Pervouchine, Timothy J Sullivan, et al. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, 2015.

[MGP68] Carolyn T MacDonald, Julian H Gibbs, and Allen C Pipkin. Kinetics of biopolymerization on nucleic acid templates. *Biopolymers*, 6(1):1–25, 1968.

[MK05] Reina E Mebius and Georg Kraal. Structure and function of the spleen. *Nature reviews immunology*, 5(8):606–616, 2005.

[MLF$^+$04] Vivian L MacKay, Xiaohong Li, Mark R Flory, Eileen Turcott, G Lynn Law, Kyle A Serikawa, XL Xu, Hookeun Lee, David R Goodlett, Ruedi Aebersold, et al. Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. *Molecular & Cellular Proteomics*, 3(5):478–489, 2004.

[MLX$^+$19] Gianni Monaco, Bernett Lee, Weili Xu, Seri Mustafah, You Yi Hwang, Christophe Carre, Nicolas Burdin, Lucian Visan, Michele Ceccarelli, Michael Poidinger, et al. Rna-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types. *Cell Reports*, 26(6):1627–1640, 2019.

[MMF17] Simon J Moore, James T MacDonald, and Paul S Freemont. Cell-free synthetic biology for in vitro prototype engineering. *Biochemical Society Transactions*, 45(3):785–791, 2017.

[MMO+20] Kevin Menden, Mohamed Marouf, Sergio Oller, Anupriya Dalmia, Daniel Sumner Magruder, Karin Kloiber, Peter Heutink, and Stefan Bonn. Deep learning–based cell composition analysis from tissue expression profiles. *Science Advances*, 6(30):eaba2619, 2020.

[Mor47] PAP Moran. The random division of an interval. *Supplement to the Journal of the Royal Statistical Society*, 9(1):92–98, 1947.

[Mor51] PAP Moran. The random division of an interval–Part II. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(1):147–150, 1951.

[Mor53] PAP Moran. The random division of an interval–Part III. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(1):77–80, 1953.

[MP84] Lawrence R Mead and Nikos Papanicolaou. Maximum entropy in the problem of moments. *Journal of Mathematical Physics*, 25(8):2404–2417, 1984.

[MVL+18] Melanie D Mumau, Ashley N Vanderbeck, Elizabeth D Lynch, Sophia B Golec, Stephen G Emerson, and Jennifer A Punt. Identification of a multipotent progenitor population in the spleen that is regulated by nr4a1. *The Journal of Immunology*, 200(3):1078–1087, 2018.

[MW47] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[NSL+19] Aaron M Newman, Chloé B Steen, Chih Long Liu, Andrew J Gentles, Aadel A Chaudhuri, Florian Scherer, Michael S Khodadoust, Mohammad S Esfahani, Bogdan A Luca, David Steiner, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37(7):773–782, 2019.

[NWHK14] Margaret L Novak, Eileen M Weinheimer-Haus, and Timothy J Koh. Macrophage activation and skeletal muscle healing following traumatic injury. *The Journal of Pathology*, 232(3):344–355, 2014.

[PK11] Joshua B Plotkin and Grzegorz Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12(1):32–42, 2011.

[PPS+11] Wendy W Pang, Elizabeth A Price, Debashis Sahoo, Isabel Beerman, William J Maloney, Derrick J Rossi, Stanley L Schrier, and Irving L Weissman. Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. *Proceedings of the National Academy of Sciences*, 108(50):20012–20017, 2011.

[PRI$^+$14] Cristina Pop, Silvi Rouskin, Nicholas T Ingolia, Lu Han, Eric M Phizicky, Jonathan S Weissman, and Daphne Koller. Causal signals between codon bias, mrna structure, and the efficiency of translation and elongation. *Molecular Systems Biology*, 10(12):770, 2014.

[PTSP18] PF Palamara, J Terhorst, YS Song, and AL Price. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, 50(9):1311–1317, 2018.

[QCSvdO15] Tessa EF Quax, Nico J Claassens, Dieter Söll, and John van der Oost. Codon bias as a means to fine-tune gene expression. *Molecular Cell*, 59(2):149–161, 2015.

[RC14] Gabriel Rosenblum and Barry S Cooperman. Engine out of the chassis: Cell-free protein synthesis and its uses. *FEBS letters*, 588(2):261–268, 2014.

[RCK07] Michael C Riley, Amanda Clare, and Ross D King. Locational distribution of gene functional classes in arabidopsis thaliana. *BMC Bioinformatics*, 8(1):112, 2007.

[Rez91] Fraydoum Rezakhanlou. Hydrodynamic limit for attractive particle systems on $\mathbb{Z}^d$. *Communications in Mathematical Physics*, 140(3):417–448, 1991.

[Ria86] Sedki M Riad. The deconvolution problem: An overview. *Proceedings of the IEEE*, 74(1):82–85, 1986.

[S$^+$99] Timo Seppäläinen et al. Existence of hydrodynamics for the totally asymmetric simple k-exclusion process. *The Annals of Probability*, 27(1):361–415, 1999.

[SAB15] Peter H Sudmant, Maria S Alexis, and Christopher B Burge. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biology*, 16(1):1–11, 2015.

[Sch05] G. Schönherr. Hard rod gas with long-range interactions: Exact predictions for hydrodynamic properties of continuum systems from discrete models. *Phys. Rev. E*, 71:026122, Feb 2005.

[SCN10] Andreas Schadschneider, Debashish Chowdhury, and Katsuhiro Nishinari. *Stochastic Transport in Complex Systems: from Molecules to Vehicles*. Elsevier, 2010.

[SDN$^+$13] Premal Shah, Yang Ding, Malwina Niemczyk, Grzegorz Kudla, and Joshua B Plotkin. Rate-limiting steps in yeast protein translation. *Cell*, 153(7):1589–1601, 2013.

[SdQ11] R. B. Stinchcombe and S. L. A. de Queiroz. Smoothly varying hopping rates in driven flow with exclusion. *Physical Review E*, 83:061113, Jun 2011.

[SHF$^+$16] Yuh Shiwa, Tsuyoshi Hachiya, Ryohei Furukawa, Hideki Ohmomo, Kanako Ono, Hisaaki Kudo, Jun Hata, Atsushi Hozawa, Motoki Iwasaki, Koichi Matsuda, et al. Adjustment of cell-type composition minimizes systematic bias in blood DNA methylation profiles derived by DNA collection protocols. *PLOS One*, 11(1):e0147519, 2016.

[SJNK17] Michael B Stout, Jamie N Justice, Barbara J Nicklas, and James L Kirkland. Physiological aging: links among adipose tissue dysfunction, diabetes, and frailty. *Physiology*, 32(1):9–19, 2017.

[Smi48] Nickolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948.

[SMV09] Howard M Salis, Ethan A Mirsky, and Christopher A Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, 27(10):946–950, 2009.

[SNCR18] Juraj Szavits-Nossan, Luca Ciandrini, and M. Carmen Romano. Deciphering mRNA sequence determinants of protein production rate. *Phys. Rev. Lett.*, 120:128101, Mar 2018.

[SNM$^+$15] Tim Snijders, Joshua P Nederveen, Bryon R McKay, Sophie Joanisse, Lex B Verdijk, Luc JC van Loon, and Gianni Parise. Satellite cells in human skeletal muscle plasticity. *Frontiers in Physiology*, 6:283, 2015.

[SR70] J Sethuraman and JS Rao. Pitman efficiencies of tests based on spacings. In M L Puri, editor, *Nonparametric Techniques in Statistical Inference*, page 405–416. Cambridge University Press, 1970.

[SS04] G Schönherr and GM Schütz. Exclusion process for particles of arbitrary extension: hydrodynamic limit and algebraic properties. *Journal of Physics A: Mathematical and General*, 37(34):8215–8231, 2004.

[SSL04] Leah B. Shaw, James P. Sethna, and Kelvin H. Lee. Mean-field approaches to the totally asymmetric exclusion process with quenched disorder and large particles. *Phys. Rev. E*, 70:021901, Aug 2004.

[Ste81] Michael A Stephens. Further percentage points for Greenwood's statistic. *Journal of the Royal Statistical Society. Series A (General)*, 144(3):364–366, 1981.

[SWGV14] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research*, 42(14):8845–8860, 2014.

[SZ00] G Schechtman and J Zinn. Concentration on the ℓ_p^n ball. In *Geometric Aspects of Functional Analysis*, pages 245–256. Springer, 2000.

[SZL03] Leah B Shaw, RKP Zia, and Kelvin H Lee. Totally asymmetric exclusion process with extended objects: a model for protein synthesis. *Physical Review E*, 68(2):021910(17), 2003.

[T$^+$20] Tabula Muris Consortium et al. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, 583(7817):590–595, 2020.

[TCV$^+$10] Tamir Tuller, Asaf Carmi, Kalin Vestsigian, Sivan Navon, Yuval Dorfan, John Zaborske, Tao Pan, Orna Dahan, Itay Furman, and Yitzhak Pilpel. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2):344–354, 2010.

[TDC$^+$19] Daphne Tsoucas, Rui Dong, Haide Chen, Qian Zhu, Guoji Guo, and Guo-Cheng Yuan. Accurate estimation of cell-type composition from gene expression data. *Nature Communications*, 10(1):1–9, 2019.

[Tra15] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome Research*, 25(10):1491–1498, 2015.

[VM13] Richard Von Mises. *Wahrscheinlichkeit, Statistik und Wahrheit: Einführung in d. neue Wahrscheinlichkeitslehre u. ihre Anwendung*, volume 3. Springer-Verlag, 2013.

[W$^+$03] Christopher Walsh et al. *Antibiotics: actions, origins, resistance.* American Society for Microbiology (ASM), 2003.

[WARF$^+$18] Gregory A Wyant, Monther Abu-Remaileh, Evgeni M Frenkel, Nouf N Laqtom, Vimisha Dharamdasani, Caroline A Lewis, Sze Ham Chan, Ivonne Heinze, Alessandro Ori, and David M Sabatini. NUFIP1 is a ribosome receptor for starvation-induced ribophagy. *Science*, 360:751–758, 2018.

[Wei56] Lionel Weiss. A certain class of tests of fit. *The Annals of Mathematical Statistics*, 27(4):1165–1170, 1956.

[Wil45] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[WJW14] Christopher C Williams, Calvin H Jan, and Jonathan S Weissman. Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. *Science*, 346(6210):748–751, 2014.

[WPS$^+$19] Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, 10(1):380, 2019.

[XKO$^+$16] Yurong Xin, Jinrang Kim, Haruka Okamoto, Min Ni, Yi Wei, Christina Adler, Andrew J Murphy, George D Yancopoulos, Calvin Lin, and Jesper Gromada. Rna sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metabolism*, 24(4):608–615, 2016.

[XSW$^+$20] Xuemei Xie, Qiang Shi, Peng Wu, Xiaoyu Zhang, Hiroto Kambara, Jiayu Su, Hongbo Yu, Shin-Young Park, Rongxia Guo, Qian Ren, et al. Single-cell transcriptome profiling reveals neutrophil heterogeneity in homeostasis and infection. *Nature Immunology*, 21(9):1119–1133, 2020.

[YDZ$^+$15] Chien-Hung Yu, Yunkun Dang, Zhipeng Zhou, Cheng Wu, Fangzhou Zhao, Matthew S Sachs, and Yi Liu. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Molecular Cell*, 59(5):744–754, 2015.

[YH17] Qianhui Yu and Zhisong He. Comprehensive investigation of temporal and autism-associated cell type composition-dependent and independent gene expression changes in human brains. *Scientific Reports*, 7(1):4121, 2017.

[ZDS11] Royce KP Zia, JiaJia Dong, and Beate Schmittmann. Modeling translation in protein synthesis with TASEP: A tutorial and recent developments. *Journal of Statistical Physics*, 144(2):405–428, 2011.

[ZT16] Hadas Zur and Tamir Tuller. Predictive biophysical modeling and understanding of the dynamics of mrna translation and its evolution. *Nucleic Acids Research*, 44(19):9031–9049, 2016.

[ZZZ$^+$19] Rui Zhou, Jingwen Zhang, Dongqiang Zeng, Huiying Sun, Xiaoxiang Rong, Min Shi, Jianping Bin, Yulin Liao, and Wangjun Liao. Immune cell infiltration as a biomarker for the diagnosis and prognosis of stage I-III colon cancer. *Cancer Immunology, Immunotherapy*, 68(3):433–442, 2019.

Appendix A

Supporting Information: Chapter 2

A.1 The hydrodynamic limit of the inhomogeneous ℓ-TASEP

We derive here the PDE governing the hydrodynamic limit of the open-boundaries inhomogeneous ℓ-TASEP. To do so we exploit a representation of its dynamics in terms of another interacting particle system, the so-called zero range process (ZRP), whose hydrodynamics can be found explicitly. This TASEP-ZRP duality provides an expedient and general tool for identifying explicit TASEP formulas; however, rigorously proving the validity of these formulas often requires more technical tools from probability theory. Since this work's emphasis is on the application of TASEP to unraveling the key parameters of translation dynamics, we will here concentrate on showcasing the TASEP-ZRP framework, and keep a rigorous existence proof of the hydrodynamic limit, combining techniques from [Rez91, CR97] and [Bah12], to a separate manuscript.

Reduction to periodic boundaries and mapping to the ZRP.

The purpose of the hydrodynamic limit is to describe the local evolution of the macroscopic particle density in the large system limit. As such, it does not explicitly rely on the precise formalism by which particles enter and exit the lattice at the boundaries (which will only later be needed to impose boundary conditions on the resulting PDE [Bah12]). In particular, we are free to choose periodic boundary conditions for our limiting procedure without changing the resulting PDE [SS04]. This has the advantage of preserving the total number of particles, which is essential for establishing the correspondence between TASEP and ZRP. In the following, we thus consider the ℓ-TASEP with M particles on a *ring* of N sites jumping to the right at rate p_i, and take $M,N \to \infty$ while M/N remains constant.

The ZRP is now obtained by reversing the roles of holes and particles: It consists of $N-M\ell$ particles (corresponding to the $N-M\ell$ holes in the TASEP) distributed across M sites (matching the TASEP particles) $\{1,\dots,M\}$, with multiple particles allowed to stack up on the same site. A ZRP configuration $(\xi_{i,t})_{1\leq i\leq M}$ describes the number of particles $\xi_{i,t}$ at each site $i \in \{1,\dots,M\}$ and time t, and can be seen as a representation of spacings between particles i and $i+1$ in the TASEP.

As a result, the TASEP dynamics are translated into ZRP dynamics as follows: If a site i at time t is occupied by at least one particle, then the topmost particle jumps to the left with rate $m_{i,t} = p_{k(i,t)}$, where $k(i,t)$ is the position of the ith TASEP particle (see formula (A.1) below) at time t. This jump occurs regardless of whether the destination site is occupied or not. That is, neither exclusion nor long range interactions are present, which will be key to establishing the hydrodynamic limit.

The correspondence between TASEP and ZRP states described above is so far only determined up to rotations of the TASEP lattice, hence we introduce one further variable $\xi_{0,t} \in \{1,\ldots,N\}$ to trace the position of particle 1. More explicitly, at time t, TASEP particle i is located at site

$$k(i,t) = \sum_{j=0}^{i-1} \xi_{j,t} + \ell(i-1) \tag{A.1}$$

on the TASEP ring. An illustration of this correspondence is given in Figure A.1.

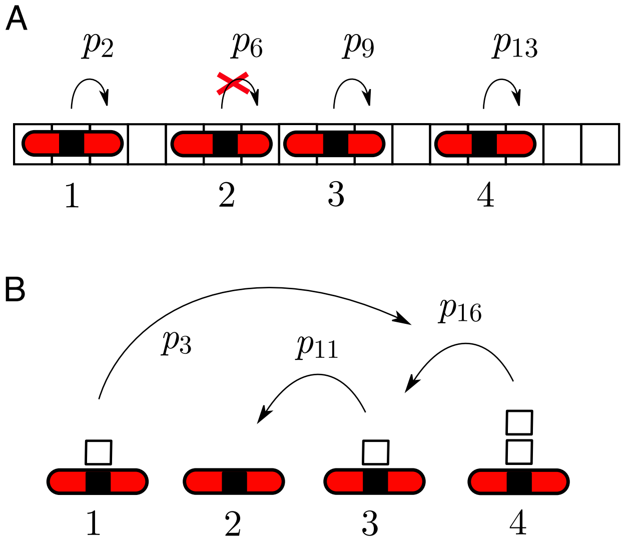


Figure A.1: **Correspondence between inhomogeneous ℓ-TASEP (A) and the ZRP (B).** ℓ-TASEP particles (rods) correspond to ZRP sites, and holes (empty squares) become ZRP particles.

The hydrodynamic limits of the ZRP and TASEP.

The connection between the TASEP and the ZRP has been fruitfully used to derive hydrodynamic limits for homogeneous systems [SS04, Sch05]. Here we generalize this approach to heterogeneous lattices and supply appropriate boundary conditions to the PDE, which become necessary when working with open rather than periodic boundaries.

We start with the master equation associated with the ZRP:

$$\partial_t \xi_{i,t} = m_{i+1,t} z_{i+1,t} - m_{i,t} z_{i,t}, \tag{A.2}$$

where $z_{i,t} = \mathbb{P}(\xi_{i,t} > 0)$ is the probability that site i is non-empty at time t. Our goal is to identify a PDE that describes the limit of (A.2) under Euler scaling, i.e., on time scale at and spatial scale ia. Denoting these scaled variables as t again in time and x, y in space such that $k = \lfloor x/a \rfloor$ and $i = \lfloor y/a \rfloor$, and assuming the existence of a continuously differentiable rate function λ such that $\lambda(x) = p_k$, the master equation (A.2) becomes

$$\begin{aligned} a\partial_t c(y,t) &= \lambda(x(y+a,t))z(y+a,t) - \lambda(x(y,t))z(y,t) \\ &= a\partial_y[\lambda(x(y,t))z(y,t)] + \frac{a^2}{2}\partial_{yy}[\lambda(x(y,t))z(y,t)] + O(a^3), \end{aligned} \tag{A.3}$$

where $c(y,t)$ and $z(y,t)$ are the continuum limits of $\xi_{i,t}$ and $z_{i,t}$, respectively. Under local stationarity [KL13], we may replace z in (A.3) using the fugacity-density relation $z = c(1+c)^{-1}$ to obtain the final hydrodynamic limit of the inhomogeneous ZRP as

$$\partial_t c = \partial_y\left(\lambda \frac{c}{1+c}\right) + \frac{a}{2}\partial_{yy}\left(\lambda \frac{c}{1+c}\right). \tag{A.4}$$

The assumption of local stationarity is essentially justified by the one-block estimates in [CR97], as long as one can ensure slow enough variation of $\lambda(x(y,t))$ in t. In our case, this smooth dependency is given, since in a small (on the Eulerian scale) time interval $N\Delta t$, we expect a particle to perform $O(N\Delta t)$ jumps, and whence $\lambda(x(y,t+N\Delta t)) - \lambda(x(y,t)) \in O(\Delta t)$.

To derive the corresponding PDE for the TASEP, we use (A.1) to establish the continuum relation between x, y and t. More precisely,

$$x(y,t) = ak(i,t) = a\left(\sum_{j=0}^{i-1}\xi_{j,t} + \ell(i-1)\right) = \int_0^y c(u,t)\,\mathrm{d}u - \frac{a}{2}\Big(c(y,t) - c(0,t)\Big) + \ell(y-a) + O(a^2). \tag{A.5}$$

Upon recognizing that particle densities are related by $\rho = (c+\ell)^{-1}$ and changing coordinates according to (A.5), (A.4) yields the hydrodynamic limit of the TASEP

$$\partial_t \rho = -\partial_x\left[\lambda(x)\rho G(\rho)\right] - \frac{a}{2}\partial_{xx}\left[\lambda(x)G(\rho)\right] + O(a^2), \tag{A.6}$$

where $G(\rho) = \frac{1-\ell\rho}{1-(\ell-1)\rho}$.

A.2 Phase diagram analysis

We now use (A.6) to provide a detailed derivation of the phase diagram described in Chapter 2.

Reduction to conservation law.

Solutions of (A.6) converge locally uniformly (under mild conditions on λ, see Section A.2) to viscosity solutions of the scalar conservation law

$$\partial_t \rho(x,t) = -\partial_x[\underbrace{\lambda(x)H(\rho(x,t))}_{J(\rho(x,t),x)}], \tag{A.7}$$

where $H(\rho) = \rho G(\rho)$, which thus determines the phase diagram in the hydrodynamic regime. Setting $\partial_t \rho = 0$ identifies the stationary profiles of the TASEP as distributions satisfying

$$J(\rho, x) = J_c, \tag{A.8}$$

where $J_c = J_c(\alpha, \beta, \lambda)$ is the critical current, set to belong to $[0, J_{\max}]$, where $J_{\max}$ is the transport capacity of the lattice

$$J_{\max} = \min_{x \in [0,1]} \max_{\rho \in [0,1/\ell]} J(\rho, x) = \frac{\lambda_{\min}}{(1+\sqrt{\ell})^2}. \tag{A.9}$$

(A.8) has two solutions (see Figure A.2A) of the form

$$\rho_{\pm}(x) = \frac{1}{2\ell} + \frac{J_c(\ell-1)}{2\ell\lambda(x)} \pm \sqrt{\left(\frac{1}{2\ell} + \frac{J_c(\ell-1)}{2\ell\lambda(x)}\right)^2 - \frac{J_c}{\ell\lambda(x)}}, \tag{A.10}$$

any mixture of which may be a potential attractor picked by the system as $t \to \infty$. Deciding precisely which mixture dominates requires analysis of the characteristic curves.

Solving the characteristic ODE.

Denoting the characteristic curves by x^t and ρ^t with initial data x^0, ρ^0, their evolution is described by the system of ODE [Eva10]

$$\frac{dx^t}{dt} = \lambda(x^t)H'(\rho^t), \tag{A.11}$$

$$\frac{d\rho^t}{dt} = -\lambda'(x^t)H(\rho^t), \tag{A.12}$$

where H' and λ' respectively denote the derivatives of H and λ with respect to their arguments. The solutions are easily verified to be

$$x^t = F^{-1}(t) \tag{A.13}$$

$$\rho^t = H^{-1}\left(\frac{J(\rho^0, x^0)}{\lambda(x^t)}\right) \tag{A.14}$$

as long as $J(\rho^0, x^0) \in [0, J_{\max}]$. The form of F follows from formally separating variables:

$$F(x) = \int_{x^0}^{x} \frac{1}{\lambda(y)H' \circ H^{-1}(J(\rho^0, x^0)/\lambda(y))}\, \mathrm{d}y, \tag{A.15}$$

while $H^{-1}(J(\rho^0, x^0)/\lambda(x^t))$ is understood to be the preimage compatible with ρ^0, see Figure A.2A. For the homogeneous ℓ-TASEP (A.13) and (A.14) depend linearly on each other, giving rise to straight line characteristic curves (see Figure A.2B). In the more general heterogeneous setting,

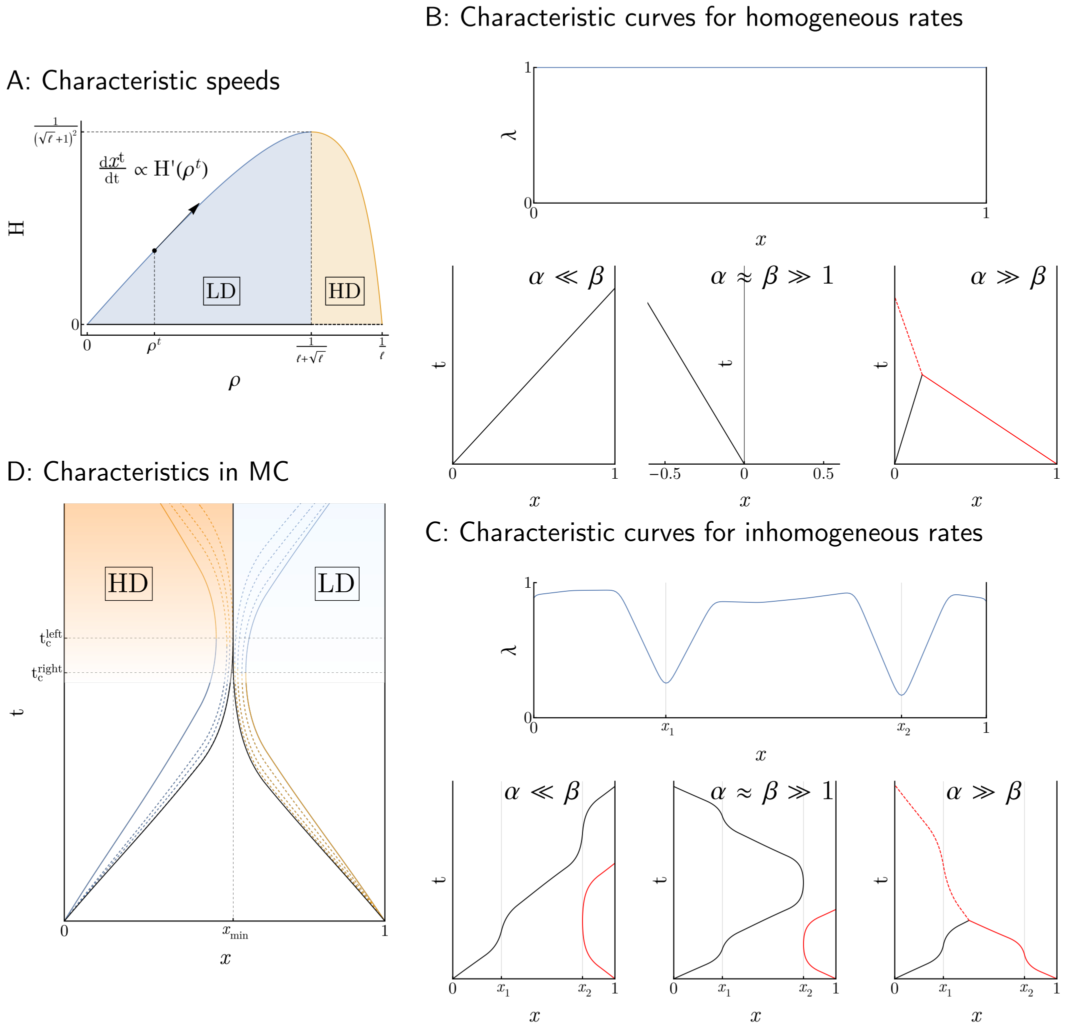


Figure A.2: $H(\rho)$ **and its effect on characteristic curves. A:** The rate-normalized flux $H(\rho) = J(\rho, x)/\lambda(x)$ is depicted, with characteristic velocity of x^t indicated. If $J(\rho^0, x^0) < J_{\max}$, ρ^t stays within the regions marked LD (blue) or HD (orange), depending on the sign of $\rho^0 - (\ell + \sqrt{\ell})^{-1}$. Otherwise, ρ^t may cross $(\ell + \sqrt{\ell})^{-1}$ forcing x^t to return to its origin x^0. **B,C:** Characteristic curves starting at lattice start (black solid curves) and end (red solid curves) for different regions of the phase diagram. Dotted curves represent shock fronts, with colors indicating which characteristic drives the shock. **B:** Homogeneous rates give rise to straight line characteristics with speed $\partial_\rho J(\rho_0)$ and $\partial_\rho J(\rho_1)$, respectively. **C:** Inhomogeneous rates produce complicated behavior, with curves slowing down (and potentially reversing direction) near the troughs (x_1 and x_2) of λ. **D:** If $J(\rho^0, x^0) > J_{\max}$ the characteristic x^t (left and right colored solid curves) reverse directions at times t_c and return to their origin. At t_c, ρ^t switches from LD (blue) to HD (orange) (if associated with $x^0 = 0$) or HD to LD (if associated with $x^0 = 1$, cf. **A**). The same happens on all associated rarefaction waves (dashed curves), which interpolate between x^t and the stationary shock of $x^t_{\max}$ (solid black curve).

however, more complicated behavior emerges (Figure A.2C). In particular, if $J(\rho^0,x^0) < J_{\max}$, then for all $t \geq 0$,

$$\frac{J(\rho^0,x^0)}{\lambda(x^t)} < \frac{1}{(1+\sqrt{\ell})^2},$$

so $\rho^t < \frac{1}{\ell+\sqrt{\ell}}$ for all t if $\rho^0 < \frac{1}{\ell+\sqrt{\ell}}$, while $\rho^t > \frac{1}{\ell+\sqrt{\ell}}$ for all t if $\rho^0 > \frac{1}{\ell+\sqrt{\ell}}$. Hence, the sign of $\frac{dx^t}{dt} = \lambda(x^t)H'(\rho^t)$ remains the same for all t, and any characteristic curve x^t starting at the left lattice boundary $x^0 = 0$ or right lattice boundary $x^0 = 1$ propagates towards the opposite end and fills the lattice entirely.

On the other hand, if $J(\rho^0,x^0) > J_{\max}$, then $\frac{J(\rho^0,x^0)}{\lambda(x_{\min})} > \frac{1}{(1+\sqrt{\ell})^2}$, where $x_{\min} = \arg\min_x \lambda(x)$, so $H^{-1}\left(\frac{J(\rho^0,x^0)}{\lambda(x_{\min})}\right) > \frac{1}{\ell}$. Recalling (A.14) and noting that it is physically not possible to have $\rho^t > \frac{1}{\ell}$, we conclude that the characteristic curve x^t cannot reach $x_{\min}$. Indeed, it follows from (A.11) and (A.12) that at some critical time t_c before reaching $x_{\min}$, the characteristic curve x^t reverses direction while ρ^t crosses $\arg\max_\rho H(\rho) = (\ell+\sqrt{\ell})^{-1}$, resulting in x^t returning to its origin. Figure 2.1E of Chapter 2 and Figure A.2D illustrate this behavior.

Computing initial densities ρ^0.

As a consequence of the above, determining phase transitions in the α-β phase diagram reduces to establishing regimes in which $J(\rho^0,x^0)$ exceeds or falls short of $J_{\max}$, which in turn is equivalent to finding an expression for ρ^0 in terms of α and β. This is done by considering each lattice end separately and balancing currents:

The right lattice end $x^0 = 1$:

As described in Chapter 2, $\rho_1 = \rho(1)$ decomposes into a sum of two contributions, the periodic part ρ_1^+ and the troughs ρ_1^- [CL04]. More explicitly,

$$\rho_1 = \frac{1}{\ell}\left[(\ell-1)\rho_1^- + \rho_1^+\right]. \tag{A.16}$$

Since the current J_c is a conserved quantity of the system, the local currents across the last lattice site, the second to last lattice site and within the last ℓ sites must all be the same:

$$J_R := J(\rho_1,1) = \beta\rho_1^+ = \lambda_1\rho_1^-. \tag{A.17}$$

Solving for ρ_1 gives exactly $\frac{1}{\ell}(1-\frac{\beta}{\lambda_1})$. Consequently, $J_R \leq J_{\max}$ iff

$$\beta < \beta^* = \frac{1}{2}\left[\lambda_1 - \frac{\ell-1}{(1+\sqrt{\ell})^2}\lambda_{\min} - \sqrt{\left(\lambda_1 - \frac{\ell-1}{(1+\sqrt{\ell})^2}\lambda_{\min}\right)^2 - \frac{4\lambda_1\lambda_{\min}}{(1+\sqrt{\ell})^2}}\right]. \tag{A.18}$$

The left lattice end $x^0 = 0$:

Computing α^* is more delicate as the effective jump rate is a combination of entrance rate and particle exclusion. To bypass this problem, we investigate the current of *holes* rather than particles, which is running in the opposite direction. With the loss of the particle-hole symmetry present in the simple 1-TASEP [DEHP93], the hole density ρ^h here assumes a more complicated form. It satisfies its own conservation law given by

$$\partial_{t^h}\rho^h = \partial_x[J^h(\rho^h, x)], \tag{A.19}$$

where

$$J^h(\rho^h, x) = \lambda(x)\rho^h \frac{1-\rho^h}{1+(\ell-1)\rho^h} \tag{A.20}$$

and $t^h = \ell t$ is the time scale of the holes, moving slower as their density is higher. Thus by balancing hole currents rather than particle currents at $x^0 = 0$, we obtain, noting that the effective exit rate (of holes) is still α (as ℓ holes need to accumulate for exiting to happen),

$$J^h(\rho_0^h, 0) = \alpha\rho_0^h. \tag{A.21}$$

Solving for ρ_0^h and using $\rho_0^h = 1 - \ell\rho_0$, we obtain $\rho_0 = \alpha/[\lambda_0 + (\ell-1)\alpha]$. Defining $J_L := J(\rho_0, 0)$, we obtain α^* by solving for α, $J_L = J_{\max}$.

Phase transitions and profiles.

Using the densities obtained from (A.17) and (A.21) in the characteristic curves (A.11) and (A.12) yields the HD and LD regimes for parameter configurations $(\alpha > \alpha^*, \beta < \beta^*)$ and $(\alpha < \alpha^*, \beta > \beta^*)$, respectively. To describe the phase transition between HD and LD, we observe that for $\alpha < \alpha^*$ and $\beta < \beta^*$ both characteristic curves move into the lattice, meet, and move along a common shock with speed

$$v_{\text{shock}} = \frac{J_R - J_L}{\rho_r - \rho_l}, \tag{A.22}$$

where ρ_l and ρ_r are the densities left and right of the shock. As $\rho_r - \rho_l > 0$ as long as $\alpha < \alpha^*$ and $\beta < \beta^*$ (cf. Figure A.2A), $v_{\text{shock}} > 0$ if and only if $J_R > J_L$. That is, the slower current pushes the faster one past the lattice boundaries and dominates the stationary behavior of the system. The HD and LD regimes are thus separated by incoming currents of equal magnitudes

$$J_L = \frac{\alpha(\lambda_0 - \alpha)}{\lambda_0 + (\ell-1)\alpha} = \frac{\beta(\lambda_1 - \beta)}{\lambda_1 + (\ell-1)\beta} = J_R. \tag{A.23}$$

Lastly, we can use the behavior of characteristic curves for $J(\rho^0, x^0) > J_{\max}$ to describe stationary profiles in the MC regime ($\alpha > \alpha^*$ and $\beta > \beta^*$): Each characteristic curve reverses direction at a critical time t_c and returns to its respective lattice boundary, while the density ρ^t it carries transitions from ρ_- to ρ_+ (on the left characteristic) or ρ_+ to ρ_- (on the right characteristic). Since the reversal of directions occurs strictly before reaching $x_{\min}$, these characteristics provide density information on only part of the lattice. The uncovered regions are determined by the simultaneously propagating rarefaction waves [Eva10], which interpolate between x^t and the characteristic curve $x^t_{\max}$ associated with $J(\rho^0, x^0) = J_{\max}$ (see Figure A.2D). Together, these observations combine to produce the high density and low density profiles to the left and right of $x_{\min}$, respectively, with critical current $J_c = J_{\max}$, as described in Chapter 2.

If λ has exactly one global minimum $x_{\min}$, this description captures the density profile on the entire lattice. In the case of multiple global minima at $\{x_{\min,1}, \ldots, x_{\min,n}\}$ however, it describes ρ on $[0, x_{\min,1}] \cup [x_{\min,n}, 1]$ only, leaving open fluctuations on the middle segment $(x_{\min,1}, x_{\min,n})$. Although unlikely to be encountered in practice, these singular rate functions exhibit interesting stochastic phenomena: The presence of high densities on the initial interval and low densities on the terminal one suggest the formation of a coexistence phase in-between. Indeed, the subsystem restricted to $[x_{\min,1}, x_{\min,n}]$ may be regarded as a TASEP with entrance and exit rates $\alpha = \beta = \lambda_{\min}/(1+\sqrt{\ell})$, positioning it at the triple point of the phase diagram, and computing the characteristics reveals one or multiple stationary shock fronts in the interior. Such macroscopic phenomenon in the homogeneous 1-TASEP has previously been associated on the microscopic level with a shock performing a random walk on the lattice with reflecting boundaries [DLS97]. Numerical simulations seem to locate these shock around local maxima disproportionately often (cf. Figure A.3), which might reflect dependencies of its diffusivity on λ.

Applicability to discrete lattices

The existence of a continuous limiting rate function $\lambda : [0,1] \to \mathbb{R}^+$ extending the discrete jump rates $p_k = \lambda(ak)$ is an important ingredient in our treatment of the hydrodynamic limit. That is, in order for density profiles to be accurately approximated by solutions to the PDE (2.3), the p_k must vary smoothly across lattice sites. Microscopic systems like the translation machinery in cells, however, are typically subjected to substantial amounts of fluctuations, resulting in far rougher elongation profiles (see Figure 2.2A). Despite this lack of regularity, the hydrodynamic limit can still be employed to describe local averages of such a system. More precisely, fixing $r \in \{1, \ldots, N\}$, we associate with an elongation rate profile $\{p_1, \ldots, p_N\}$ and the corresponding density profile $\{\rho_1, \ldots, \rho_N\}$ their smoothed profiles $\{\overline{p}_1, \ldots, \overline{p}_{N-r+1}\}$ and $\{\overline{\rho}_1, \ldots, \overline{\rho}_{N-r+1}\}$, respectively, obtained through a moving r-codon average: $\overline{p}_k = \sum_{i=k}^{k+r-1} p_i/r$, and $\overline{\rho}_k = \sum_{i=k}^{k+r-1} \rho_i/r$. Moreover, we define $\{\sigma_1, \ldots, \sigma_{N-r+1}\}$ to be the steady state density profile under the elongation rates $\{\overline{p}_k\}$. If $\{p_k\}$ extends to a smooth $\lambda : [0,1] \to \mathbb{R}^+$, then since $|\overline{p}_k - p_k| \in O(N^{-1})$, $\{\overline{p}_k\}$ extends to this same λ, and hence $\{\rho_k\}, \{\overline{\rho}_k\}$ and $\{\sigma_k\}$ all converge to the solution ρ of (2.3). When $\{p_k\}$ does not extend to a continuous limit, then $\{\rho_k\}$ generally does not either. However, by the same reasoning that establishes the hydrodynamics for the 1-TASEP with quenched disorder [S+99], $\{\overline{\rho}_k\}$ should still be close to $\{\sigma_k\}$, which, due to the greater regularity of $\{\overline{p}_k\}$, is well approximated by the hydrodynamic density profile under $\{\overline{p}_k\}$. Thus, $\{\overline{\rho}_k\}$ is ultimately well approximated by the hydrodynamic limit under $\{\overline{p}_k\}$.

To confirm this, we carried out an extensive simulation study on elongation rate profiles obtained from ribosome profiling data of yeast (see Section A.2 for more details on data). Specifically, we

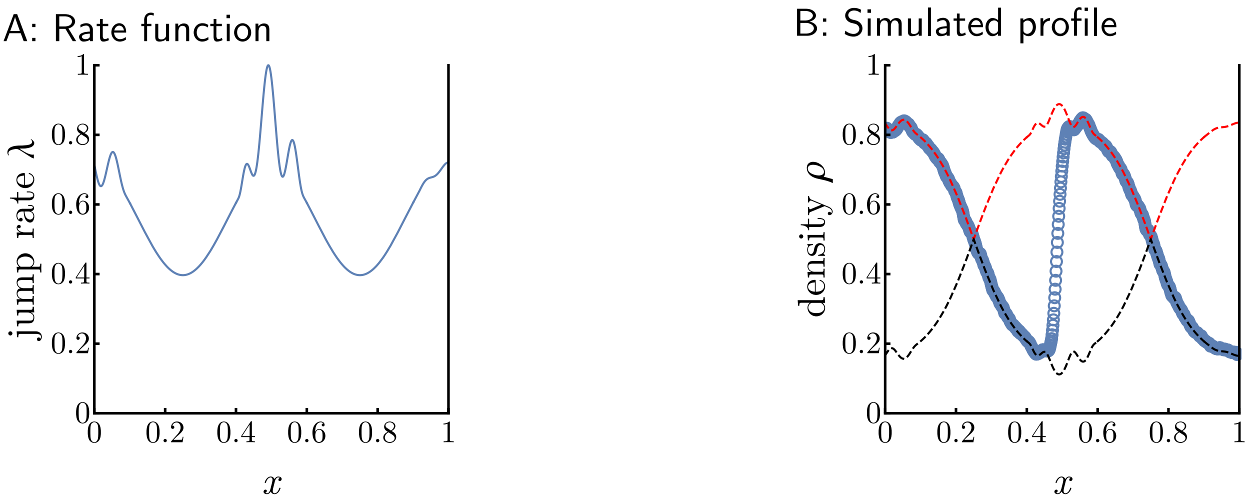


Figure A.3: **Atypical behaviour of MC branch switching in the presence of two global minima.** Hydrodynamic predictions suggest that branch switching is bound to occur between any two global minima, but do not provide explicit information about the precise location of these singularities. Simulations indicate that branch switching is preferentially situated around local maxima. **A:** Elongation rates. **B:** Circles are averaged counts over 5×10^7 Monte-Carlo steps after 10^7 burn-in cycles on a lattice of size $N = 2000$ with parameters $\alpha = \beta = \ell = 1$ and elongation rate function shown in **A** We compare these simulated densities to the theoretical profile obtained from the upper (red) and lower (black) branch solutions (described in (A.10)).

performed the smoothing $\{p_k\} \to \{\overline{p}_k\}$ (Figure 2.2A,B), simulated density profiles $\{\rho_k\}$ under $\{p_k\}$ (Figure 2.2A,C), and compared the corresponding smoothed densities $\{\overline{\rho}_k\}$ with the hydrodynamic prediction under $\{\overline{p}_k\}$ (Figure 2.2D). A choice of $r = 10$, which is equal to the particle (ribosome) size ℓ in translation and the smallest window size guaranteeing smoothness of $\{\overline{p}_k\}$ due to the ℓ-periodicity induced by traffic jams, resulted in excellent agreement both in densities and currents uniformly across transcripts while maintaining local structure.

Boundary conditions

The computation of initial densities in Section A.2 yielded precise boundary values for $x = 0$ in the LD regime and $x = 1$ in the HD regime, respectively. Using the same principle of balancing currents, boundary conditions for all locations in the phase diagram can be computed. The results are listed in Table A.1, which extend previous results obtained in [LC03] (who derived entries (1,1), (2,2) and (2,3) of Table A.1). More precise information about the boundary layers can be gleaned from direct analysis of (A.6) rather than its limit (A.7).

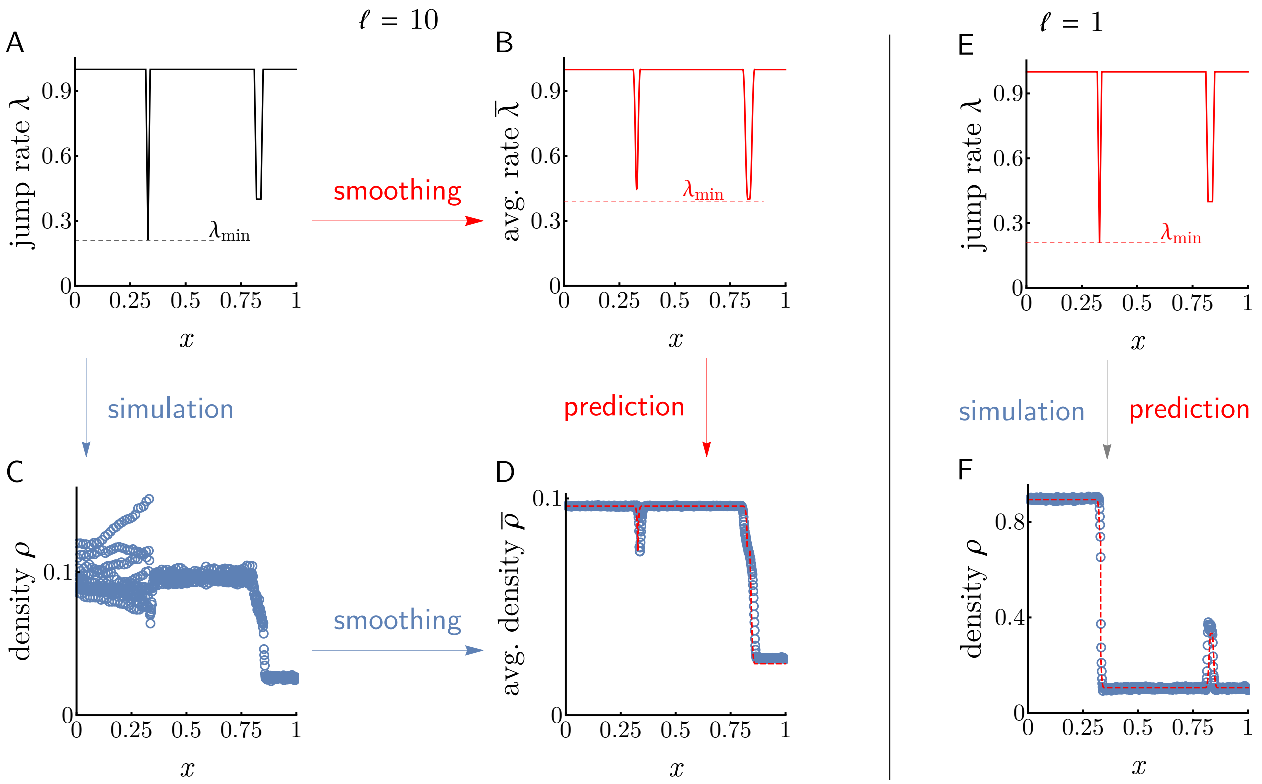


Figure A.4: **MC branch switching is determined by locally averaged elongation rates rather than raw elongation rates (A-D), and the averaging scale depends on the particle size (E,F).** Both the value as well as the location of the minimal elongation rate may differ significantly when measured with respect to the discrete elongation profile (Panel **A**) or smoothed elongation profile (Panel **B**). Panels **C** and **D** demonstrate that our hydrodynamic prediction is very accurate, and show that MC branch switching is governed by the smoothed elongation profile rather than its discrete counterpart. Several of the yeast transcripts we analyzed are affected by this phenomenon, suggesting that a codon's local neighborhood is a stronger determinant of translation dynamics than the absolute elongation rate at that site. Whether smoothed elongation rates (as opposed to unsmoothed rates) describe the translation dynamics more accurately is strongly linked to the particle size (ℓ) and the long-range correlations (in particular, ℓ-periodicity after traffic jams) it induces. To demonstrate this point, we performed the same analysis as in Panels **A-D** using particles of size $\ell = 1$. We found that our hydrodynamic predictions based on the raw, unsmoothed elongation rates (Panel **E**) does indeed provide an accurate approximation of simulated densities (Panel **F**). In short, the fact that the ribosome occupies 10 codons (i.e., the "particle" size is $\ell = 10$) provides another reason (in addition to alleviating the irregularity of elongation rates that cause analytical difficulty) for why smoothing the elongation rates is the right thing to do when applying the hydrodynamic limit of the ℓ-TASEP to study mRNA translation.

Table A.1: **Boundary conditions by phase.** Expected densities at the left ($x = 0$) and right ($x = 1$) end boundaries of the lattice.

Phase	ρ_0	ρ_1^+	ρ_1^-
LD	$\frac{\alpha}{\lambda_0+(\ell-1)\alpha}$	$\frac{1}{\beta}\left[\frac{\alpha(\lambda_0-\alpha)}{\lambda_0+(\ell-1)\alpha}\right]$	$\frac{1}{\lambda_1}\left[\frac{\alpha(\lambda_0-\alpha)}{\lambda_0+(\ell-1)\alpha}\right]$
HD	$\frac{1}{\ell} - \frac{1}{\ell\alpha}\left[\frac{\beta(\lambda_1-\beta)}{\lambda_1+(\ell-1)\beta}\right]$	$\frac{\lambda_1-\beta}{\lambda_1+(\ell-1)\beta}$	$\frac{1}{\lambda_1}\left[\frac{\beta(\lambda_1-\beta)}{\lambda_1+(\ell-1)\beta}\right]$
MC	$\frac{1}{\ell}-\frac{1}{\ell\alpha}\left[\frac{\lambda_{\min}}{(1+\sqrt{\ell})^2}\right]$	$\frac{1}{\beta}\left[\frac{\lambda_{\min}}{(1+\sqrt{\ell})^2}\right]$	$\frac{1}{\lambda_1}\left[\frac{\lambda_{\min}}{(1+\sqrt{\ell})^2}\right]$

Data processing

Initiation, elongation, and termination rates were obtained from an earlier work [DDS18], where the rates were estimated from ribosome profiling data of *S. cerevisiae* for a set of 850 genes selected based on length and footprint coverage. The initiation and termination rates (α and β) were taken directly from that previous work. To compute the elongation rates relevant to the hydrodynamic limit, we applied a ten-codon moving average to their elongation rates (see Section A.2). To demonstrate replicability on larger datasets, we took ribosome profiles directly from [WJW14] and [PRI$^+$14] (combined with polysome profiling from [MLF$^+$04] for normalization purposes, yielding 3098 and 2536 genes, respectively), smoothed them by moving averages of length $\ell = 10$, and inverted the solution of (2.3) to obtain initiation rates, termination rates, and smoothed elongation profiles.

A.3 Agreement between theoretical prediction and simulation

In order to empirically verify our theoretical justification of the hydrodynamic limit, we simulated ribosome profiles and currents for all 850 *S. cerevisiae* genes studied in [DDS18]. For each gene, we considered four conditions: LD, HD, MC, and under the actual initiation and termination rates inferred in [DDS18]; these four conditions correspond to different rows in Figure A.5. Absolute errors in ribosome density profiles and currents (first and last columns of Figure A.5) are accurately predicted across all gene lengths—with a slight increase in prediction accuracy for longer genes (as expected, since the hydrodynamic limit becomes exact in the infinite length limit)—and across all regimes of the phase diagram. Due to two or more bottlenecks occasionally competing on the same transcript (i.e., when $|\{x\,:\,\lambda(x)=\lambda_{\min}\}|>1$, cf., last paragraph of Section A.2), error distributions in MC exhibit heavier tails than in LD and HD. However, overall these outliers do not affect the quality of our theoretical prediction significantly. In particular, correlations between simulated and theoretical transcript-by-transcript quantities—ribosome density profiles and mean occupancies (middle column), as well as currents (last column)—are consistently high, demonstrating good predictive power of our hydrodynamic framework.

In HD, predicted and simulated ribosome density profiles had quite low mean squared differences (second row, first column of Figure A.5), but poor correlation (histograms in second row, second column). This seemingly contradictory result can be explained by typical fluctuations in theoretical density profiles being of the same order as typical fluctuations in the random noise (mean ratio of fluctuations $= 0.037$). That is, generic HD profiles are close to flat, allowing uncorrelated site-by-site noise to substantially reduce overall correlations.

Quantification and statistical analysis

To establish significance of a subset X of genes with respect to a statistic f (e.g., α, J or x_{min}) relative to a background set Y, we performed hypothesis testing on the median m_f of f over samples in X. Under the null distribution of X being drawn uniformly at random, the probability of this test statistic exceeding m equals the probability of a hypergeometric variable with parameters $N = |Y|, K = 2|Y_m|, n = |X|$, where Y_m is the set of genes in Y whose f exceeds m, exceeding $\lfloor |X/2| \rfloor$. This p-value can be computed explicitly. Sets of ribosomal and stress response genes were taken from the Saccharomyces Genome Database [CHA+11].

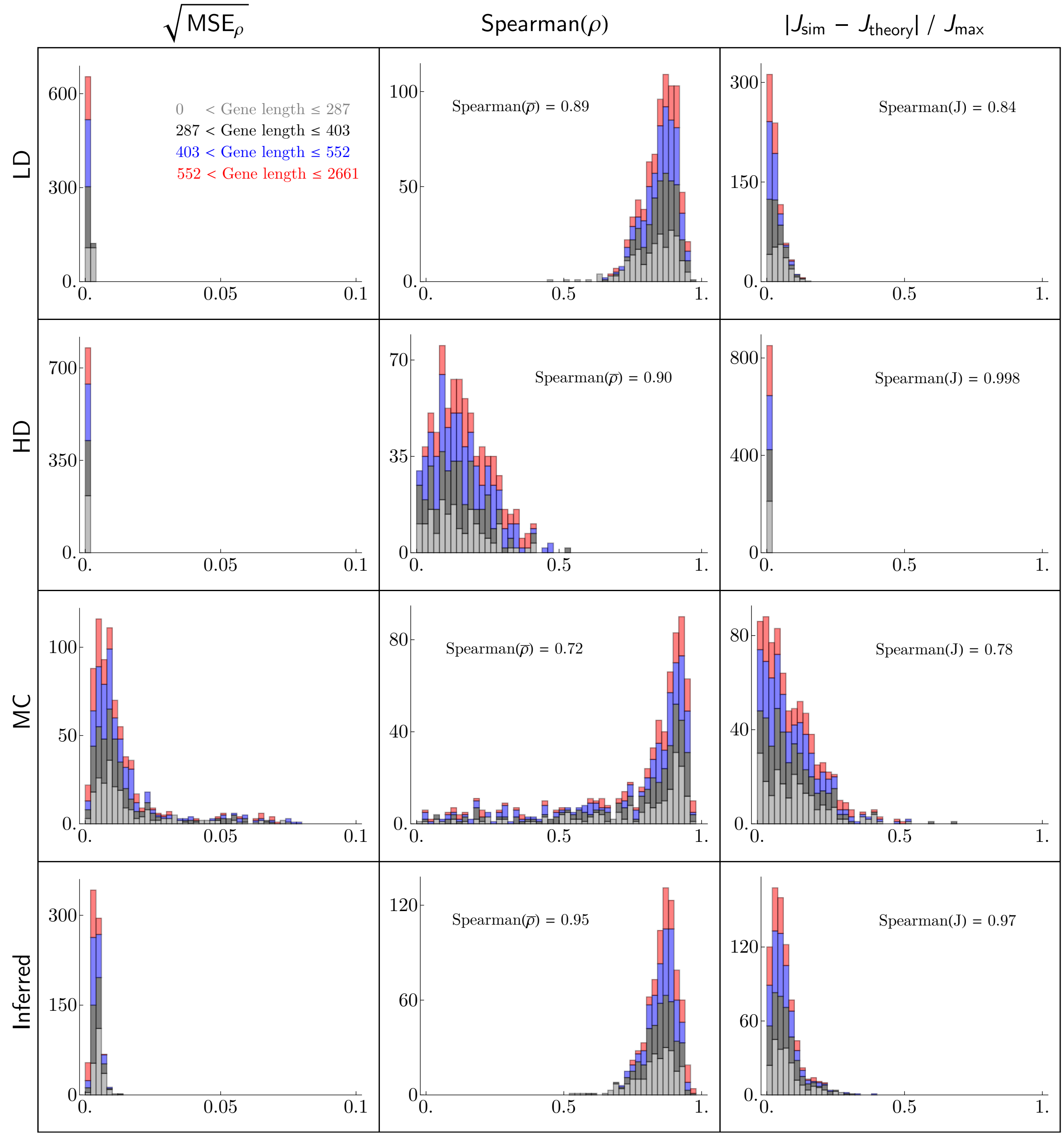


Figure A.5: **Comparison between simulation and theoretical prediction of our hydrodynamic approximation.** Errors in ρ (first column) and J (third column) are low for all gene lengths, for different regimes (first three rows), and for biologically relevant initiation and termination rates (last row) inferred in [DDS18]. Moreover, ρ and mean ribosome occupancies $\overline{\rho}$ correlate well between simulated data and our hydrodynamic predictions (middle column), as do currents (third column).

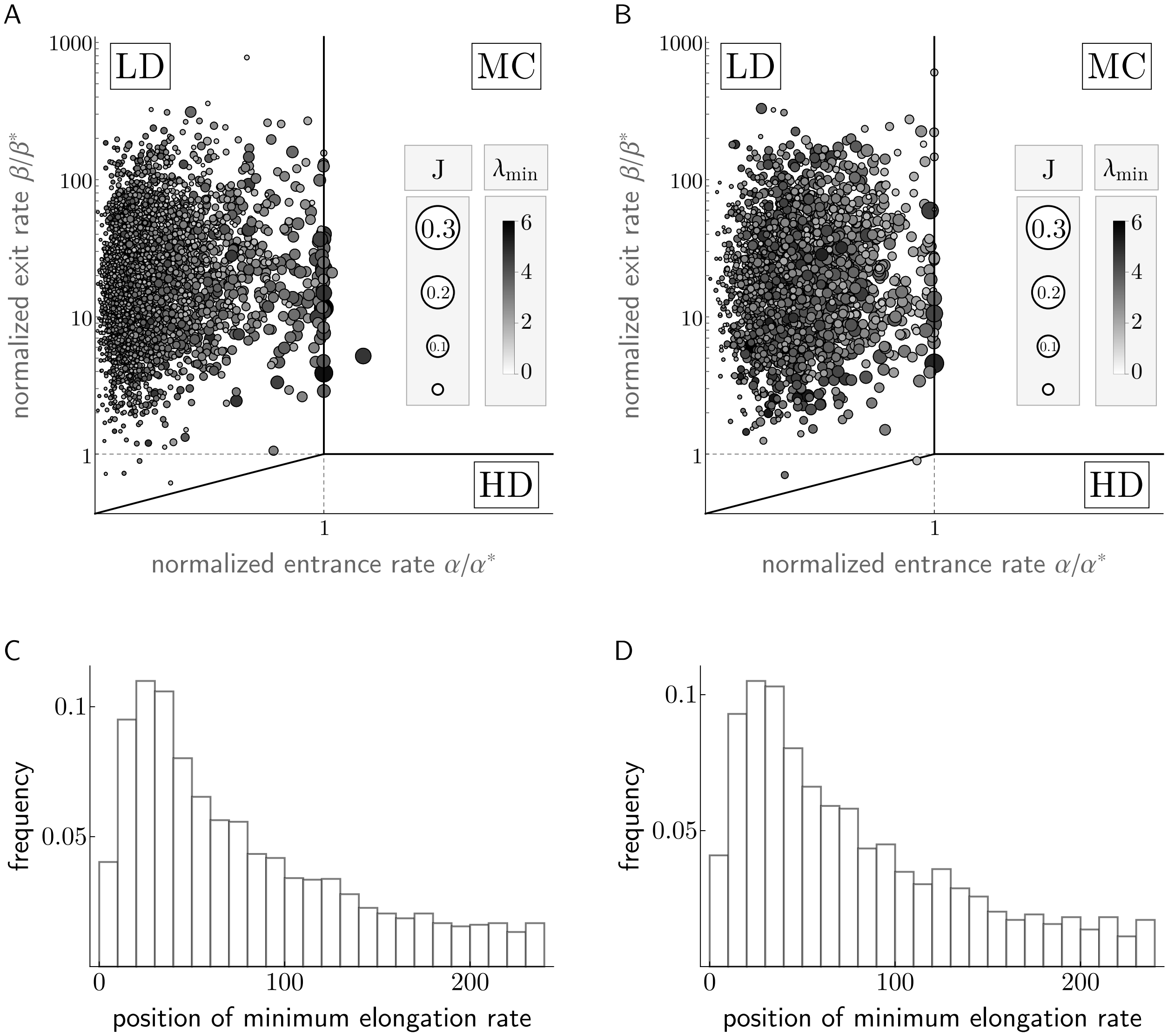


Figure A.6: **Inferences on the efficiency of yeast's translational system are consistent across datasets.** To test the replicability of our analysis using the previously inferred elongation rates in [DDS18] and to exclude any possible systematic biases, we repeated our inference on elongation rates obtained by inverting (2.3) on two independent ribosome profiling datasets: One compiled by [WJW14] (**A, C**, total of 3098 genes), and one by [PRI$^+$14] (**B, D**, total of 2536 genes). The clear localization of genes within LD and at the LD/MC boundary, together with a characteristic ramp-shaped distribution of the minimum elongation location remain apparent, lending support to our proposed design principles holding true not only on the 850 genes analyzed in Chapter 2, but more generally as a framework governing translation efficiency.

Appendix B

Supporting Information: Chapter 3

B.1 Methods

Below we present the mathematical details that justify our design choices in formulating RNA-Sieve. For the reader interested in more high-level guidance on when to use RNA-Sieve and what potential preprocessing steps to take, we compiled Table B.3 as an accessible overview.

Notation

To ease parsing of technical equations, we briefly introduce our notation here. We generally refer to vector quantities with boldfaced lowercase letters, while plain lower- and uppercase symbols are reserved for scalars (or scalar functions) and matrices, respectively. The k^{th} column vector of a matrix $A = \left(a_{ij}\right)_{ij}$ is written as $\boldsymbol{a}_k$, and inner products between vectors $\boldsymbol{v}, \boldsymbol{w}$ are typically denoted $\langle \boldsymbol{v}, \boldsymbol{w} \rangle$. To distinguish observed, random quantities from their underlying deterministic, ground truth objects we add tildes to the former and asterisks to the latter; e.g., $\tilde{\boldsymbol{b}}$ are observed bulk gene expressions, while $\boldsymbol{b}^*$ are the true bulk gene expression means. Estimates of latent parameters carry hats; e.g., $\hat{\boldsymbol{\alpha}}$ is the vector of mixture weights inferred by our deconvolution procedure. Finally, we denote by $[n]$ the set of n elements $\{1, \dots, n\}$, and by $\Delta^{K-1} = \left\{x \in \mathbb{R}^K : \|x\|_1 = 1 \text{ and } x_k \geq 0 \text{ for all } k\right\}$ the $K-1$ dimensional simplex.

Mathematical Model

We assume that for each gene $g \in [G]$ and cell type $k \in [K]$, there exists a distribution $\nu_{g,k}$ describing the expression of gene g in cell type k. As multiple cell types comprise any given organ/tissue, the expression of gene g in a cell drawn at random from a organ/tissue is governed by the mixture distribution

$$\rho_g = \sum_{k=1}^{K} \alpha_k^* \nu_{g,k}, \tag{B.1}$$

where $\boldsymbol{\alpha}^* = (\alpha_k^*)_{k\in[K]} \in \Delta^{K-1}$ contains the proportions of each cell type in the organ/tissue of interest. Despite the *a priori* infinite-dimensional setting, if $G > K$ and ρ_g, $\{\nu_{g,k}\}_{k\in[K]}$ are fully known and sufficiently distinct, the convex combination of (B.1) immediately implies that $\boldsymbol{\alpha}^*$ can be recovered as the unique solution of the finite-dimensional problem

$$\underbrace{\begin{bmatrix} f(\nu_{1,1}) & \dots & f(\nu_{1,K}) \\ \vdots & \vdots & \vdots \\ f(\nu_{G,1}) & \dots & f(\nu_{G,K}) \end{bmatrix}}_{M} \cdot \underbrace{\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix}}_{\boldsymbol{\alpha}} = \underbrace{\begin{bmatrix} f(\rho_1) \\ \vdots \\ f(\rho_G) \end{bmatrix}}_{\boldsymbol{b}}, \tag{B.2}$$

where f is any suitable linear function on the space of probability distributions on $\mathbb{R}$ (i.e., $f(\sum_j w_j \mu_j) = \sum_j w_j f(\mu_j)$ for any convex combination of distributions μ_j). Natural f to consider include point evaluations at $x \in \mathbb{R}$; i.e., $f(\nu) = F_\nu(x)$, where F_ν denotes the cumulative distribution function (CDF) of ν, or its i^{th} moments $f(\nu) = \int x^i \nu(dx)$, both of which enjoy a wealth of statistical theory and proposed estimators.

In experimental settings, exact gene expression distributions are not accessible and instead must be estimated, so utilizing easily and robustly inferrable f becomes crucial. In addition to not having direct access to $\{\rho_g\}_{g\in[G]}$, any analysis is further complicated by the fact that bulk sequencing only yields gene expression levels over whole samples and not for particular cells or cell types. That is, the output is effectively a random variable $X_g = \sum_{i=1}^n X_{g,i}$ where $X_{g,i} \overset{iid}{\sim} \rho_g$ gives the measured expression of gene g aggregated over the $n \in \mathbb{N}$ individual cells comprising the sample. It is thus expedient to choose an f in (B.2) that is not only linear on the space of probability distributions, but also for sums of random variables. The essentially unique such f is the expectation $f(\nu) = \mathbb{E}_{X\sim\nu} X$, which turns (B.2) into

$$M\boldsymbol{\alpha} = \frac{\boldsymbol{b}}{n}, \qquad \text{where} \quad m_{g,k} = \mathbb{E}_{Y\sim\nu_{g,k}} Y \quad \text{and} \quad b_g = \mathbb{E} X_g. \tag{B.3}$$

Incorporating the fact that we only observe noisy bulk samples X_g instead of b_g directly results in

$$\frac{\tilde{\boldsymbol{b}}}{n} = \frac{(\boldsymbol{b} + \varepsilon_{\boldsymbol{b}})}{n} = M\boldsymbol{\alpha} + \frac{\varepsilon_{\boldsymbol{b}}}{n}, \tag{B.4}$$

where $(\varepsilon_{\boldsymbol{b}})_g \sim X_g - b_g = \sum_{i=1}^n (X_{g,i} - b_g/n) \sim \mathcal{N}(0, n \cdot \sigma_g^2(M, \boldsymbol{\alpha}, S))$ for large n by the central limit theorem (CLT), with $\sigma_g^2(M, \boldsymbol{\alpha}, S) := \mathrm{Var}(\rho_g)$ as a function of $M, \boldsymbol{\alpha}$, and $S = \left(s_{g,k}\right)_{g,k} := \mathrm{Var}(\nu_{g,k})$.

Incorporating the dependence of σ_g^2 on $\boldsymbol{\alpha}$: If the dependence of σ_g^2 on $\boldsymbol{\alpha}$ is ignored, (B.4) lends itself to a simple (weighted) non-negative least squares scheme solving

$$M\boldsymbol{\alpha} = \frac{\tilde{\boldsymbol{b}}}{n}. \tag{B.5}$$

This yields a solution $\hat{\boldsymbol{\alpha}}_{LS}$ of roughly $\|\hat{\boldsymbol{\alpha}}_{LS}\|_1 \approx 1$ that simply requires re-scaling to fit onto the simplex. This together with data-driven modifications is the approach pursued in [DTU+20, TDC+19, WPS+19], where it is argued that (B.5) outperforms previous methods.

The first improvement of RNA-Sieve over previous approaches stems from explicitly incorporating the dependence of σ_g^2 on $\boldsymbol{\alpha}$. More concretely, we first make explicit this $\boldsymbol{\alpha}$-dependence by computing

$$\begin{aligned}\sigma_g^2 = \sigma_g^2(M, \boldsymbol{\alpha}, S) = \mathrm{Var}(\rho_g) &= \mathbb{E}_{X\sim\rho_g} X^2 - \left(\mathbb{E}_{X\sim\rho_g} X\right)^2 \\ &= \left(\sum_{k=1}^{K} \alpha_k^* \mathbb{E}_{Y\sim\nu_{g,k}} Y^2\right) - b_g^2 \\ &= \left(\sum_{k=1}^{K} \alpha_k^* \left[s_{g,k} + m_{g,k}^2\right]\right) - b_g^2. \end{aligned} \tag{B.6}$$

The likelihood of observing data $\tilde{\boldsymbol{b}}$ then follows straightforwardly from the central limit theorem:

$$\mathbb{P}_{M,S}^{\boldsymbol{\alpha},n}\left(\tilde{\boldsymbol{b}} \in d\boldsymbol{p}\right) = \prod_{g=1}^{G} \frac{1}{\sqrt{2\pi n \sigma_g^2(M, \boldsymbol{\alpha}, S)}} \exp\left\{\frac{-[p_g - n(M\boldsymbol{\alpha})_g]^2}{2n\sigma_g^2(M, \boldsymbol{\alpha}, S)}\right\}. \tag{B.7}$$

Accounting for uncertainty in the design matrix: The above assumes exact knowledge of the individual distributions $\nu_{g,k}$ (or rather their expectations $m_{g,k}$), which is implausible in experimental settings. Instead, M needs to be estimated from data through some estimator $\tilde{M}$, which we conveniently take to be the sample mean of expression across cells within each cell type, $\tilde{m}_{g,k} = \frac{1}{c_k}\sum_{i=1}^{c_k} C_{g,k}^i$, where $C_{g,k}^i \overset{iid}{\sim} \nu_{g,k}$, and c_k denotes the number of single cells of cell type k. With this additional correction, (B.4) becomes

$$\frac{\tilde{\boldsymbol{b}}}{n} = \tilde{M}\boldsymbol{\alpha} + \frac{\varepsilon_{\boldsymbol{b}}}{n} \quad \text{and} \quad \tilde{M} = M + \varepsilon_M \tag{B.8}$$

where ε_M is a matrix of entries $(\varepsilon_M)_{g,k}$ independently following $\mathcal{N}(0, s_{g,k}/c_k)$ distributions. The second major difference between RNA-Sieve and existing tools (especially those based on least-squares methods) is the correction of the least-squares-type likelihood (B.7) by this stochasticity in the design matrix:

$$\begin{aligned}\mathbb{P}_{M,S}^{\boldsymbol{\alpha},n,\boldsymbol{c}}\left(\tilde{\boldsymbol{b}} \in d\boldsymbol{p}, \tilde{M} \in dO\right) = &\prod_{g=1}^{G} \frac{1}{\sqrt{2\pi n \sigma_g^2(M, \boldsymbol{\alpha}, S)}} \exp\left\{\frac{-[p_g - n(M\boldsymbol{\alpha})_g]^2}{2n\sigma_g^2(M, \boldsymbol{\alpha}, S)}\right\} \\ &\times \prod_{g\in[G],k\in[K]} \frac{1}{\sqrt{2\pi s_{g,k}/c_k}} \exp\left\{\frac{-(o_{g,k} - m_{g,k})^2}{2s_{g,k}/c_k}\right\}. \end{aligned} \tag{B.9}$$

Our method utilizes the likelihood shown in (B.9), the suitability of which depends on a few implicit assumptions that are worth examining. The first is that the large number of cells assayed in an experiment permits us to use asymptotic theory and apply the classical CLT. As a result, we can write down a likelihood for our observations using normal distributions as long as $\mathrm{Var}(\nu_{g,k}) < \infty$, which is true since gene expression profiles are necessarily bounded. Secondly, we suppose that the errors arising from estimating $\boldsymbol{b}$ and M are independent. This is appropriate as the bulk and single-cell experiments are performed separately. We additionally presume that expression levels in different genes are independent, as are those in different cells. It is unclear whether the latter is completely true in practice, though there is little evidence to the contrary. On the other hand, expression levels across genes within samples (either bulk or individual cells) are liable to be somewhat dependent due to expression co-regulation and the nature of the sampling process performed in RNA-seq. Given the large number of genes assayed, the latter co-dependence is apt to be fairly small. Meanwhile, co-expression estimation in single cells remains an open problem independent of deconvolution tasks, and so is not accounted for in RNA-Sieve. Once correlation structure is known however, it is straightforwardly incorporated into the likelihood we propose.

Joint deconvolution of multiple bulk samples: If it is known that multiple bulk gene expression vectors share the same constituent cell type expression profiles, we can gain statistical strength and decrease the computational burden by inferring their mixture proportions jointly rather than individually. Assuming statistical independence of the bulk sample observations, we must simply augment the likelihood in (B.9) by including the $N-1$ additional mixtures in $A = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N) \in \mathbb{R}^{K\times N}, \tilde{B} = (\tilde{\boldsymbol{b}}_1, \ldots, \tilde{\boldsymbol{b}}_N) \in \mathbb{R}^{G\times N}$ and $\boldsymbol{n} = (n_1, \ldots, n_N)$:

$$\mathbb{P}^{\boldsymbol{\alpha},\mathbf{n},\boldsymbol{c}}_{M,S}\left(\tilde{B}\in dP, \tilde{M}\in dO\right) = \prod_{b=1}^{N}\prod_{g=1}^{G} \frac{1}{\sqrt{2\pi n_b \sigma_g^2(M,\boldsymbol{\alpha}_b,S)}} \exp\left\{\frac{-[(\boldsymbol{p}_b)_g - n_b(M\boldsymbol{\alpha}_b)_g]^2}{2n_b\sigma_g^2(M,\boldsymbol{\alpha}_b,S)}\right\}$$
$$\times \prod_{g\in[G],k\in[K]} \frac{1}{\sqrt{2\pi s_{g,k}/c_k}}\exp\left\{\frac{-(o_{g,k}-m_{g,k})^2}{2s_{g,k}/c_k}\right\}. \qquad \text{(B.10)}$$

This increase in power depends solely on the statistical independence of distinct bulk samples rather than their respective cell type compositions. In fact, for the purposes of denoising the reference M, samples of dissimilar compositions are preferable because they provide non-redundant information. Conversely, bulk samples exhibiting heterogeneity in gene expression patterns (e.g., through differential expression) without corresponding reference matrices M amount to model misspecification, and thus may negatively impact inference. This impediment is information-theoretically unavoidable and therefore a challenge for all deconvolution methods. In our particular applications we did not find a strong effect of sample heterogeneity on our results; for instance, simultaneous deconvolution with mice of different ages yielded highly similar results to when we stratified by age. In the case of cell types with strong expression differences across different phenotypes, this may not hold, however.

Data Pre-processing Procedure

Due to the well-known influence of technical variability in scRNA-seq data, we suggest that users of RNA-Sieve perform their own quality control filtering of cells and genes prior to running our software in addition to their preferred normalization. Given the potential complexity of these patterns in general, we feel that manual cleaning is more reliable than automated procedures.

Nonetheless, we implement a simple, largely optional, cell filtering and normalization scheme to ensure the accuracy of results when the user has chosen not to perform their own quality control. Our procedure attempts to do the following:

1. Remove low-quality cells with anomalously low or high library sizes (≥ 3 median absolute deviations away from the median value of total number of reads per cell in each cell type)

2. Normalize read counts in cells (re-scale reads so that all cells have the median number of reads from across all cells);

3. Identify and remove cells which may be mislabeled or are simply extremely different from other cells with the same cell type label (≥ 3 median absolute deviations away from the median value of inter-cellular pairwise distances in each cell type)

4. Identify and retain genes which are expressed sufficiently often ($\geq 20\%$ non-zero measurements in at least one cell type).

We note that the first three steps are optional whereas step 4 is necessary to remove lowly expressed genes, whose presence may result in poor optimization outcomes due to creating biologically implausible expressions (a non-zero bulk expression can never be realized as a convex combination of zero or almost zero, low variance, single cell expressions).

We implemented two additional layers of gene filtering which we found improved robustness to cross-protocol differences in reference and bulk gene expression measurements. The motivation behind these steps is as follows:

1. By virtue of being a convex combination of expression levels from different cell types (under our generative model (B.9)), a gene's *true* expression b_g must necessarily lie between its smallest and largest corresponding expressions $m_{g,k}$ across cell types $k \in [K]$. That is,

$$b_g \in \left[\min_{k \in [K]} m_{g,k}, \max_{k \in [K]} m_{g,k} \right], \tag{B.11}$$

which naturally motivates a filtering scheme based on violations of these constraints. Of course, these inequalities do not necessarily hold in the presence of observational noise, which may push a gene's bulk expression outside of its theoretical extremes. However, a stochastic version of (B.11) persists in that

$$\mathbb{P}_{M,S}^{\boldsymbol{\alpha},n,\boldsymbol{c}} \left[\delta \left(\tilde{b}_g, \left[\min_{k \in [K]} \tilde{m}_{g,k}, \max_{k \in [K]} \tilde{m}_{g,k} \right] \right) \geq t \right] \tag{B.12}$$

decays in t with sub-Gaussian tails (with constants depending on $\left\{ \sigma_{g,k} \right\}_{k \in [K]}$), where $\delta(p,A) = \inf_{a \in A} |p - a|$ is the shortest distance of the point p to a set A. It is thus plausible to filter out all genes for which (B.12) is sufficiently small (in principle, computing the precise tail bounds (B.12) requires access to the true parameter $\boldsymbol{\alpha}$, which prior to deconvolution is not available; however, reasonable upper bounds of (B.12) can be calculated independently of $\boldsymbol{\alpha}$).

2. Gene expression profiles may experience (occasionally drastic) shifts when measured with distinct protocols. For example, mean and variance estimates of some gene expression levels may correlate little, or even not at all, across data generated using Smart-Seq2, UMI-based, or bulk RNA-seq technologies. To identify and remove these genes, we resort to a handful of empirically effective filtering steps. Specifically, we remove a gene if it presents as an outlier (as measured by median absolute deviations from the median) in any of the following summary statistics:

$$T_M(g) = \max_{k\in[K]} \hat{m}_{g,k}, \qquad\qquad T_S(g) = \max_{k\in[K]} \hat{s}_{g,k},$$
$$R_{S/M}(g) = \max_{k\in[K]} \frac{\hat{s}^{\frac{1}{2}}_{g,k}}{\hat{m}_{g,k}} = \max_{k\in[K]} c_V(C_{g,k}), \qquad\qquad R_{\boldsymbol{b}/M}(g) = \min_{k\in[K]} \frac{\hat{b}_g}{\hat{m}_{g,k}}, \tag{B.13}$$

where $c_V(C_{g,k})$ denotes the coefficient of variation associated with expression profiles of gene g in cell type k. While the choice of these summary statistics was primarily guided by empirical considerations, they do reveal intuitively plausible and previously observed patterns: T_M, T_S and $R_{S/M}$ reflect the fact that severe over- or under-expression, or high degrees of variability in expression are not well-preserved across protocols, whereas $R_{\boldsymbol{b}/M}(g) = \hat{b}_g / \max_{k\in[K]} \hat{m}_{g,k}$ directly assesses any abnormal conversion factors between bulk and reference protocols.

In our experience, applying these filters based on (B.12) and (B.13) on top of the basic cell filter retains between $3,000 - 12,000$ genes on which to perform the deconvolution task.

Optimization and Estimation

We estimate $\boldsymbol{\alpha}$, the cell type proportions for a given bulk sample, using the MLE which arises from maximizing (B.9). Given the number of free parameters ($GK + K$ in total, corresponding to $M, \boldsymbol{\alpha}$, and n) and structure of the likelihood, this is non-trivial, with standard optimization schemes commonly failing or returning sub-optimal solutions. On its face, the shape of (B.9) is reminiscent of loss functions appearing in so-called Total Least Squares formulations (see, e.g., [GVL80]), whose minimizers can typically be found through SVD-decompositions. However, the presence of entry-wise uncertainties ε_M and the dependence of $\varepsilon_{\boldsymbol{b}}$ on $\boldsymbol{\alpha}$ render such spectral tools inapplicable to our setting; indeed, the corresponding linear algebraic problem consists of finding low-rank approximations to the concatenation of M and $\boldsymbol{b}$ in a Frobenius norm with $\boldsymbol{\alpha}$-dependent weights, for which no satisfactory theory exists. We thus propose an alternating maximization scheme which iteratively estimates and updates $\boldsymbol{\alpha}$, M, and n (and consequently σ_g^2) via a combination of quadratic programming and gradient descent. Despite the increased computational burden relative to W-NNLS or similar techniques, we find that convergence times remain reasonable, requiring between 15-40 minutes on typical data sets of $10,000+$ genes and 6 cell types using a modern laptop computer. We sketch an overview of our optimization procedure below in Algorithm 1 (where $\mathbb{P}^{\boldsymbol{\alpha},n,\boldsymbol{c}}_{M,S|\boldsymbol{\sigma}^2(M',\boldsymbol{\alpha}',S)}$ refers to (B.9) with $\boldsymbol{\sigma}^2(M,\boldsymbol{\alpha},S)$ kept fixed at $\boldsymbol{\sigma}^2(M',\boldsymbol{\alpha}',S)$).

Algorithm 1: Find MLE of $\boldsymbol{\alpha}$

```
Data: Single cell expression vectors $\{\tilde{\boldsymbol{v}}_{k,i}\}_{k\in[K],i\in[c_k]} \subset \mathbb{R}^G$, bulk gene expression vector
      $\tilde{\boldsymbol{b}} \in \mathbb{R}^G$
Result: Mixture proportions $\{\hat{\alpha}_k\}_{k\in[K]}$ of cell types in the bulk, number of cells $\hat{n} \in \mathbb{R}_+$ in
        bulk, mean expression $\hat{M} \in \mathbb{R}^{G\times K}$ of cell types
1  begin
2      $\tilde{m}_{g,k} \longleftarrow \frac{1}{c_k}\sum_{i=1}^{c_k}(\tilde{v}_g)_{k,i}$, $s_{g,k} \longleftarrow \frac{1}{c_k}\sum_{i=1}^{c_k}\left((\tilde{v}_g)_{k,i} - \tilde{m}_{g,k}\right)^2$
3      $\boldsymbol{\alpha}_0 \longleftarrow \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}_+^K} \|\tilde{M}\boldsymbol{\alpha} - \tilde{\boldsymbol{b}}\|_2^2$, $n_0 \longleftarrow \|\boldsymbol{\alpha}_0\|_1$, $\boldsymbol{\alpha}_0 \longleftarrow \boldsymbol{\alpha}_0/\|\boldsymbol{\alpha}_0\|_1$ $M_0 \longleftarrow \tilde{M}$
4      while $\mathbb{P}_{M_{j+1},S}^{\boldsymbol{\alpha}_{j+1},n_{j+1},\boldsymbol{c}}\left(d\tilde{M}, d\tilde{\boldsymbol{b}}\right) - \mathbb{P}_{M_j,S}^{\boldsymbol{\alpha}_j,n_j,\boldsymbol{c}}\left(d\tilde{M}, d\tilde{\boldsymbol{b}}\right) > \delta$ do
5          $M_{j+1} \longleftarrow \arg\max_M \mathbb{P}_{M,S|\boldsymbol{\sigma}^2(M_j,\boldsymbol{\alpha}_j,S)}^{\boldsymbol{\alpha}_j,n_j,\boldsymbol{c}}\left(d\tilde{M}, d\tilde{\boldsymbol{b}}\right)$
6          $\boldsymbol{\alpha}_{j+1} \longleftarrow \arg\max_{\boldsymbol{\alpha}\in\Delta^{K-1}} \mathbb{P}_{M_{j+1},S|\boldsymbol{\sigma}^2(M_{j+1},\boldsymbol{\alpha}_j,S)}^{\boldsymbol{\alpha},n_j,\boldsymbol{c}}\left(d\tilde{M}, d\tilde{\boldsymbol{b}}\right)$
7          $n_{j+1} \longleftarrow \arg\max_{n\in\mathbb{R}_+} \mathbb{P}_{M_{j+1},S}^{\boldsymbol{\alpha}_{j+1},n,\boldsymbol{c}}\left(d\tilde{M}, d\tilde{\boldsymbol{b}}\right)$
8      end
9      $(\boldsymbol{\alpha}_\ell, M_\ell, n_\ell) \longleftarrow$ Last $(\boldsymbol{\alpha}_j, M_j, n_j)$ iterate returned in line 7
10     while $\mathbb{P}_{M_{\ell+1},S}^{\boldsymbol{\alpha}_{\ell+1},n_{\ell+1},\boldsymbol{c}}\left(d\tilde{\Omega}, d\tilde{\boldsymbol{b}}\right) - \mathbb{P}_{M_\ell,S}^{\boldsymbol{\alpha}_\ell,n_\ell,\boldsymbol{c}}\left(d\tilde{\Omega}, d\tilde{\boldsymbol{b}}\right) > \delta$ do
11         $M_{\ell+1} \longleftarrow \arg\max_M \mathbb{P}_{M,S}^{\boldsymbol{\alpha}_\ell,n_\ell,\boldsymbol{c}}\left(d\tilde{\Omega}, d\tilde{\boldsymbol{b}}\right)$
12         $\boldsymbol{\alpha}_{\ell+1} \longleftarrow \arg\max_{\boldsymbol{\alpha}\in\Delta^{K-1}} \mathbb{P}_{M_{\ell+1},S}^{\boldsymbol{\alpha},n_\ell,\boldsymbol{c}}\left(d\tilde{\Omega}, d\tilde{\boldsymbol{b}}\right)$
13         $n_{\ell+1} \longleftarrow \arg\max_{n\in\mathbb{R}_+} \mathbb{P}_{M_{\ell+1},S}^{\boldsymbol{\alpha}_{\ell+1},n,\boldsymbol{c}}\left(d\tilde{M}, d\tilde{\boldsymbol{b}}\right)$
14     end
15     $(\hat{\boldsymbol{\alpha}}, \hat{M}, \hat{n}) \longleftarrow$ Last iterate returned in line 13
16     $\hat{\boldsymbol{\alpha}} \longleftarrow \arg\min_{\boldsymbol{\alpha}\in\Delta^{K-1}} \|\hat{M}\boldsymbol{\alpha} - \tilde{\boldsymbol{b}}\|_{\boldsymbol{\sigma}^2(\hat{M},\hat{\boldsymbol{\alpha}},S)}^2$
17     Return $(\hat{\boldsymbol{\alpha}}, \hat{M}, \hat{n})$.
18 end
```

Implementations of Algorithm 1 are currently available in Python and Mathematica. While these implementations differ slightly, they both agree on the following fundamental design choices:

1. Instead of maximizing $\mathbb{P}_{M,S}^{\boldsymbol{\alpha},n,\boldsymbol{c}}$, we minimize $-\log\mathbb{P}_{M,S}^{\boldsymbol{\alpha},n,\boldsymbol{c}}$, rendering lines 5 and 6 as quadratic programs which can be solved efficiently.

2. Lines 7 and 13 can be solved explicitly by differentiating (B.9) and finding the zeros of the resulting algebraic fractions in n. Thus, these steps do not require any explicit optimization scheme.

3. The optimizations in lines 11 through 13 proceed via gradient descent (or a variation thereof), and so could possibly require long runtimes. However, the coarser maximization (minimization, cf. item 1) in lines 5–7 typically improves the objective function to such an extent that

only two or three more iterations are required. Moreover, both sets of optimizations are amenable to parallelization.

4. Algorithm 1 straightforwardly generalizes to the setting of jointly inferring mixture proportions in an arbitrary number N of bulk samples (cf., the remark in Section B.1). Both of our implementations support this generalized deconvolution.

Lastly, we note that although the alternating optimization in lines 5–7 is not guaranteed to converge, the second round of maximization in lines 11–13 is a proper coordinate descent and is therefore guaranteed to reach a local minimum.

Confidence Intervals

As indicated in Section 3.2, the explicit generative modeling of (B.9) allows us to not only compute precise point estimators of $\boldsymbol{\alpha}$ and n, but also to quantify this precision through confidence regions. More concretely, since our model is well-behaved in the sense of satisfying all assumptions in, say, Theorem 9.14 of [Kee11], we expect our estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{n}$ to be distributed normally around the true configuration $\boldsymbol{\alpha}^*, n^*$ with covariance matrix given by the inverse of the Fisher information $I^{\boldsymbol{c}}_{M,S}(\boldsymbol{\alpha}^*, n^*) \approx I^{\boldsymbol{c}}_{\hat{M},S}(\hat{\boldsymbol{\alpha}}, \hat{n})$. Given such asymptotic normality, it is straightforward to construct both marginal confidence intervals (from the diagonal entries of $\left[I^{\boldsymbol{c}}_{\hat{M},S}\right]^{-1}$) as well as K-dimensional confidence regions around $\hat{\boldsymbol{\alpha}}$. Generically, there are infinitely many possibilities for choosing such confidence regions from $I^{\boldsymbol{c}}_{\hat{M},S}(\hat{\boldsymbol{\alpha}}, \hat{n})$, which we acknowledge by providing the entire (inverse) Fisher information to the user to allow computation of the confidence volume of their choosing. One option is to calculate the canonical (that is, Lebesgue volume-minimizing) q-confidence region $C_q = \{\boldsymbol{\alpha} \in \mathbb{R}^{K-1}_{+} : \sum_{k=1}^{K-1} a_k \leq 1, \|\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\|^2_{I^{\boldsymbol{c}}_{\hat{M},S}} \leq F^{-1}_{\chi^2_{K-1}}(q)\}$, where $\|\boldsymbol{v}\|_\Sigma = \langle \boldsymbol{v}, \Sigma^{-1}\boldsymbol{v} \rangle$ is the Mahalanobis norm of $\boldsymbol{v}$ associated with covariance matrix Σ, $F_{\chi^2_{K-1}}$ denotes the CDF of a χ^2 variable with $K-1$ degrees of freedom, and where we reparameterize $\boldsymbol{\alpha}$ in order to account for the simplex constraint in our computation of the Fisher information matrix; this option is included as the default.

Confidence intervals derived in this manner are, as a consequence of the aforementioned Theorem 9.14 in [Kee11], necessarily well-calibrated *if* data adhere to our generative model (B.9). As we have observed in Section A.2, this may not hold when the use of different protocols in the reference and bulk experiments induce significant distributional shifts. Nonetheless, we can still provide conservative, yet well-calibrated, confidence regions generalizing the Fisher information $I^{\boldsymbol{c}}_{\hat{M},S}(\hat{\boldsymbol{\alpha}}, \hat{n})$ to the Godambe information matrix $G^{\boldsymbol{c}}_{\hat{M},S}(\hat{\boldsymbol{\alpha}}, \hat{n})$ of the data [God60]. If protocol mismatches result in the true generating distribution $\mathbb{Q}$ of the data not lying within our model family $\mathscr{M} = \left\{\mathbb{P}^{\boldsymbol{\alpha},n,\boldsymbol{c}}_{M,S}\right\}_{M,\boldsymbol{\alpha},n\in\Theta}$, where $\Theta \subset \mathbb{R}^{G\cdot K+(K-1)+1}$ is the space of all possible parameter configurations, then $G^{\boldsymbol{c}}_{\hat{M},S}(\hat{\boldsymbol{\alpha}}, \hat{n})$ describes the Gaussian fluctuations of $(\hat{\boldsymbol{\alpha}}, \hat{n})$ around the relative-entropy-projection of $\mathbb{Q}$ onto $\mathscr{M}$; i.e., $(\boldsymbol{\alpha}^\pi, n^\pi, M^\pi) = \arg\min_{\boldsymbol{\alpha},n,M\in\Theta} KL(\mathbb{Q} \parallel \mathbb{P}^{\boldsymbol{\alpha},n,\boldsymbol{c}}_{M,S})$, where $KL(\nu \parallel \mu)$ denotes the relative entropy (also known as Kullback-Leibler divergence) between two probability distributions ν and μ. Thus confidence regions for $\hat{\boldsymbol{\alpha}}$ based on $G^{\boldsymbol{c}}_{\hat{M},S}(\hat{\boldsymbol{\alpha}}, \hat{n})$ are still well-calibrated, assuming that $\boldsymbol{\alpha}^* \approx \boldsymbol{\alpha}^{\boldsymbol{\pi}}$, which is plausible given that distributional shifts induced by protocol differences appear to affect expression means (the entries of M) primarily through global scaling. In the absence of distributional mismatches, the Godambe information matrix $G^{\boldsymbol{c}}_{\hat{M},S}(\hat{\boldsymbol{\alpha}},\hat{n})$ automatically collapses to the (empirical) Fisher information matrix $I^{\boldsymbol{c}}_{\hat{M},S}(\hat{\boldsymbol{\alpha}},\hat{n})$, and so our confidence region estimation proceeds through $G^{\boldsymbol{c}}_{\hat{M},S}(\hat{\boldsymbol{\alpha}},\hat{n})$ in both within- and cross-protocol settings by default, though can be adjusted manually if so desired. Occasionally, especially when constituent cell types are closely related to each other, the resulting covariance matrices may be nearly singular, in which case their inversion poses computational difficulties. To sidestep potential numerical instabilities, we subsample genes according to their incurred residual values. This reduces the probability of collinearities and produces more well-behaved confidence intervals. Simulations both across and within protocols confirm the utility of our confidence regions, and therefore the validity of the $\boldsymbol{\alpha}^{\boldsymbol{\pi}} \approx \boldsymbol{\alpha}^*$ assumption, assessed in this manner (cf. Figure 3.7). Namely, our estimated 95%-confidence intervals contain the true cell type proportions 96.7% and 91.8% of the time in the *Tabula Muris Senis* pseudobulks, 95.8% of the time in the Monaco et al. bulk samples, and 90.3% of the time in the Newman et al. bulk samples.

A Note on n

One of the parameters inferred by our model is n, the number of cells in the bulk sample. This parameter is accurate and physically meaningful in within-protocol deconvolutions, or cross-protocol experiments where relative amplification factors are explicitly known. However, it loses interpretability when the relative scales across protocols are unclear, and so we sought to verify that both our inferred proportions and computed confidence intervals are robust even in such situations. Numerical experiments in which we re-scaled the bulk samples to artificially manipulate the inference of n showed no degradation in performance over a wide range of values, providing additional support beyond the observed high-quality results in both *in silico* and real bulk cross-protocol deconvolutions (Figure B.10). Moreover, theoretical computations suggest a fairly weak dependence of confidence intervals on n, which are instead driven primarily by the total number of genes available for deconvolution.

Comparison of runtimes

We found that all considered deconvolution algorithms could be successfully run in no more than a few hours on a laptop computer for the data sets we considered. RNA-Sieve runtimes ranged from 15-40 minutes, as did those of Scaden. Because of the straightforward manner in which we construct the signature and variance matrices for cell types, RNA-Sieve's runtime is not sensitive to the size of the scRNA-seq reference. This is not the case for DWLS, whose runtime we found grew quickly with the size of the data set due to the model fitting involved in its signature gene inference procedure. For most cases, DWLS runtimes were also in the 15-40 minute range, but for some of the larger single-cell reference panels with many cell types, the runtime could extend to a few hours. CIBERSORTx typically ran in 5-15 minutes. The remaining methods (SCDC, MuSiC, Bisque, and NNLS) were quite fast, with runtimes of no more than a couple of minutes, though SCDC and MuSiC may take a few extra minutes if their tree-based deconvolution modes are used.

Benchmarking procedures

We employed two distinct approaches to benchmark computational deconvolution methods. The first, the construction of so-called “pseudobulk" experiments, is a common strategy which aggregates scRNA-seq measurements across cells in order to construct gene expression mixtures which we treat as bulk samples with known ground-truth cell type proportions. For this task, we used data from the *Tabula Muris Senis* Consortium which covers many organs/tissues and cell types in two different single-cell experimental protocols–Smart-Seq2 and 10x Genomics Chromium. Specifically, we utilized bladder, kidney, large intestine, limb muscle, liver, lung, mammary gland, marrow, pancreas, skin, thymus, tongue, and trachea in these *in silico* experiments (see Table B.1). For each tissue, four different deconvolutions were performed. For cross-protocol deconvolutions, one in which the reference came from Smart-Seq2 data with 10x Chromium pseudobulk and one in the reverse configuration. For within-protocol deconvolutions, the reference and pseudobulk were built using (non-overlapping) cells from the same protocol. For all pseudobulk deconvolution scenarios, a single reference set and pseudobulk was constructed. All eligible cells from each protocol were used. For scRNA-seq data from *Tabula Muris Senis*, the cell filtering procedure described in Section A.2 was applied.

Our second approach exploited the availability of bulk RNA-seq data sets with knowledge about true cell type proportions. For the PBMC and neutrophil scRNA-seq data sets, cells were filtered after manual inspection. Due to the large number of neutrophils in the available data set, 250 cells were randomly sampled from one individual for use with the *Newman et al.* reference and 1250 across three individuals were randomly sampled for use with the 10x Genomics reference. We considered four different scenarios:

1. Breast cancer and fibroblast cell lines and mixture from [DTU+20];
2. Reference PBMCs and neutrophils from [NSL+19] and [XSW+20], respectively, with bulk whole blood from [NSL+19];
3. Reference PBMCs and neutrophils from 10x Genomics and [XSW+20], respectively, with bulk whole blood from [MLX+19];
4. Pancreatic islets from [XKO+16] and [FVL+14].

The same data were used for all algorithms in each deconvolution, and all were run as described in their respective tutorials using default settings unless otherwise noted. When MuSiC was run, NNLS results were taken from MuSiC’s implementation; otherwise, the DWLS implementation was used. The corresponding scRNA-seq and bulk RNA-seq data files will be available at the Song Lab GitHub repository: `https://github.com/songlab-cal/rna-sieve`.

We chose to utilize the L_1, L_2, and L_∞ distances, in addition to the Kullback-Leibler (KL) divergence, as our performance metrics for their ease of interpretation and ability to capture different aspects of algorithm performance. While the L_1 and L_2 distances, which we further average across cell types, are related to common notions of error such as the mean absolute deviation and root mean square error, the L_∞ distance measures the largest difference between true and inferred values across all cell types and quantifies the worst-case performance in a deconvolution task. The KL divergence is a popular manner by which to compare probability distributions and so fits nicely with the deconvolution setting in which cell type proportions can be thought of as the sampling probability for an individual cell. It is also more sensitive to rarer cell types than the other considered metrics. We specifically chose to compute $KL(\hat{\boldsymbol{\alpha}} \parallel \boldsymbol{\alpha}^*)$ because it corresponds to the false positive rate when testing $H_0 : \boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ against $H_1 : \boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$ through a likelihood ratio test, and so is more relevant than $KL(\boldsymbol{\alpha}^* \parallel \hat{\boldsymbol{\alpha}})$.

B.2 Supporting Tables and Figures

Organs	**# cell types**	**Cell types**
Bladder	2	bladder cell, bladder urothelial cell
Kidney	7	B cell, epithelial cell of proximal tubule, fenestrated cell, kidney collecting duct principal cell, kidney loop of Henle ascending limb epithelial cell, macrophage, T cell
Large intestine	3	enterocyte of epithelium of large intestine, epithelial cell of large intestine, intestinal crypt stem cell
Limb muscle	6	B cell, endothelial cell, macrophage, mesenchymal stem cell, skeletal muscle satellite cell, T cell
Liver	5	B cell, endothelial cell of hepatic sinusoid, hepatocyte, Kupffer cell, myeloid leukocyte
Lung	12	adventitial cell, B cell, bronchial smooth muscle cell, CD4+ $\alpha\beta$ T cell, CD8+ $\alpha\beta$ T cell, classical monocyte, fibroblast of lung, myeloid dendritic cell, neutrophil, natural killer cell, non-classical monocyte, vein endothelial cell
Mammary gland	3	basal cell, luminal epithelial cell of mammary gland, stromal cell
Marrow	9	granulocyte, granulocytopoietic cell, immature B cell, late pro-B cell, macrophage, megakaryocyte-erythroid progenitor cell, naive B cell, precursor B cell, promonocyte
Pancreas	3	pancreatic A cell, pancreatic B cell, pancreatic D cell
Skin	2	basal cell of epidermis, epidermal cell
Thymus	2	DN4 thymocyte, thymocyte
Tongue	2	basal cell of epidermis, keratinocyte
Trachea	5	basal epithelial cell of tracheobronchial tree, chondrocyte, endothelial cell, fibroblast, macrophage

Table B.1: **Cell types for each organ in pseudobulk experiments.** These were the cell types used in pseudobulk experiments with the *Tabula Muris Senis* data. The order in which they are listed here matches their order in any figures based off of these experiments.

(a) Smart-seq2 reference and 10x Chromium pseudobulk

	RNA-Sieve	Bisque	CIBERSORTx	DWLS	MuSiC	NNLS	Scaden	SCDC
Bladder	0.081	**0.047**	0.082	0.072	0.106	0.378	0.099	0.113
Kidney	0.095	0.109	**0.028**	0.055	0.110	0.249	0.062	0.083
Large intestine	0.076	0.082	0.300	0.123	0.108	0.226	**0.042**	0.136
Limb muscle	0.199	0.108	0.037	0.039	0.199	0.310	**0.030**	0.144
Liver	0.137	0.129	0.030	0.054	0.139	0.340	0.076	**0.027**
Lung	0.056	0.078	0.071	0.064	0.056	0.149	0.057	**0.029**
Mammary gland	**0.020**	0.258	0.072	0.029	0.047	0.371	0.083	0.058
Marrow	0.061	0.101	0.071	0.073	0.070	0.166	0.072	**0.049**
Pancreas	**0.011**	0.029	0.050	0.030	0.067	0.130	0.059	0.067
Skin	**0.019**	0.270	0.048	0.123	0.098	0.462	0.182	0.128
Thymus	**0.017**	0.050	0.098	0.331	0.127	0.482	0.030	0.120
Tongue	**0.016**	0.289	0.068	0.293	0.047	0.448	0.217	0.017
Trachea	0.108	**0.097**	0.166	0.165	0.142	0.252	0.110	0.154

(b) 10x Chromium reference and Smart-seq2 pseudobulk

	RNA-Sieve	Bisque	CIBERSORTx	DWLS	MuSiC	NNLS	Scaden	SCDC
Bladder	**0.002**	0.066	0.059	0.085	0.156	0.044	0.167	0.261
Kidney	0.082	0.045	0.036	**0.028**	0.113	0.173	0.044	0.046
Large intestine	0.117	0.158	0.089	0.152	0.186	0.448	0.066	**0.007**
Limb muscle	0.137	0.132	0.037	**0.013**	0.122	0.142	0.102	0.177
Liver	0.107	0.056	**0.032**	0.050	0.126	0.164	0.052	0.070
Lung	0.092	0.069	0.029	**0.021**	0.130	0.153	0.074	0.045
Mammary gland	**0.009**	0.244	0.062	0.013	0.160	0.228	0.196	0.157
Marrow	**0.070**	0.110	0.124	0.097	0.113	0.147	0.111	0.121
Pancreas	0.121	0.085	0.137	0.054	**0.023**	0.117	0.111	0.173
Skin	**0.037**	0.191	0.162	0.050	0.168	0.676	0.109	0.192
Thymus	**0.002**	0.110	0.114	0.208	0.298	0.317	0.036	0.275
Tongue	**0.006**	0.254	0.022	0.143	0.672	0.672	0.191	0.672
Trachea	0.092	0.098	**0.067**	0.080	0.166	0.151	0.105	0.153

Table B.2: **Deconvolution errors for different algorithms in pseudobulk experiments.** Deconvolutions were performed using the specified methods in thirteen organs using both Smart-seq2 and 10x Chromium data from the *Tabula Muris Senis* experiment. Presented errors show the L_1 distance between the ground truth and inferred values divided by the number of present cell types. These values correspond to Table 1 and Figure 2 of Chapter 3.

Data attribute	**RNA-Sieve requirements**
Cell counts	The asymptotic analysis of RNA-Sieve relies primarily on the Central Limit Theorem, and so any cell counts that allow its application are sufficient. Because most gene expression counts reasonably follow Poisson or negative binomial distributions, having at least 30 cells is typically sufficient for accurate approximations. Unusually skewed distributions may necessitate ∼100-400 cells. X_g, necessary cell counts can be computed through the Berry-Esséen Theorem which (conservatively) bounds the deviation from Gaussian tail probabilities by $0.475 \cdot \mathbb{E}\vert X_g\vert^3 / \left(\sqrt{\#\text{ cells}} \cdot \mathrm{Var}^{3/2} X_g\right)$. The above reasoning equally applies to the number of cells in the reference matrix (denoted c in Chapter 3) as well as the bulk cell count (n).
Number of reference individuals	RNA-Sieve does not rely on the presence of multiple individuals in the reference and performs inference reliably with any number of individuals. If multiple reference individuals are available, RNA-Sieve simply operates on the pooled mean and variance matrices. We currently do not recommend mixing data from different experimental protocols in the reference.
Reference and bulk protocols	In the case of differences in the data due to protocol mismatch in the scRNA-seq reference and bulk samples, potential nonlinear distributional shifts may need to be accounted for (linear differences are absorbed into the inference of n, see the A NOTE ON n section in Appendix B.1). Empirically, we found such the largest driver of such nonlinear shifts to be differences in the rates of null inflation. In some cases, this is compensated for by increased sequencing depth. Thus, deeply sequenced libraries can be analyzed without further correction, while sparser data sets may benefit substantially from the filtering steps detailed in the DATA PREPROCESSING PROCEDURE section of Appendix B.1.
Jointly deconvolving multiple bulks	If each cell type is expressed similarly across bulk samples (i.e., cells are not differentially expressed in different bulk samples), joint deconvolution is recommended as it increases statistical power regardless of any heterogeneity in mixture proportions. If cell types display differential expression (due to biological or technical reasons), model misspecification becomes a concern and inference results may depend on the nature of the misspecification. In such cases, it is advisable to deconvolve different bulk samples separately.

Table B.3: **Guidance on RNA-Sieve usage across diverse data sets.** RNA-Sieve's is based on a generative model in an asymptotic regime. The mild criteria outlined above guarantee that the data to be deconvolved behaves in accordance with this generative model.

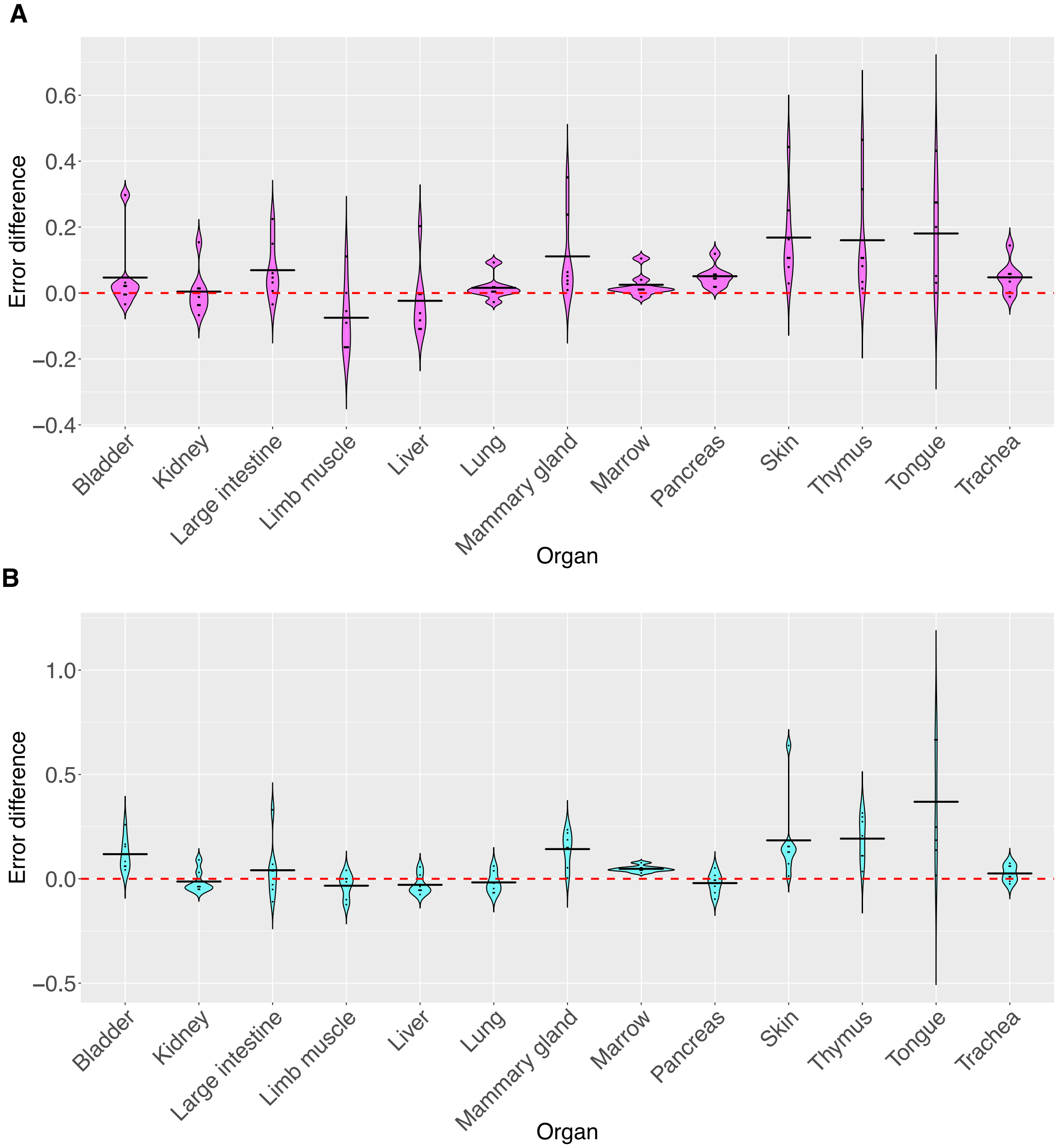


Figure B.1: **Comparison of other methods to RNA-Sieve across 13 murine organs.** Pseudobulk experiments were performed using data from the *Tabula Muris Senis* experiment, and average L_1 errors across cell types computed. For each organ, the difference in errors was computed between other methods and RNA-Sieve. **A**: Smart-seq2 reference, 10x Chromium pseudobulk; **B**: 10x Chromium pseudobulk, Smart-seq2 pseudobulk. Horizontal black bars correspond to the mean difference in error, and positive values indicate better comparative performance for RNA-Sieve.

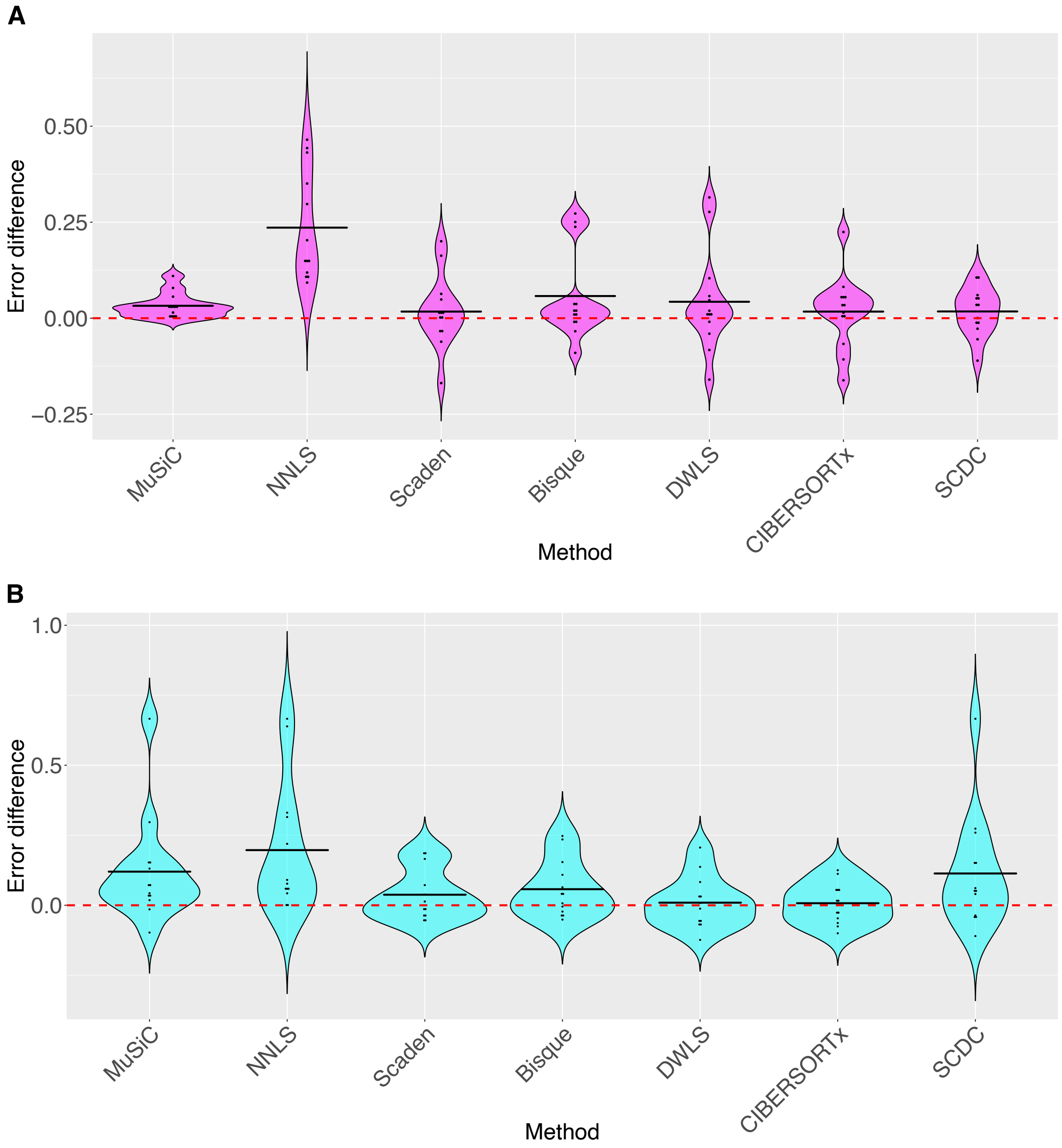


Figure B.2: **Direct comparison of other methods to RNA-Sieve.** Pseudobulk experiments were performed in 13 different organs using data from the *Tabula Muris Senis* experiment. Errors were computed as the average L_1 error across cell types in each organ. For each method, the difference in errors was computed between it and RNA-Sieve across each of the 13 organs. **A**: Smart-seq2 reference, 10x Chromium pseudobulk; **B**: 10x Chromium pseudobulk, Smart-seq2 pseudobulk. Horizontal black bars correspond to the mean difference in error, and positive values indicate better comparative performance for RNA-Sieve.

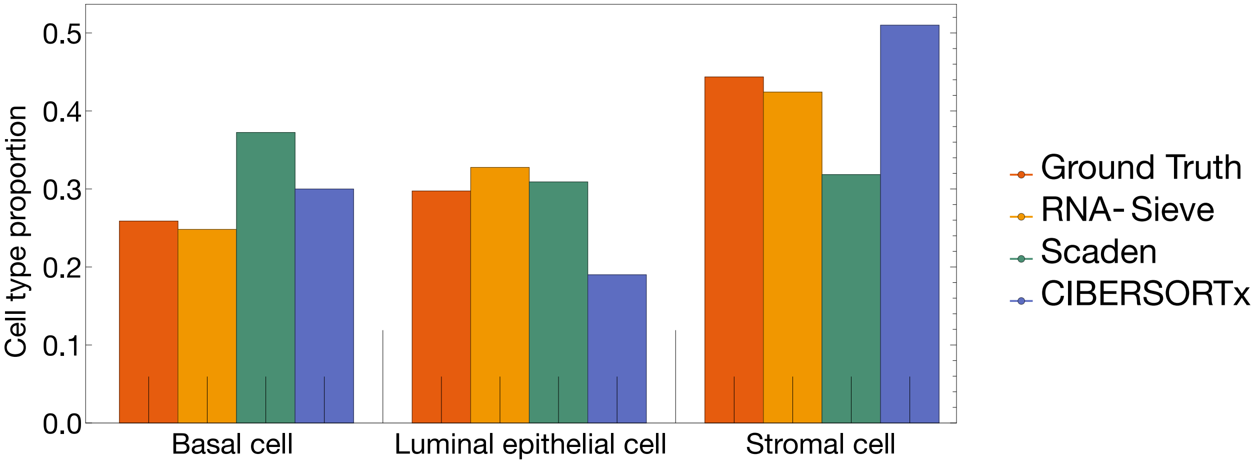


Figure B.3: **Minor per-cell-type differences may result in major individual-cell-type deviations.** The average improvement of RNA-Sieve artificially appears minor because of our chosen error metric (average deviation from the true values) and averaging across cell types. This can be seen in the above (real) example of deconvolving a 10x mammary gland bulk from a Smart-seq2 reference in which RNA-Sieve (0.02), Scaden (0.08), and CIBERSORTx (0.07) may appear to perform similarly when only the raw error values are compared. However, closer inspection reveals that Scaden and CIBERSORTx exhibit large errors for some cell types whereas RNA-Sieve does not.

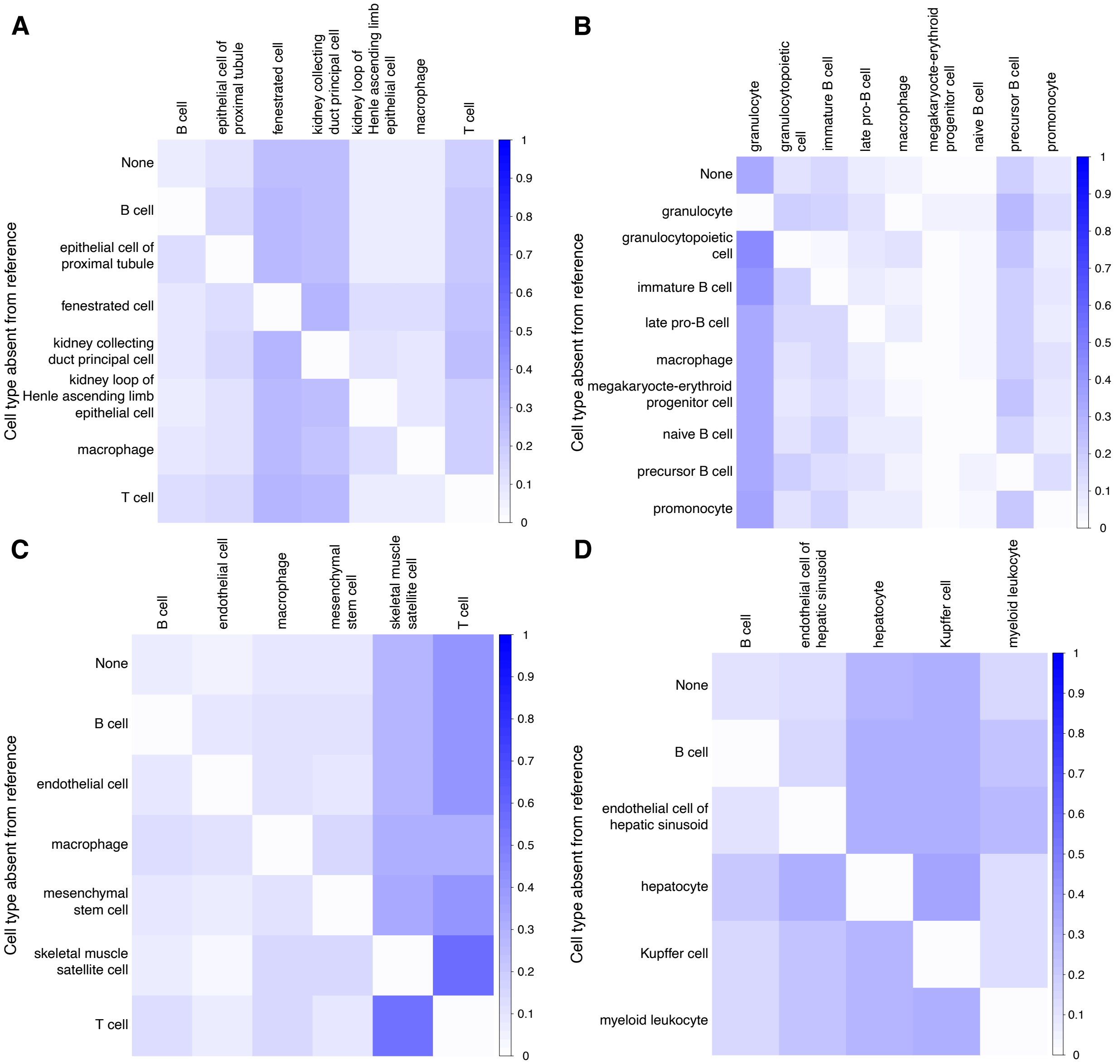


Figure B.4: **Deconvolution with extra cell types in the reference matrix.** Deconvolution was performed in pseudobulk experiments in four different organs (**A** – Kidney; **B** – Marrow; **C** – Limb muscle; **D** – Liver). For each organ, we followed a leave-one-out procedure in which one cell type is removed from the pseudobulk at a time. Deconvolution was then performed with this extra cell type in the reference in order to examine RNA-Sieve's specificity. The top row shows the inferred proportions with no extra reference cell types. Darker colors indicate a higher estimated proportion value. Here we used 10x Chromium data for the reference and Smart-seq2 for the pseudobulk.

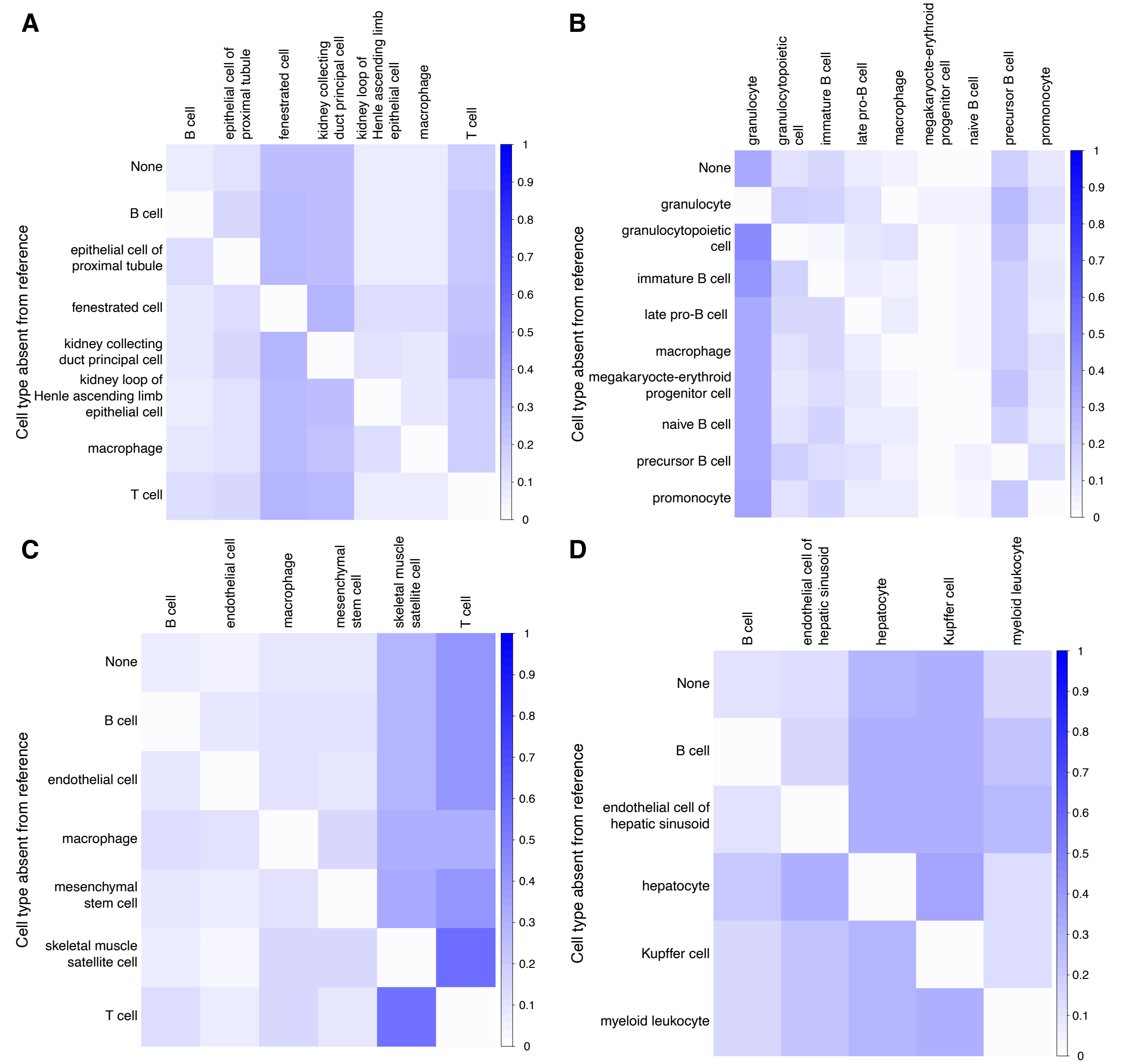


Figure B.5: **Deconvolution with missing cell types in the reference matrix.** Deconvolution was performed in pseudobulk experiments in four different organs (**A** – Kidney; **B** – Marrow; **C** – Limb muscle; **D** – Liver). For each organ, we followed a leave-one-out procedure in which one cell type is removed from the reference at a time. Deconvolution was then performed with an extra cell type in the pseudobulk in order to examine RNA-Sieve's ability to handle such a misspecification. The top row shows the inferred proportions with no missing reference cell types. Darker colors indicate a higher estimated proportion value. Here we used 10x Chromium data for the reference and Smart-seq2 for the pseudobulk.

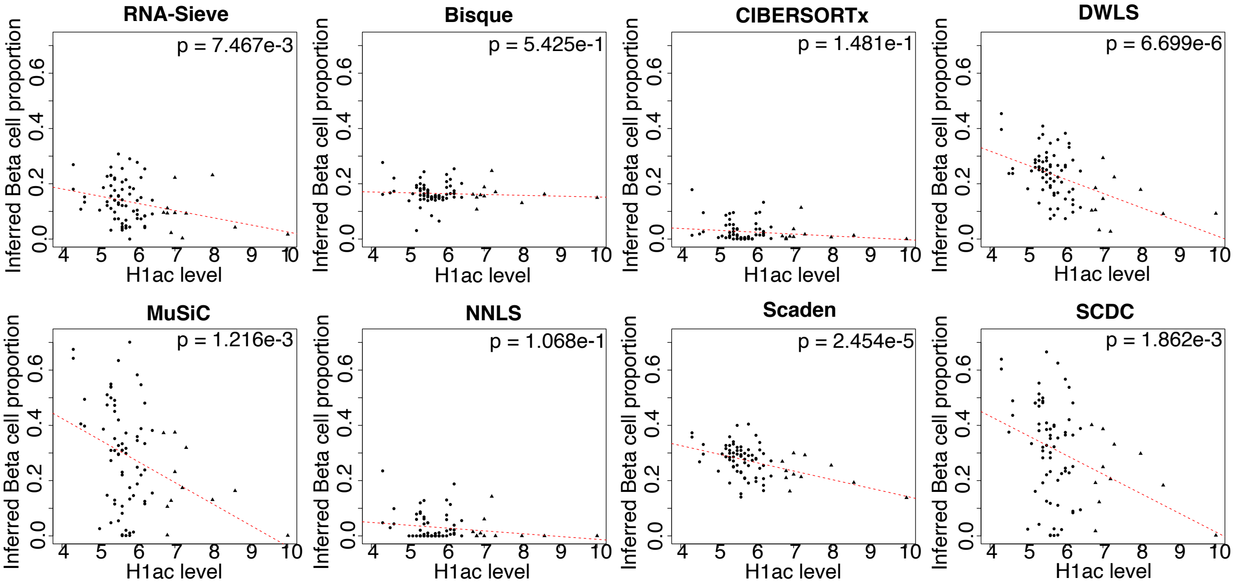


Figure B.6: **Deconvolution results on validation data.** Single-cell expression data in pancreatic islets from Xin et al. (2016) was used as reference to deconvolve bulk RNA-seq data from Fadista et al. (2014). Each point represents the estimated beta pancreatic islet cell proportion one of 77 bulks with recorded HbA1c levels. The *p*-value is for a univariate regression on the estimated proportions. Circles correspond to healthy samples while triangles represent samples from diabetic patients.

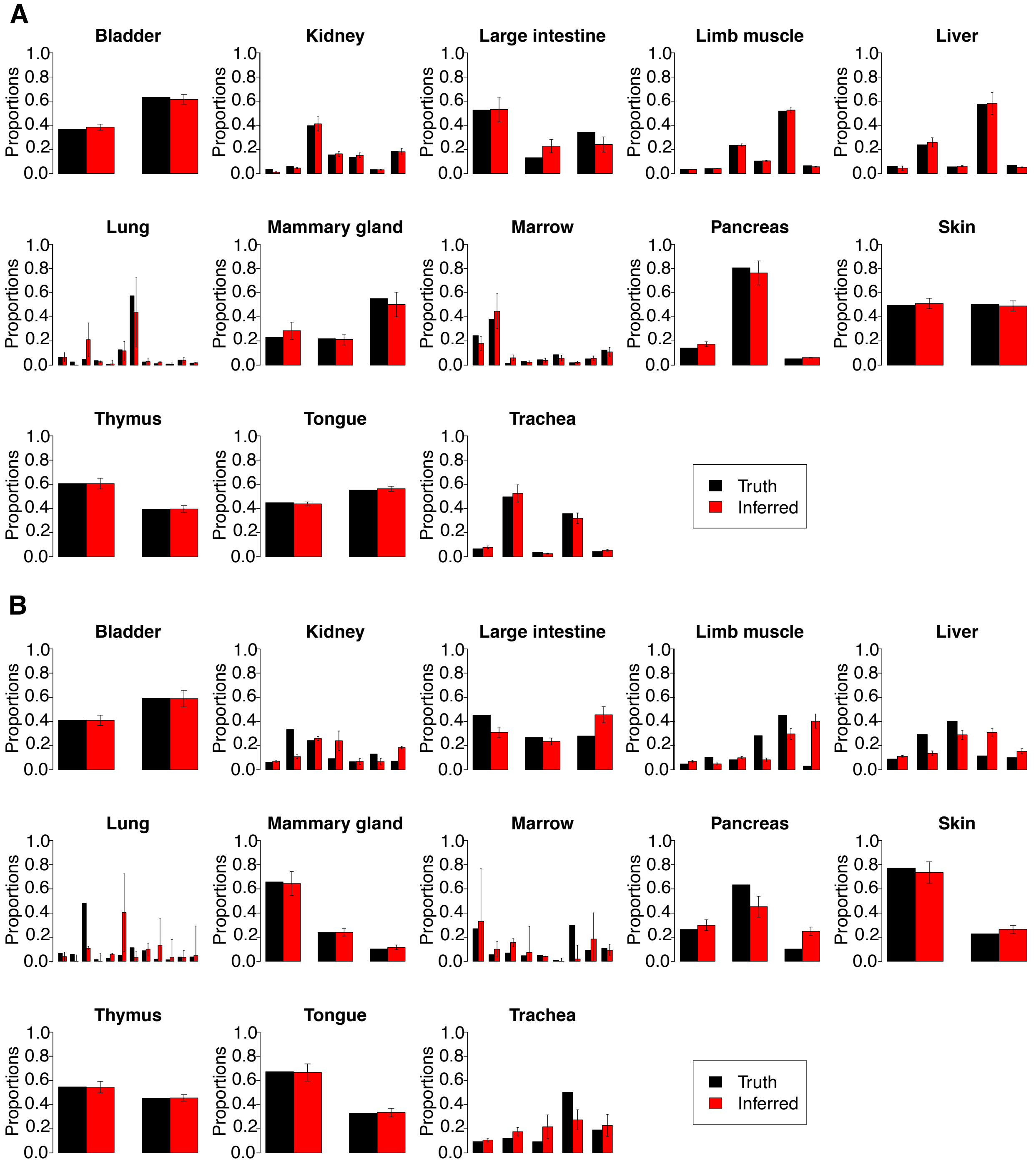


Figure B.7: **RNA-Sieve results with confidence intervals in pseudobulk experiments.** Deconvolution results for pseudobulk experiments. **A**: Within-protocol, 10x Chromium data; **B**: Across-protocol, 10x Chromium reference and Smart-seq2 pseudobulk.

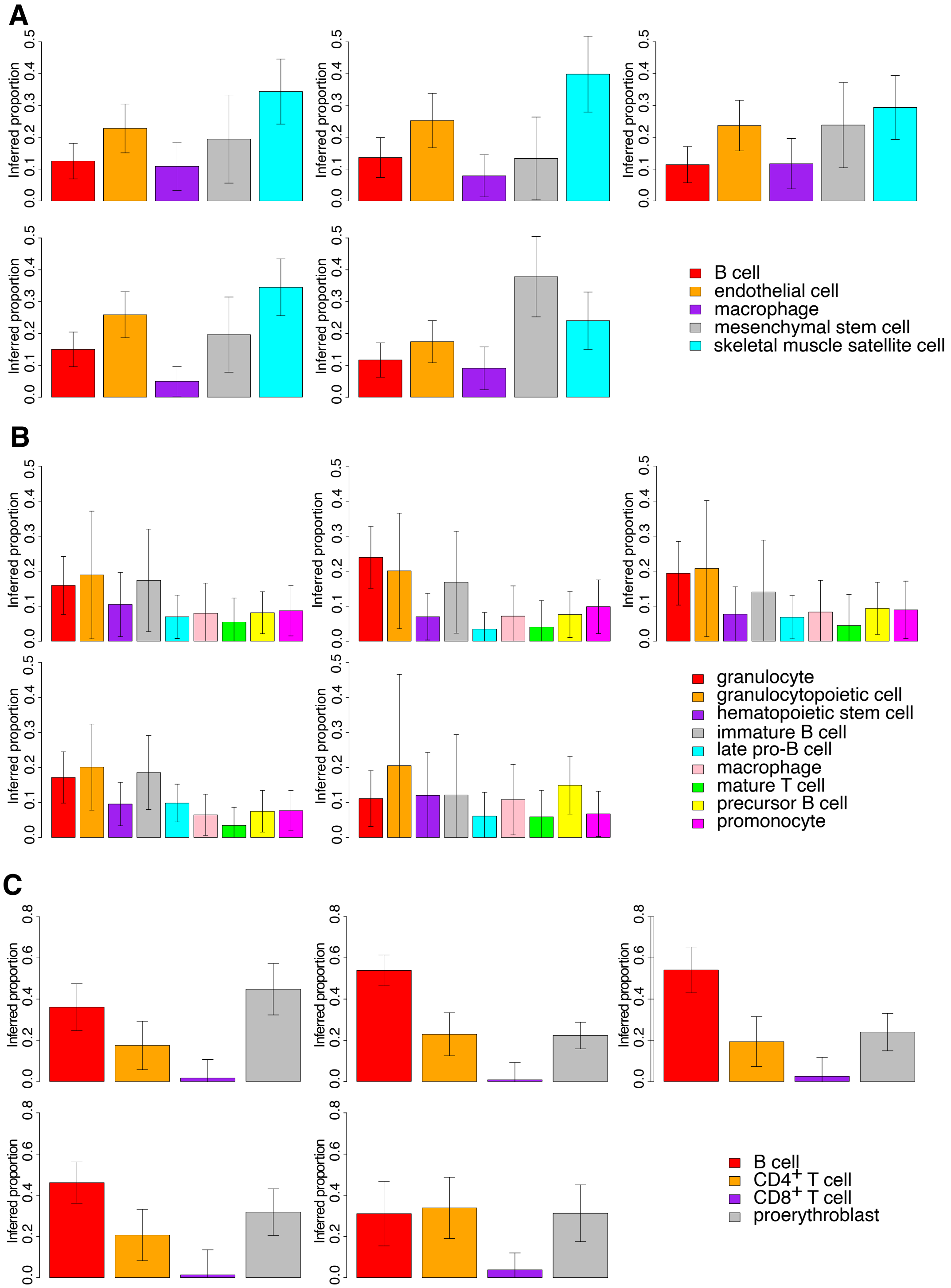


Figure B.8: **Confidence intervals with real bulk samples.** **A**–Limb muscle; **B**–Marrow; **C**–Spleen, with five randomly chosen samples (out of ∼40) each.

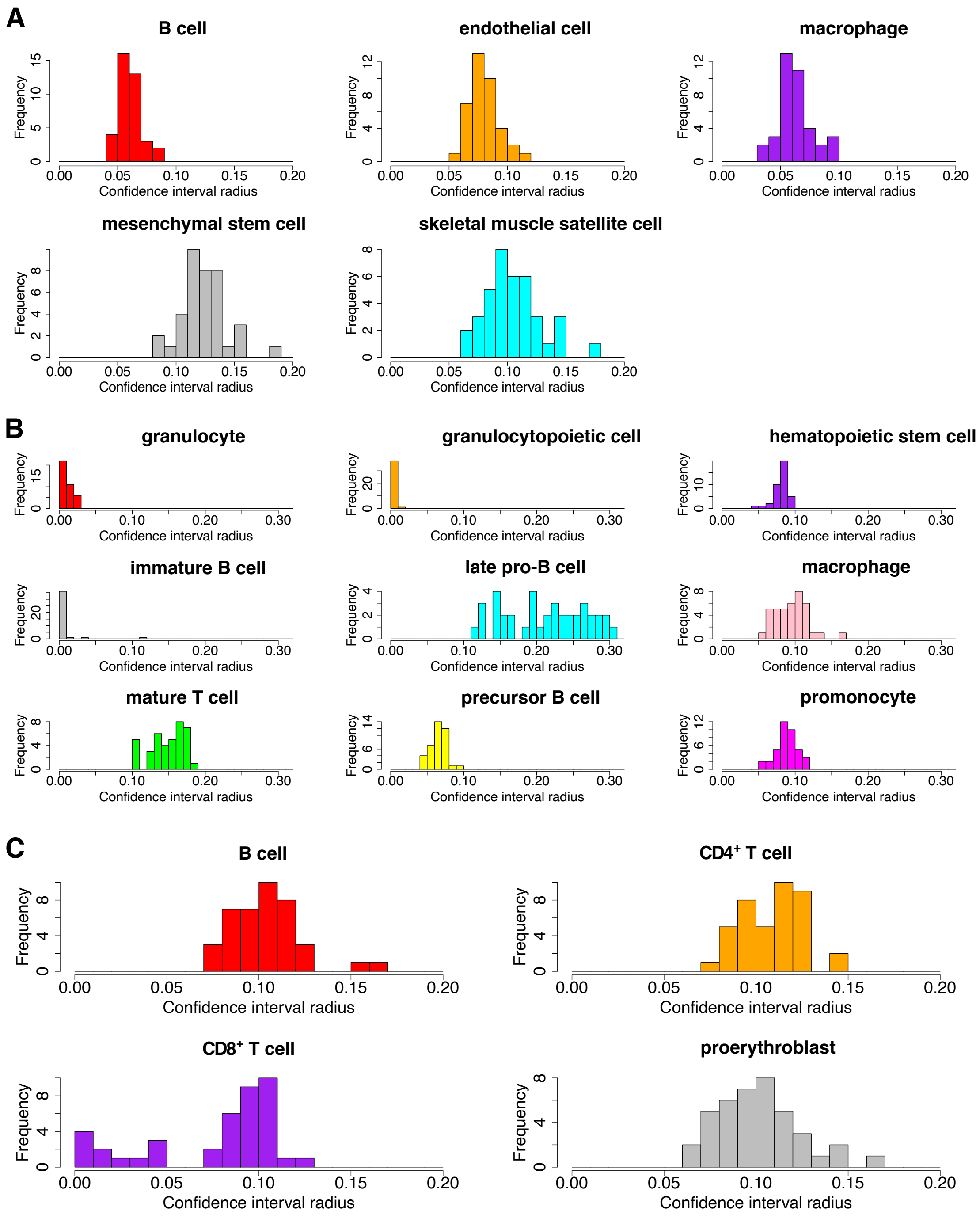


Figure B.9: **Histograms of CI radii with real bulk samples.** The radius of the 95% confidence interval for inferred cell type proportions was computed using RNA-Sieve for each real bulk sample in the listed organs (∼40 per organ). **A**–Limb muscle; **B**–Marrow; **C**–Spleen.

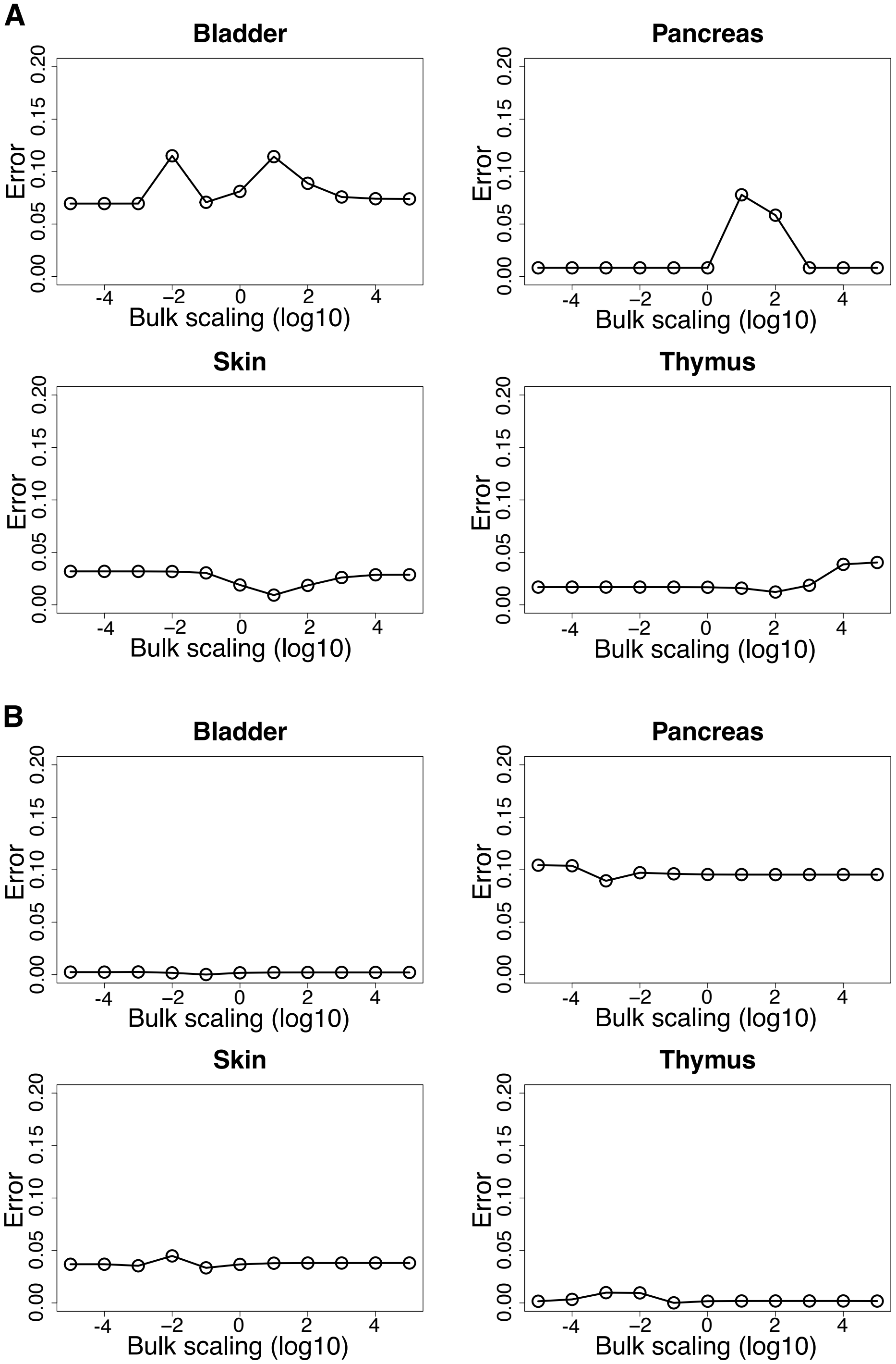


Figure B.10: **Effects of different bulk scalings.** Pseudobulks were amplified by various factors and RNA's robustness investigated. **A**–Smart-seq2 reference, 10x Chromium pseudobulk; **B**–10x Chromium reference, Smart-seq2 pseudobulk.

Appendix C

Supporting Information: Chapter 4

C.1 Computing Taylor coefficients

Proposition 7 (Taylor coefficients)**.** *Let $f_k^{p,\boldsymbol{w}}$ be the density (with respect to Lebesgue measure) of $\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}^p$, then on $[1/2,1]$, we have*

$$f_k^{p,\boldsymbol{w}} = \sum_{j=k-2}^{\infty} c_j^{\boldsymbol{w}} (1-x)^j, \tag{C.1}$$

where $c_r^{\boldsymbol{w}}$ can be computed in $O\left(\frac{r}{p}\log\frac{r}{p}\log k + [r\log r]^2\right)$ time.

Proof. We recall from (4.27) that $\mu_m = \mathbb{E}\left(\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}\right)^{pm}$ can be written as

$$\mu_m = \frac{1}{\binom{pm+k-1}{k-1}} \sum_{\boldsymbol{\eta}\in D_{m,k}} \frac{\binom{m}{\eta_1,\ldots,\eta_k}}{\binom{pm}{p\eta_1,\ldots,p\eta_k}} \prod_{j=1}^{k} w_j^{\eta_j} = \frac{s_m^{\boldsymbol{w}}}{\binom{pm+k-1}{k-1}}, \tag{C.2}$$

where $s_m^{\boldsymbol{w}} = \sum_{j=0}^{\infty} \sigma_j^{\boldsymbol{w}}(m)\cdot m^{-j}$ with $\sigma_j^{\boldsymbol{w}}(m)$ remaining constant $\sigma_j^{\boldsymbol{w}}$ past some threshold m_j^{w}. By Lemma 1 and the geometric interpretation of $\|\boldsymbol{S}_k\|_{p,\boldsymbol{w}}$, $f_k^{p,\boldsymbol{w}}$ is analytic on $[1/2,1]$, and hence we also have

$$\begin{aligned}\mu_m = \int_0^1 x^m f_k^{p,\boldsymbol{w}}(x)\,\mathrm{d}x &= \sum_{j=0}^{\infty} c_j^{\boldsymbol{w}} \int_0^1 x^m (1-x)^j\,\mathrm{d}x + O\left(e^{-m}\right) \\ &= \sum_{j=0}^{\infty} c_j^{\boldsymbol{w}} \left[(m+j+1)\binom{m+j}{j}\right]^{-1} + O\left(e^{-m}\right),\end{aligned} \tag{C.3}$$

which suggests that by matching coefficients in (C.2) and (C.3) we should be able to translate between $\sigma_j^{\boldsymbol{w}}$ and $c_j^{\boldsymbol{w}}$. For this to be helpful, we need to understand $\sigma_j^{\boldsymbol{w}}$:

Lemma 2 ($\sigma_j^{\boldsymbol{w}}$ recursion)**.** *Defining $b_r^j = [m^{-r}] \binom{m}{j} / \binom{pm}{pj}$ and employing notation as in (4.3), we have*

$$\sigma_r^{\boldsymbol{w}} = \sum_{j=0}^{r'} \left(\sigma_j^{(w_k,0)} \cdot \sigma_{r-j}^{\boldsymbol{w}_{-k}} + \mathbb{1}_{w_k=1} s_j^{\boldsymbol{w}_{-k}} \cdot b_r^j \right), \tag{C.4}$$

with initial condition $\sigma_r^{w_1,w_2} = \sum_{j=0}^{r'} b_r^j \left(\mathbb{1}_{w_2=1} w_1^j + \mathbb{1}_{w_1=1} w_2^j \right)$, where $r' = \lfloor r/(p-1) \rfloor$. In particular, we can compute $\sigma_r^{\boldsymbol{w}}$ in $O\left(r' \log r' \log k + [r \log r]^2\right)$ time.

Proof of Lemma 2. Slightly abusing notation, we have

$$\begin{aligned}
\sigma_r^{\boldsymbol{w}} &= \left[m^{-r}\right] s_m^{\boldsymbol{w}} = \left[m^{-r}\right] \sum_{\boldsymbol{\eta} \in D_{m,k}} \frac{\binom{m}{\eta_1,\ldots,\eta_k}}{\binom{pm}{p\eta_1,\ldots,p\eta_k}} \prod_{j=1}^{k} w_j^{\eta_j} \\
&= \left[m^{-r}\right] \sum_{\omega=0}^{m} \frac{\binom{m}{\omega}}{\binom{pm}{p\omega}} w_k^{\omega} \sum_{\boldsymbol{\eta} \in D_{m-\omega,k-1}} \frac{\binom{m-\omega}{\eta_1,\ldots,\eta_{k-1}}}{\binom{p(m-\omega)}{p\eta_1,\ldots,p\eta_{k-1}}} \prod_{j=1}^{k-1} w_j^{\eta_j} \\
&= \left[m^{-r}\right] \sum_{\omega=0}^{m} \frac{\binom{m}{\omega}}{\binom{pm}{p\omega}} w_k^{\omega} \cdot s_{m-\omega}^{\boldsymbol{w}_{-k}} \\
&= \left[m^{-r}\right] \sum_{\omega=0}^{r'} \frac{\binom{m}{\omega}}{\binom{pm}{p\omega}} w_k^{\omega} \cdot s_m^{\boldsymbol{w}_{-k}} + \left[m^{-r}\right] \sum_{\omega=0}^{r'} \frac{\binom{m}{\omega}}{\binom{pm}{\omega}} w_k^{m-\omega} \cdot s_{\omega}^{\boldsymbol{w}_{-k}} \\
&= \left[m^{-r}\right] \sum_{\omega=0}^{r'} s_m^{\boldsymbol{w}_{-k}} \sum_{j=0}^{\infty} b_j^{\omega} w_k^{\omega} m^{-j} + \left[m^{-r}\right] \sum_{\omega=0}^{r'} w_k^{m-\omega} s_{\omega}^{\boldsymbol{w}_{-k}} \sum_{j=0}^{\infty} b_j^{\omega} m^{-j} \\
&= \sum_{j=0}^{r'} \sigma_{r-j}^{\boldsymbol{w}_{-k}} \cdot \sigma_j^{w_k,0} + \mathbb{1}_{w_k=1} \sum_{\omega=0}^{r'} s_{\omega}^{w_{[1:k-1]}} \cdot b_j^{\omega} \\
&= \sum_{j=0}^{r'} \left(\sigma_j^{w_k,0} \cdot \sigma_{r-j}^{\boldsymbol{w}_{-k}} + \mathbb{1}_{w_k=1} s_j^{\boldsymbol{w}_{-k}} \cdot b_r^j \right),
\end{aligned}$$

as desired. To see that (C.4) can be computed in $O\left(r' \log r' \log k + [r \log r]^2\right)$ time, we notice that calculation of $s_r^{\boldsymbol{w}_{-k}}$ is $O\left(r' \log r' \log k\right)$ by the same reasoning as in Theorem 2, and b_r^j, written as,

$$\begin{aligned}
b_r^j = \left[m^{-r}\right] \frac{\binom{m}{j}}{\binom{pm}{pj}} &= (pj-1)!_p \left[m^{-r}\right] \prod_{\substack{\ell=p(m-j)+1 \\ p \nmid \ell}}^{pm-1} \frac{1}{pm} \cdot \frac{1}{1-\frac{\ell}{pm}} \\
&= (pj-1)!_p \left[m^{-r}\right] \prod_{\substack{\ell=p(m-j)+1 \\ p \nmid \ell}}^{pm-1} R\left(\frac{\ell}{pm}\right),
\end{aligned} \tag{C.5}$$

where $R(x)=\sum_{j=0}^{\infty}x^j$ is again a convolution of $(p-1)\cdot r'=r$ polynomials and hence computable in $O\left([r\log r]^2\right)$. □

With a proper understanding of $\sigma_j^{\boldsymbol{w}}$ at hand from Lemma 2, we may rewrite (C.2) as

$$\mu_m=\sum_{j=0}^{\infty}\left(\sum_{\omega=0}^{j}a_{\omega}^{k}\cdot\sigma_{j-\omega}^{\boldsymbol{w}}\right)m^{-j}+O\left(e^{-m}\right), \tag{C.6}$$

where $a_{\omega}^{k}=[m^{-\omega}]\binom{pm+k-1}{k-1}^{-1}$. Similarly, expanding (C.3) yields

$$\mu_m=\sum_{j=0}^{\infty}\left(\sum_{\omega=0}^{j-1}d_j^{\omega}\cdot c_{\omega}^{\boldsymbol{w}}\right)m^{-j}+O\left(e^{-m}\right), \tag{C.7}$$

where $d_j^{\omega}=\left[m^{-j}\right]\left[(m+\omega+1)\binom{m+\omega}{\omega}\right]^{-1}$. Consequently, matching the r^{th} coefficients in (C.6) and (C.7) allows to solve for $c_r^{\boldsymbol{w}}$:

$$c_r^{\boldsymbol{w}}=\frac{1}{r!}\left[\sum_{j=k-1}^{r+1}a_j^k\cdot\sigma_{r+1-j}^{\boldsymbol{w}}-\sum_{j=k-2}^{r-1}d_{r+1}^{j}c_j^{\boldsymbol{w}}\right], \tag{C.8}$$

where in the choice of summation indices we have used the fact that $a_j^k=0$ for $j\in\{0,\dots,k-2\}$ and $c_j^{\boldsymbol{w}}=0$ for $j\in\{0,\dots k-3\}$ by Proposition 3. Now $\{d_{r+1}^{0},\dots,d_{r+1}^{r-1}\}$ can be found in $O\left(r(\log r)^2\right)$ time, and given a,b and d, the recursion is solved in $O\left(r^2\right)$ steps, amounting to a total complexity of $O\left(r(\log r)^2+r^2+r'\log r'\log k+[r\log r]^2\right)=O\left(r'\log r'\log k+[r\log r]^2\right)$. □

C.2 Alternative scaling limits

Proof of Proposition 5. We will show that the moments of $Z_{n,k,p,\boldsymbol{w}}$ converge to the respective limiting moments of a $\mathcal{N}(0,1)$ variable, which together with Theorem 30.2 of [Bil95] implies the desired result. Hence we investigate

$$\begin{aligned}\mathbb{E}\left(Z_{n,k,p,\boldsymbol{w}}^{m}\right)&=\sigma_{n,k,p,\boldsymbol{w}}^{-m/2}\mathbb{E}\left(\|\boldsymbol{S}_{n,k}\|_{p,\boldsymbol{w}}^{p}-\mu_{n,k,p,\boldsymbol{w}}\right)^{m}\\&=\sigma_{n,k,p,\boldsymbol{w}}^{-m/2}\mathbb{E}\left(\sum_{j=1}^{k}w_j\underbrace{\left[\left(S_{n,k}[\![j]\!]\right)^{p}-\mu_{n,k,p}\right]}_{=:X_j}\right)^{m}\\&=\sigma_{n,k,p,\boldsymbol{w}}^{-m/2}\sum_{t=1}^{m}\sum_{\boldsymbol{a}\in D_{m,t}^{\leq}}C_{\boldsymbol{a}}\sum_{\substack{i_1,\dots,i_t\\\text{distinct}}}\mathbb{E}\left[(w_{i_1}X_{i_1})^{a_1}\cdots(w_{i_t}X_{i_t})^{a_t}\right],\end{aligned} \tag{C.9}$$

where $C_{\boldsymbol{a}}$ is a combinatorial factor to be determined later, and $D^{\leq}_{m,t}$ is the set of *ordered* (strong) compositions of m into t parts (which is in bijection with the set of partitions of m into t parts). We will argue that for case 1 of (4.37), all summands in (C.9) vanish in the limit of $n \to \infty$, while only the $t = m/2, a_1 = a_2 = \cdots = a_t = 2$ term survives in the $k = \alpha n$ regime. We begin by determining the asymptotics of $\mu_{n,k,p}$:

Lemma 3 (Asymptotics of $\mu_{n,k,p}$)**.** *We have*

$$\mu_{n,k,p} = \frac{n \cdot q_p(n,k)}{\langle k \rangle_p} = O\left(\frac{n^p}{k^p}\right), \tag{C.10}$$

where $q_p(n,k) = O(n^{p-1})$ is a polynomial in n and k, and $\langle k \rangle_p = k \cdot (k+1) \cdots (k+p-1)$ is the rising factorial.

Proof of Lemma 3. Let us first rewrite $\mu_{n,k,p}$ into a form more amenable to extract asymptotics:

$$\begin{aligned}\mu_{n,k,p} &= k^{-1}\mathbb{E}\|\boldsymbol{S}_{n,k}\|^p_{p,\mathbf{1}_k} = k^{-1}\sum_{j=1}^{k}\mathbb{E}\left(S_{n,k}[\![j]\!]\right)^p = \mathbb{E}\left(S_{n,k}[\![1]\!]\right)^p \\ &= \frac{1}{\binom{n+k-1}{k-1}}\sum_{j=0}^{n} j^p \binom{n-j+k-2}{k-2},\end{aligned} \tag{C.11}$$

whose RHS sum we claim can in general form be expressed as

$$\begin{aligned}\sum_{j=0}^{n} j^p \binom{n-j+\ell}{\ell} &= \sum_{j=\ell}^{n+\ell}(n-j+\ell)^p\binom{j}{\ell} \\ &= \binom{n+\ell-1}{\ell}\frac{(n+\ell)(n+\ell+1)}{\langle \ell+1 \rangle_{p+1}} q'_p(n,\ell),\end{aligned} \tag{C.12}$$

for some polynomial $q'_p(n,\ell) = O\left(n^{p-1}\right)$ if $p \geq 1$ and $q'_0(n,\ell) = n^{-1}$. We prove (C.12) by induction on p.

Base case: When $p = 0$, (C.12) simply becomes the hockey stick identity

$$\begin{aligned}\sum_{j=\ell}^{n+\ell}\binom{j}{\ell} = \binom{n+\ell+1}{\ell+1} &= \binom{n+\ell-1}{\ell}\frac{(n+\ell)(n+\ell+1)}{(\ell+1)} \cdot n^{-1} \\ &= \binom{n+\ell-1}{\ell}\frac{(n+\ell)(n+\ell+1)}{\langle \ell+1 \rangle_1} q'_0(n,\ell).\end{aligned} \tag{C.13}$$

Inductive step: Using the binomial recursion $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$, we find for $p \geq 1$

$$\sum_{j=\ell}^{n+\ell}(n-j+\ell)^p\binom{j}{\ell}$$

$$
\begin{aligned}
&= \sum_{j=\ell}^{n+\ell}(n-j+\ell)^p\binom{j+1}{\ell+1} - \sum_{j=\ell}^{n+\ell}(n-j+\ell)^p\binom{j}{\ell+1} \\
&= \sum_{j=\ell+1}^{n+\ell}(n+\ell+1-j)^p\binom{j}{\ell+1} \\
&\quad - \sum_{j=\ell+1}^{n+\ell}\left[\sum_{i=0}^{p}\binom{p}{i}(n+\ell+1-j)^i(-1)^{p-i}\right]\binom{j}{\ell+1} \\
&= \sum_{i=1}^{p-1}\binom{p}{i}(-1)^{p-i-1}\left[\sum_{j=\ell+1}^{n+\ell+1}(n+\ell+1-j)^i\binom{j}{\ell+1}\right] \\
&\quad + (-1)^{p-1}\sum_{j=\ell+1}^{n+\ell}\binom{j}{\ell+1} \\
&= \left[\sum_{i=1}^{p-1}\binom{p}{i}(-1)^{p-i-1}\binom{n+\ell}{\ell+1}\frac{(n+\ell+1)(n+\ell+2)}{\langle\ell+2\rangle_{i+1}}q'_i(n,\ell+1)\right] \\
&\quad + (-1)^{p-1}\left[\binom{n+\ell+2}{\ell+2}-\binom{n+\ell+1}{\ell+1}\right] \\
&= \binom{n+\ell-1}{\ell}\frac{(n+\ell)(n+\ell+1)}{\langle\ell+1\rangle_{p+1}}\left[\langle\ell+3\rangle_{p-1}\right. \\
&\quad \left. + \sum_{i=1}^{p-1}\binom{p}{i}(-1)^{p-i-1}(n+\ell+2)\langle\ell+3+i\rangle_{p-1-i}q'_i(n,\ell+1)\right] \\
&= \binom{n+\ell-1}{\ell}\frac{(n+\ell)(n+\ell+1)}{\langle\ell+1\rangle_{p+1}}q'_p(n,\ell),
\end{aligned}
\tag{C.14}
$$

where the second equality follows from binomial expansion of $(n-j+\ell)^p = (n+\ell+1-j-1)^p$, while the fourth equality follows from applying the inductive hypothesis. Note that $q'_p(n,\ell)$ is a sum of polynomials in $O(n^{p-1})$, and hence it itself is a polynomial in $O(n^{p-1})$.

To finish the proof of the lemma, it suffice to observe from (C.10) and (C.12) with $\ell = k-2$ that

$$
\mu_{n,k,p} = \frac{\binom{n+k-3}{k-2}}{\binom{n+k-1}{k-1}}\frac{(n+k-2)(n+k-1)}{\langle k-1\rangle_{p+1}}q'_p(n,k-2) = \frac{n\cdot q_p(n,k)}{\langle k\rangle_p},
\tag{C.15}
$$

where $q_p(x,y) = q'_p(x,y-2)$. □

With the dependence of $\mu_{n,k,p}$ on n and k in hand, we can elucidate the asymptotics of the summands in (C.9):

Lemma 4 (Asymptotics of summands)**.** *For $\ell \le t \le m$ and $\boldsymbol{a} \in D_{m,t}^{\le}$ such that $a_1 = \cdots = a_\ell = 1$ and $a_j \ge 2$ for all $j \in \{\ell+1,\dots,t\}$, we have*

$$\sigma_{n,k,p,\boldsymbol{w}}^{-m} \sum_{\substack{i_1,\dots,i_t\\ \text{distinct}}} \mathbb{E}\left[\prod_{j=1}^{\ell} w_{i_j}X_{i_j} \prod_{j=\ell+1}^{t} (w_{i_j}X_{i_j})^{a_j}\right] = O\left[\left(\frac{k}{n}\right)^{p(m-t)} \frac{k^{t-m/2}}{n^{\ell}}\right]. \tag{C.16}$$

Proof. Three cumbersome applications of (C.12) to the computation of the variance show that $\sigma_{n,p,k,w} = O\left(\left(\frac{n}{k}\right)^p \sqrt{k}\right)$, which takes care of the denominator on the LHS of (C.16). To treat the enumerator, we use exchangeability of the X_i as well as the compact support of the weights w_j to upper bound the magnitude of the sum on the LHS by the magnitude of

$$\begin{aligned} k^t W_{\max}^m \cdot \mathbb{E}&\left[\prod_{j=1}^{\ell} X_j \prod_{j=\ell+1}^{t} X_j^{a_j}\right] \\ &= k^t W_{\max}^m \mathbb{E}\left\{\mathbb{E}\left[X_1 \mid X_2,\dots,X_t\right] \prod_{j=2}^{\ell} X_j \prod_{j=\ell+1}^{t} X_j^{a_j}\right\}. \end{aligned} \tag{C.17}$$

For fixed n and growing k, we expect the bin occupations to decorrelate, and hence the conditional expectation on the RHS of (C.17) to vanish. Indeed, referring once more to (C.12) and writing $X_2^t := \sum_{j=2}^t X_j$, we have

$$\begin{aligned} \mathbb{E}\left[X_1 \mid X_2,\dots,X_t\right] = \mathbb{E}\left[X_1 \mid X_2^t\right] &= \mu_{n-X_2^t,k-t+1,p} - \mu_{n,k,p} \\ &= \frac{n-X_2^t}{\langle k-t+1\rangle_p} q_p\left(n-X_2^t, k-t+1\right) - \frac{n}{\langle k\rangle_p} q_p(n,k) \\ &= O\left(\frac{n^{p-1}}{k^p}\right), \end{aligned} \tag{C.18}$$

as long as $X_2^t = \sum_{j=2}^t S_{n,k}[\![j]\!] = o(n)$. Whence the magnitude of (C.17) is bounded above by

$$\begin{aligned} k^t W_{\max}^m \mathbb{E}&\left\{O\left(\frac{n^{p-1}}{k^p}\right) \mathbb{1}_{X_2^t=o(n)} \prod_{j=2}^{\ell} X_j \prod_{j=\ell+1}^{t} X_j^{a_j} + \mathbb{1}_{X_2^t \neq o(n)} \prod_{j=1}^{\ell} X_j \prod_{j=\ell+1}^{t} X_j^{a_j}\right\} \\ &\le k^t W_{\max}^m O\left(\frac{n^{p-1}}{k^p}\right) \mathbb{E}\left[\mathbb{1}_{X_2^t=o(n)} \prod_{j=2}^{\ell} X_j \prod_{j=\ell+1}^{t} X_j^{a_j}\right] + k^t W_{\max} \mathbb{P}\left[X_2^t \neq o(n)\right] n^m \\ &= k^t W_{\max}^m O\left(\frac{n^{p-1}}{k^p}\right) \underbrace{\mathbb{E}\left[\prod_{j=2}^{\ell} X_j \prod_{j=\ell+1}^{t} X_j^{a_j}\right]}_{A} + 2n^m k^t W_{\max} \mathbb{P}\left[X_2^t \neq o(n)\right]. \end{aligned} \tag{C.19}$$

Repeating the same reasoning used in (C.17) and (C.19) ℓ times on A yields an upper bound for the magnitude of (C.16) of

$$k^t W_{\max}^m O\left(\frac{n^{\ell(p-1)}}{k^{\ell p}}\right)\underbrace{\mathbb{E}\left[\prod_{j=\ell+1}^{t} X_j^{a_j}\right]}_{B} + 2(\ell-1)k^t n^m \underbrace{\mathbb{P}\left[S_{n,k}[\![1]\!] \neq o(n)\right]}_{C}. \tag{C.20}$$

We will argue below in Lemma 5 that bin sizes concentrate around their means n/k, which implies that C is exponentially small in k, while B scales like $(n/k)^{p(t-\ell)}$. Combining these with the $O(\sqrt{k}(n/k)^p)$ scaling of $\sigma_{n,k,p,\mathbf{w}}$ gives the final upper bound

$$O\left[k^t \frac{1}{k^{m/2}}\left(\frac{k}{n}\right)^{pm}\left(\frac{k}{n}\right)^{\ell-p\ell}\frac{1}{k^{\ell}}\left(\frac{k}{n}\right)^{p(\ell-t)}\right], \tag{C.21}$$

which simplifies to (C.16) as desired. □

To substantiate the claims about B and C in the proof of Lemma 4, we establish the following lemma:

Lemma 5 (Large deviations for bin sizes)**.** *If $k = O(n^\beta)$, then for all $0 < \varepsilon < \beta$ there exists $C_\varepsilon > 0$ so that*

$$\mathbb{P}\left[X_j \geq n^{1-\varepsilon}\right] < Ce^{-n^{\beta-\varepsilon}}. \tag{C.22}$$

Proof. We can compute explicitly

$$\begin{aligned}\mathbb{P}\left[S_{n,k}[\![1]\!] \geq n^{1-\varepsilon}\right] &= \frac{\binom{n-n^{1-\varepsilon}+k-1}{k-1}}{\binom{n+k-1}{k-1}} = \left[\prod_{j=n-n^{1-\varepsilon}+1}^{n}\left(1+\frac{k-1}{j}\right)\right]^{-1} \\ &\leq C\exp\left[-(k-1)\log\left(\frac{1}{1-n^{1-\varepsilon}}\right)+O\left(\frac{k^2}{n^{1+\varepsilon}}\right)\right] \\ &\leq C\exp\left[-n^{\beta-\varepsilon}+O\left(n^{\beta-2\varepsilon}+n^{2\beta-1-\varepsilon}\right)\right],\end{aligned} \tag{C.23}$$

which is dominated by the $n^{\beta-\varepsilon}$ term as long as $\varepsilon < \beta$. □

At last, we are now in shape to establish Proposition 5 from (C.16), for we see that

1. if $t < m/2$, the RHS is of $O\left[\left(\frac{k}{n}\right)^{m/2} k^{t-m/2}\right]$ and hence vanishes as $n \to \infty$;
2. if $t > m/2$, then since $\ell \geq 2t-m$ (because $\ell + 2(t-\ell) \leq m$), the RHS is of $O\left[(k/n^2)^{t-m/2}\right]$ and vanishes as $n \to \infty$;
3. if $t = m/2$ and $k = o(n)$, we obtain asymptotics of $O\left[(k/n)^{pm/2}\right]$, which vanishes once again as $n \to \infty$.

Hence, the only terms in (C.9) contributing to the limiting moments are those for which $t = m/2$ (and consequently $\ell = 0, a_j = 2$ for all $j \in \{1, \dots, t\}, C_{2,\dots,2} = (m-1)!!$ and m must be even), when $k = \Theta(n)$. That is,

$$\mathbb{E}\left(Z^m_{n,\alpha n,p,\boldsymbol{w}}\right) = (m-1)!! \cdot \mathbb{E} \prod_{j=1}^{m/2} \left[\frac{(S_{n,k}[\![j]\!])^p - \mu_{n,kp}}{k^{-1/2}\sigma_{n,k,p,w}} \right]^2, \tag{C.24}$$

which by decorrelation computations very similar to (C.18) is readily seen to converge to $(m-1)!!$. These are precisely the normal moments, which proves the first half (4.37) of Proposition 5. To show the second half (4.38), we sidestep individual moment computations and resort directly to Lévy's continuity theorem. To wit, we have

$$\begin{aligned}
\mathbb{E}e^{-t\|\boldsymbol{S}_{n,k}\|^p_{p,\boldsymbol{w}}} &= \mathbb{P}\left(\|\boldsymbol{S}_{n,k}\|_\infty \leq 1\right) \cdot \mathbb{E}\left[e^{-t\|\boldsymbol{S}_{n,k}\|^p_{p,\boldsymbol{w}}} \mid \|\boldsymbol{S}_{n,k}\|_\infty \leq 1\right] \\
&\quad + \mathbb{P}\left(\|\boldsymbol{S}_{n,k}\|_\infty > 1\right) \cdot \mathbb{E}\left[e^{-t\|\boldsymbol{S}_{n,k}\|^p_{p,\boldsymbol{w}}} \mid \|\boldsymbol{S}_{n,k}\|_\infty > 1\right] \\
&= \frac{1}{\binom{k}{n}} \sum_{\substack{S \subset [k] \\ \#S = n}} e^{-t\sum_{j \in S} w_j} + O\left(k^{-1}\right) \\
&= \frac{1}{k(k-1)\cdots(k-n+1)} \sum_{s_1=1}^{k} \sum_{\substack{s_2=1 \\ s_2 \neq s_q}}^{k} \cdots \sum_{\substack{s_n=1 \\ s_n \notin \{s_1,\dots,s_{n-1}\}}}^{k} e^{-t\sum_{j=1}^n w\left(\frac{s_j}{k}\right)} \\
&\quad + O\left(k^{-1}\right),
\end{aligned} \tag{C.25}$$

which, since w is bounded and continuous almost everywhere (and hence Riemann integrable) converges, as $k \to \infty$, to

$$\left(\int_0^1 e^{-tw(x)} \, \mathrm{d}x\right)^n = \mathbb{E}e^{-t\sum_{j=1}^n w(U_j)} \tag{C.26}$$

as desired. □