Econometric models of treatment with application to labour market outcomes

University College London

Jeffrey Rowley

Department of Economics

Submitted towards the degree of Doctor of Philosophy (PhD) in Economics

I, Jeffrey Rowley, confirm that the work presented in this thesis is my own. Where information has been derived from other sources or where work has been completed in collaboration with others, I confirm that this has been indicated in the thesis and that the contribution of each author has been recognised.

Determining assignment and identifying effect

Abstract

This thesis is concerned with the related problems of whether individuals benefit from treatment and determining who should receive treatment—both from an empirical and a theoretical viewpoint—and is broadly divided into two parts (with Chapter A outlining the notation that is used throughout).

The first part is mainly theoretical and examines the problem of how to allocate individuals to treatment in order to maximise a welfare criterion. Chapter B collects several results pertaining to modified Bessel functions of the first kind that are then used in Chapter C to derive a characterisation of the statistical divergence of a von Mises-Fisher distribution and its asymptotic behaviour. Chapter D develops a framework for estimating stochastic—rather than deterministic—assignment rules that are drawn from von Mises-Fisher distributions. This method for estimating assignment rules is motivated by a variational Bayes approximation of the infeasible optimal posterior distribution (i.e., the optimal stochastic assignment rule). The optimal posterior distribution is obtained from the maintained prior via an updating procedure that is based upon an empirical welfare criterion and that reflects the objective of a utilitarian social planner with access to the results of a randomised control trial or otherwise suitable data. The use of stochastic assignment rules has strong theoretical justification, and the proposed method is shown to achieve low regret with high probability. Experimental data from the National Job Training Partnership Act Study is used to illustrate the implementation and performance of this framework.

The second part is mainly empirical and examines the problem of identifying how additional children affect maternal labour supply. Chapter E studies a model of maternal behaviour that allows for a rich set of behaviours by mothers and that is applicable to a range of other empirical problems. Census (and related) data is used to estimate whether maternal labour supply increases or decreases in the presence of additional children.

Impact statement

Several novel results are introduced that can be built upon by or directly used in future academic research, both inside and outside of the field of economics. To give a specific example, von Mises-Fisher distributions have been studied extensively in the field of directional statistics, and statistical divergences appear frequently in the field of machine learning as a means of quantifying information; characterisation of the statistical divergence of a von Mises-Fisher distribution and its asymptotic behaviour is arguably of interest to students in both fields. Indeed, these and many of the other concepts that are discussed

are likely more familiar to students of other fields than to those of economics despite their applicability to many types of economic research. Aside from the intrinsic value of these results, their introduction to the field of economics and the suggestion of alternative approaches to empirical and theoretical research can be of benefit to students of economics—especially to those studying cyclical data, or settings that are homomorphic or isomorphic to spheres.

The tools that are developed are with specific economic environments in mind, and are motivated by particular questions that have a basis in policy design and evaluation. Despite this specificity, however, they are suitable for application to a broad range of other economic environments that have a similar composition. The evidence that these tools provide about labour market outcomes is intended to inform public policy on this subject. How should educational and skills training be directed at the unemployed? How do additional children affect maternal employment at the extensive margin? Such questions are of national and international importance, and the evidence that these tools provide is particularly credible due to the parsimony that both frameworks exhibit. In the case of the first framework, this manifests as a lack of specificity about how education and previous earnings influence wages, and is accompanied by the interpretability of the approximating class of assignment rules and their suitability for application to a broader population. In the case of the second framework, this manifests as a lack of restriction upon how mothers choose their employment status following the birth of additional children. It is important to emphasise that parsimony does not simply lend a tool credibility; it is likewise useful for understanding what information further restrictions can yield and for reconciling any existing evidence that is contradictory.

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published works

This declaration is supported by a further statement (see the Table of Contents).

1. For a research manuscript that has already been published (if not yet published, please skip to section 2):

   (a) What is the title of the manuscript?

   (b) Please include a link to or doi for the work:

   (c) Where was the work published?

   (d) Who published the work?

   (e) When was the work published?

   (f) List the manuscript's authors in the order they appear on the publication:

   (g) Was the work peer reviewed?

   (h) Have you retained the copyright?

   (i) Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi

   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3):

   (a) What is the current title of the manuscript?

   von Mises-Fisher distributions and their statistical divergence

   Stochastic Treatment Choice with Empirical Welfare Updating

(b) Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?

If 'Yes', please please give a link or doi:

arXiv:2202.05192

arXiv:2211.01537

(c) Where is the work intended to be published?

As a technical note in a lower tier statistics journal

As an article in a higher tier economics journal

(d) List the manuscript's authors in the intended authorship order:

Toru Kitagawa and Jeff Rowley

Toru Kitagawa, Hugo Lopez and Jeff Rowley

(e) Stage of publication:

In submission

In submission

3. For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4):

Kitagawa (senior author and supervisor)—Contributed to the introduction, and to the main theoretical results (suggested the use of the moment-generating function in place of the general definition of a moment to simplify mathematics). Reviewed.
Rowley—All other work.

Kitagawa (senior author and supervisor)—Contributed to the introduction, to the framework, to the main theoretical results (first theorem and its proof, second theorem and its proof, third theorem and its proof, and first lemma and its proof), and to the implementation. Reviewed.
Lopez—Contributed to the introduction, to the framework, to the main theoretical results (first theorem and its proof, second theorem and its proof, third theorem and its proof, and first lemma and its proof), and to the implementation. Prepared initial draft and edited.
Rowley—Contributed to the main theoretical results (proof of the second theorem, third theorem

and its proof, second lemma and its proof, third lemma and its proof), and to the empirical application. Prepared subsequent draft and edited.

4. In which chapter(s) of your thesis can this material be found?

   Chapter C

   Chapter D

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

Candidate:
Date:

Supervisor/Senior Author signature (where appropriate):
Date:

Acknowledgements

As I look back now upon my time at University College London, it is with a sense of detachment—and the perspective that affords. Throughout my journey from young undergraduate to the man that I am now, I am fortunate to have learnt from some truly exceptional individuals.

These notably include—in what is certainly a non-exhaustive list: Professors Martin Cripps and Vasiliki Skreta—for the challenge of teaching a changing and stimulating syllabus of microeconomics; Professor Roger Koenker—for insights into computation and programming; Professor Guy Laroque—for discussion of a broad range of topics; Mr Hugo Lopez—for an enjoyable and rewarding collaboration; Professor Thomas Sargent—for curing a natural aversion to macroeconomics; and Dr Michela Tincani.

Professors Andrew Chesher and Toru Kitagawa, in particular, have been central to my academic development, and I would like to extend my most heartfelt thanks to both for the patient and supportive guidance that they have offered me. Their influence on the way that I think about economics cannot be understated—one quote stands out above all others, for the many meanings it holds for me.

> If mean independence then why not median independence; and if median independence why not full statistical independence?

I recognise that so many of the opportunities that I—and others—have enjoyed is because of them both: Andrew through his patronage of the discipline via The Centre for Microdata Methods and Practice that he established and continues to advance; and Toru through his patronage of annual intercollegiate and interdisciplinary reading groups. I know that so many hold Andrew in high regard not only for his excellent scholarship but also for his generosity—I count myself amongst them, the warmth with which he greeted me when we ran into each other at the excellent Bristol Econometric Study Group Annual Conference last year just one example—and the very same can be said of Toru.

Indeed, to Toru, I would like to extend a special word of thanks for employing me as a research assistant over the past several years. That employment—and, preceding that, an Economic and Social Research Council grant (Number 1329842; Application of minimally restrictive econometric models)—has been crucial to my undertaking this doctoral programme for the financial support that it has given me, but also for the diverse tasks that Toru has exposed me to and the training that these tasks have provided. Quite simply, I cannot overstate how much I—and this work—owes to Toru, his faith in me, and what I have learnt from him. It has been a great privilege of mine to work for and with Toru, and I wish him and his family every happiness as they build their life together in Rhode Island.

I acknowledge and thank these individuals—and those too that I have not mentioned by name, but whom I have been taught by or worked with—for their influence on my journey.

Along the way I have made some dear and lifelong friends—schoolfriends—with whom I have shared so many precious experiences, and with whom I hope to make many more memories. Some have even ventured with me on this last part of my journey: Dr Rui Costa—especially for a memorable trip to Lisbon; Dr Gavin Kader; and Dr Jeremy McCauley. More great adventures await each of us, I am sure. For Mr Noel Thomas too—whose comradeship was so reassuring—with his retirement.

My family have, of course, been with me throughout my journey, and before too. To them, I have so much that I want to say, and so much to thank them for—I hope that finally seeing my work in print, to know that my struggle is at an end, is reward enough for now.

As my journey ends, I look forward to the next with my beloved fiancée, Debbie, and little Percival and Peregrine. The greatest opportunity that I have had—and ever will—was meeting you.

Dedicated to the memory of my grandmother

Rita Rowley

Contents

—| Notation |—

Conventions and typography

This chapter constitutes a record of the various mathematical and typographical conventions that I have adopted—a sort of standards manual if you will. I include some additional information that is otherwise taken to be implicit and underline when I feel that additional emphasis is warranted (something that I continue to do throughout this thesis). Some of these works are full or partial reproductions of collaborations with with co-authors. Where this is the case, I include only that material that I have contributed towards or else that is necessary for the reader's comprehension—omitting, for instance, any related proofs—and use the third-person singular.

I adopt one and a half spacing throughout. All margins are set to 3/4 inch except the binding (side) margin, which is set to 3/2 inch.

The main text is typeset in Computer Modern regular serif font (Figure A.1), or in Computer Modern italicised serif font (Figure A.1) whenever this is appropriate. Figures and tables are centred, framed and titled, with the title appearing above each object and typeset in Computer Modern capitalised serif font (Figure A.1). A caption is included below a figure or a table when I feel that further description is necessary, and typeset in Computer Modern italicised serif font. All external references are typeset in Computer Modern capitalised serif typeface, whilst internal references are regularly typeset. Paragraphs are marked using vertical line spacing.

Mathematical expressions are typeset in one of two fonts and several typographical effects depending upon what they are intended to represent, and are either presented inline (if they comprise a simple equality or inequality relation, or are the product of no more than two objects) or as equations. Equations are left-aligned and labelled, and are preceded by the conditions under which they hold. For instance, an equation might hold for all or for some values of an argument. Where these conditions are omitted then equations should be taken to hold for all values of their arguments. Equations are not punctuated as a matter of style.

FIGURE A.1
The typography used in the main text

Computer Modern regular serif font

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
a	b	c	d	e	f	g	h	i	j	k	*l*	m	n	o	p	q	r	s	t	u	v	w	x	y	z

Computer Modern italicised serif font

A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*W*	*X*	*Y*	*Z*
a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*	*q*	*r*	*s*	*t*	*u*	*v*	*w*	*x*	*y*	*z*

Computer Modern capitalised serif font

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	B	C	D	E	F	G	H	I	J	K	*l*	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z

Underpinning my analysis and associated exposition is The Probability Approach in Econometrics (HAAVELMO, 1944), and I assume throughout that there exists a (standard) probability space equipped with the Borel algebra, and upon which are defined a multitude of stochastic variables that stand-in for real economic variates.

I denote stochastic variables using the Latin alphabet, distinguishing single-valued stochastic variables from their multiple-valued counterparts (Figure A.2). I refer to single-valued stochastic variables as random variables, which I denote using Computer Modern italicised serif font; and I refer to multiple-valued stochastic variables as random sets or random vectors, which I denote using Computer Modern bold italicised serif font. I reserve the upper-case alphabet for stochastic variables, whilst I reserve the lower-case alphabet for their realisations (actual or hypothetical). Where I refer to more than one realisation of the same stochastic variable, I distinguish these two realisations using a combination of accents and subscripted characters from the Latin alphabet. So as to minimise overlap or possible confusion, I avoid using those characters from the alphabet that are difficult to distinguish from Arabic numerals or that I use extensively for other purposes,[1] such as to represent mathematical objects and operators that I rely upon.

I denote the collection of values that a stochastic variable can take using the Latin alphabet combined with Fraktur font (Figure A.3). By support, I mean one of two things, which I do not distinguish between: the collection of values that a stochastic variable can take; and the collection of values that a stochastic variable does take (i.e., the collection of values that occur with positive probability). I reserve the upper-case alphabet for the supports of stochastic variables, whilst I reserve the lower-case alphabet for their cardinalities. Where I am concerned with the support of a stochastic variable conditional on

1 I present the full alphabet in Figure A.2 in any case, for completeness.

FIGURE A.2
The typography used to denote random variables, sets and vectors

Single-valued stochastic variables—Random variables

$$\begin{array}{cccccccccccccccccccccccc} A & B & C & D & E & F & G & H & I & J & K & L & M & N & O & P & Q & R & S & T & U & V & W & X & Y & Z \\ a & b & c & d & e & f & g & h & i & j & k & \ell & m & n & o & p & q & r & s & t & u & v & w & x & y & z \end{array}$$

Multiple-valued stochastic variables—Random sets and vectors

$$\begin{array}{cccccccccccccccccccccccc} \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{C} & \boldsymbol{D} & \boldsymbol{E} & \boldsymbol{F} & \boldsymbol{G} & \boldsymbol{H} & \boldsymbol{I} & \boldsymbol{J} & \boldsymbol{K} & \boldsymbol{L} & \boldsymbol{M} & \boldsymbol{N} & \boldsymbol{O} & \boldsymbol{P} & \boldsymbol{Q} & \boldsymbol{R} & \boldsymbol{S} & \boldsymbol{T} & \boldsymbol{U} & \boldsymbol{V} & \boldsymbol{W} & \boldsymbol{X} & \boldsymbol{Y} & \boldsymbol{Z} \\ \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{c} & \boldsymbol{d} & \boldsymbol{e} & \boldsymbol{f} & \boldsymbol{g} & \boldsymbol{h} & \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} & \boldsymbol{\ell} & \boldsymbol{m} & \boldsymbol{n} & \boldsymbol{o} & \boldsymbol{p} & \boldsymbol{q} & \boldsymbol{r} & \boldsymbol{s} & \boldsymbol{t} & \boldsymbol{u} & \boldsymbol{v} & \boldsymbol{w} & \boldsymbol{x} & \boldsymbol{y} & \boldsymbol{z} \end{array}$$

The upper-case alphabet is reserved for stochastic variables, whilst the lower-case alphabet is reserved for their realisations as well as for certain deterministic objects.

FIGURE A.3
The typography used to denote the supports of stochastic variables

$$\begin{array}{cccccccccccccccccccccccc} \mathfrak{A} & \mathfrak{B} & \mathfrak{C} & \mathfrak{D} & \mathfrak{E} & \mathfrak{F} & \mathfrak{G} & \mathfrak{H} & \mathfrak{I} & \mathfrak{J} & \mathfrak{K} & \mathfrak{L} & \mathfrak{M} & \mathfrak{N} & \mathfrak{O} & \mathfrak{P} & \mathfrak{Q} & \mathfrak{R} & \mathfrak{S} & \mathfrak{T} & \mathfrak{U} & \mathfrak{V} & \mathfrak{W} & \mathfrak{X} & \mathfrak{Y} & \mathfrak{Z} \end{array}$$

Cardinality or dimension

$$\begin{array}{cccccccccccccccccccccccc} \mathfrak{a} & \mathfrak{b} & \mathfrak{c} & \mathfrak{d} & \mathfrak{e} & \mathfrak{f} & \mathfrak{g} & \mathfrak{h} & \mathfrak{i} & \mathfrak{j} & \mathfrak{k} & \mathfrak{l} & \mathfrak{m} & \mathfrak{n} & \mathfrak{o} & \mathfrak{p} & \mathfrak{q} & \mathfrak{r} & \mathfrak{s} & \mathfrak{t} & \mathfrak{u} & \mathfrak{v} & \mathfrak{w} & \mathfrak{x} & \mathfrak{y} & \mathfrak{z} \end{array}$$

The cardinality or dimension of a set or vector can be finite or (countably or uncountably) infinite.

the realisations of one or more other stochastic variables then I use the syntax

$$\text{Support of stochastic variable} \leftarrow\!\shortmid \text{ (Realisations of other stochastic variables)} \tag{A.1}$$

omitting the included parentheses if there is a single realisation that is being conditioned upon; I regard conditioning on multiple realisations of one or more other stochastic variables as a union operation over the support conditional on each realisation. For illustration, I write the support of A conditional on b as $\mathfrak{A} \leftarrow\!\shortmid b$ and the support of A conditional on b and c as $\mathfrak{A} \leftarrow\!\shortmid (b, c)$, with $\mathfrak{a} \leftarrow\!\shortmid b$ and $\mathfrak{a} \leftarrow\!\shortmid (b, c)$ the respective cardinalities. The notation that I use for the support of a random variable is consistent with the notation that I use for the co-domain or support of a function.

The support of a random variable can be of finite or of (countably or uncountably) infinite cardinality;[2] where the support of a random variable is finite, it is often convenient to write the support as the Arabic numerals one through the cardinality. The supports of random sets and random vectors inherit their cardinalities from the random variables that they are, in essence, constructed around. Every stochastic variable that I consider inhabits one of several common spaces that I denote using the Latin alphabet combined with blackboard-effect Computer Modern serif font (Figure A.4). These spaces include general spaces such as the complex numbers, the integer numbers, the natural numbers and the real numbers,

2 I assume that the axiom of choice holds.

but also include some special spaces such as the spherical manifold, the unit continuum and the unit simplex. These special spaces warrant some additional description. The spherical manifold is defined for higher dimensional Euclidean spaces as the collection of unit vectors; the unit continuum is defined for higher dimensional Euclidean spaces as the collection of vectors whose elements are all between zero and one; and the unit simplex is defined for higher dimensional Euclidean spaces as the collection of vectors of non-negative elements whose elements sum to one. I otherwise refer to the spherical manifold as the hypersphere, to the unit continuum as the hypercube and to the unit simplex as the simplex. I subscript spaces that can contain zero by an asterisk to render the space exclusive of zero. I superscript the space by infinity to render the power set of (or Borel sets over) the space; I superscript the space by a finite integer (or by i, as a placeholder for an unspecified finite integer) to render the product space; and I use a colon separator (plus symbol preceding infinity) to render the power set over the product space. To summarise, I use the syntax

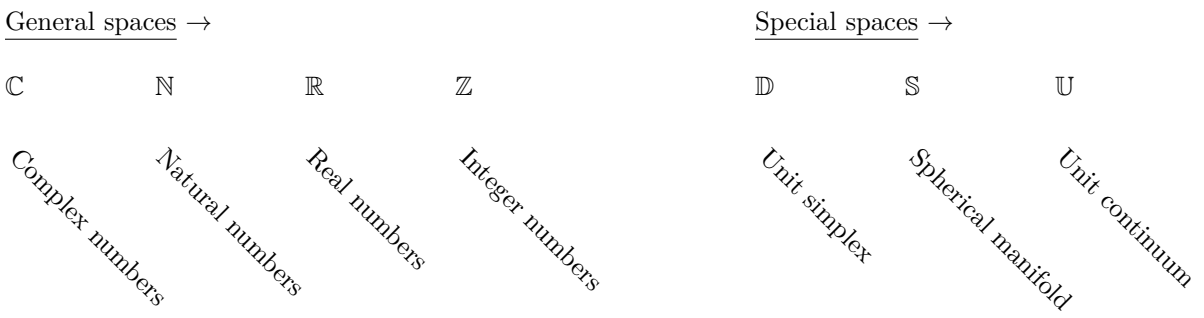
$$\begin{aligned} \text{Space}^{\infty} &= \text{the power set of the space} \\ \text{Space}^{i} &= \text{the product set of the space} \\ \text{Space}_{*}^{i:\infty} &= \text{the power set of the product set of the space exclusive of zero} \end{aligned} \tag{A.2}$$

which is a convention that I also adopt when considering the support of a stochastic variable (in place of a space). I am not specific as to whether these spaces or their subsets are closed or open or neither unless providing this detail is useful. The notation that I use for the space that a random variable inhabits is consistent with the notation that I use for the space that a function inhabits.

To distinguish cases where a random set or random vector is such that several of their elements can take the same value from cases where each of their elements is distinct, I use $\in$ (the in symbol) and $\subset$ (the subset symbol), respectively. To avoid any confusion, particularly with respect to closedness and openness, I use inequality relations to define intervals rather than brackets and parentheses; I use brackets to group objects (in advance of applying an operation to them say) and parentheses to enclose function arguments.

Where I am concerned with the particular elements of a set or vector, I use subscripted Arabic numerals to index and reference individual elements. I denote the concatenation of elements of a set or of a vector using braces, separating individual elements by a comma or a semi-colon depending upon which is clearer. Both sets and vectors are, in essence, lists of varying dimension or length, and the notation that I use does not distinguish between them in view of this. Often a set or vector comprises elements

FIGURE A.4
The typography used to denote some common spaces



that have a common definition or form, for which I use the syntax

$$\text{List} = \{\text{List}_i : \text{Definition of the } i^{\text{th}} \text{ element (and all other elements) of the list}\} \tag{A.3}$$

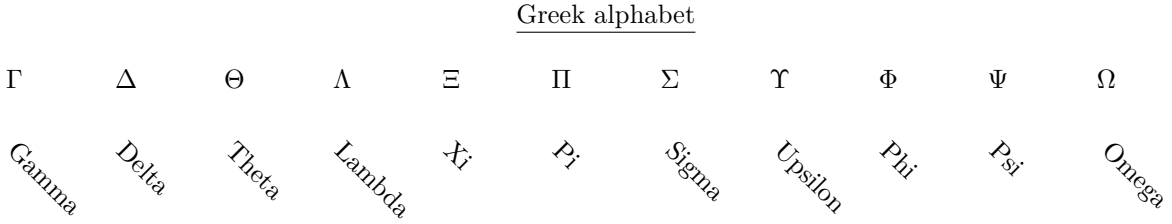
additionally specifying the cardinality of these objects or other information when I feel that this is warranted. To transpose a list, I use $\top$ (the transpose symbol), which I write in superscript.

Some of the characters of the Latin alphabet already have a well-established meaning in the body of existing economic work. For instance, the characters Y and T are typically interpreted as observable endogenous economic variates (response and treatment, respectively); the characters Z and X as observable exogenous economic variates (instrument and covariates, respectively); and the character U as a latent exogenous economic variate (heterogeneity). I maintain this convention for the most part, whether as random variables as is written above or as random sets and random vectors.

What distinguishes endogenous economic variates from their exogenous counterparts is how they are determined. I say that an economic variate is endogenous if it is determined by other economic variates in the overarching system, and is exogenous if it is given (i.e., it is determined outside of the overarching system, or is determined by an economic agent other than the one whose choice is of principal focus). The relationships between the endogenous and exogenous components of the economic environment are described by structural functions, for which I adopt the same conventions as I do for functions.

I denote functions (and operators) using characters from the Greek or Latin alphabets combined with either Computer Modern italicised serif font or Computer Modern regular serif font, omitting those characters from the Greek alphabet that overlap with or appear similar to characters from the Latin alphabet (Figure A.5). The reader is undoubtedly familiar with many of the functions that are associated with these characters, but some may be unfamiliar. I enclose the arguments of functions inside parentheses. In cases where several functions share a purpose, I denote these functions using the same character; to

FIGURE A.5
The typography used to denote functions



distinguish between each function, I employ a combination of accents and characters from the Latin alphabet. For example, in the classical model of endogenous choice described by the equations

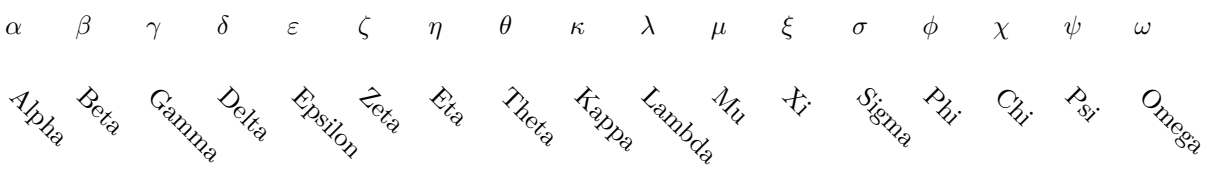
$$\begin{aligned} y &= h_y\left(t, x, u\right) \\ t &= h_t\left(z, x, u\right) \end{aligned} \tag{A.4}$$

I write h to refer to h_y and h_t together. This example also serves to illustrate how I refer to functions in the main text. Whilst I typeset non-Latin characters exclusively in regular font, I typeset Latin characters in a mixture of italicised and regular font; as a general rule, I use regular font to denote common (perhaps more accurately described as standard, since some of the objects that I work with are uncommon) constants, functions and operators, otherwise using italicised font. Particular examples include the characters D (the differential operator), E (the elementary vector, the expectation operator, or the exponential constant), I (the identity matrix or vector, or the modified Bessel function of the first kind), M (the annihilation matrix), O (the order operator), and P (the projection matrix), to list but a few examples.

Often, the functions that I define are characterised or multiplied by one or several parameters. Where these parameters are of finite dimension, I term the associated function parametric; where these parameters are of infinite dimension, I term the associated function non-parametric; and where some of these parameters are of finite dimension and others of infinite dimension, I term the associated function semi-parametric. I denote parameters using the Greek alphabet combined with Computer Modern italicised serif font (Figure A.6), omitting those characters that overlap with or appear similar to characters in the Latin alphabet or that are mathematical constants that I rely upon. To distinguish the parameters of a function from its other inputs, I use a semi-colon separator.

In some limited cases, the functions that I define rely upon other functions. That is, they are functions of functions—otherwise known as functionals. To distinguish the functions of a functional from its other

FIGURE A.6
The typography used to denote parameters



inputs and parameters, I use a semi-colon separator. This syntax can be summarised as

$$\begin{aligned}&\text{Output} = \text{Function}\,(\text{Input};\,\text{Parameter})\\&\text{Output} = \text{Functional}\,(\text{Function};\,\underline{\text{Functional}}\ \text{input};\,\underline{\text{Functional}}\ \text{parameter})\end{aligned} \tag{A.5}$$

with non-applicable arguments simply omitted as and when this is required. I otherwise do not treat functionals differently from other types of functions.

I emphasise that functionals are <u>not</u> the same thing as composite functions. Where I sequentially apply more than two functions and no detail is omitted by doing so (i.e., a later function does not require additional arguments to be specified beyond the output of the preceding function), I use $\circ$ (the <u>composition</u> symbol). This syntax can be summarised as

$$\begin{aligned}&\text{Output} = \text{Second function}\,(\text{First function}\,(\text{Input}))\\&\text{Output} = \text{Third function} \circ \text{Second function}\,(\text{First function}\,(\text{Input}))\end{aligned} \tag{A.6}$$

with the composition symbol replaced by more standard parentheses if the third function is reliant on another input or on a parameter. The composition symbol is not to be confused with $\cdot$ (the <u>dot</u> symbol) that I use for multiplication. I also use the composition symbol to make notation more concise when several functions with the same input are added, divided, multiplied or subtracted.

The functions that I define are, for the most part, regular, in that they map from one or several single-valued inputs (whether deterministic or random) to a single-valued output. I do, however, allow for a function's generalisation to set- or vector-valued inputs that, accordingly, map to set- or vector-valued outputs. For example, returning to the classical model of endogenous choice described by Equation A.4, I require that h_y satisfies such equality relations as

$$h_y\,(t, x, \boldsymbol{u}) = \left\{\boldsymbol{y}_i : \boldsymbol{y}_i = h_y\,(t, x, \boldsymbol{u}_i)\right\} \tag{A.7}$$

say, so that h_y is well-defined irrespective of the dimension of its inputs. I require that functions have well-defined inverses; with the exception of the inverse operator, I write operations applied to functions

FIGURE A.7
The typography used to denote intermediate objects

Cyrillic alphabet

Ђ	Ж	Љ	Њ	Ю	Я	Æ	Ґ	Ӈ	Ѫ
Dje	*Zhe*	*Lje*	*Nje*	*Yu*	*Ya*	*Ae*	*Ghe*	*En*	*Yus*

in superscript following both the function and the parentheses enclosing its arguments.

Differentiation and integration are two operations that I apply to functions but do not use superscripted notation for; rather, I write the differential and integral operators in full, making explicit the variable of differentiation or integration.

To characterise the behaviour of functions as the principal argument takes extreme values I use either the big-O operator (from big-O notation, otherwise referred to as Bachmann-Landau notation in LATTIMORE and SZEPESVÁRI, 2020) or $\simeq$ (the approximately equal symbol). Specifically, I use the big-O operator to characterise the behaviour of functions as their principal argument becomes large (i.e., approaches infinity) and the approximately equal symbol to characterise the behaviour of functions as their principal argument becomes small (i.e, approaches zero). In each case, I express the function in terms of another function. This syntax can be summarised as

$$\begin{aligned} &\text{Function}\,(\text{Principal argument}) = \text{O}\,(\text{Known function}\,(\text{Principal argument})) \\ &\text{Function}\,(\text{Principal argument}) \simeq \text{Known function}\,(\text{Principal argument}) \end{aligned} \tag{A.8}$$

which should be taken to mean that the function on the left-hand side grows or decays at a rate that is proportional to the known function on the right-hand side, or is else is approximately equal to the known function on the right-hand side; the growth or decay rate of the known function on the right-hand side is then what I refer to as the rate of the function on the left-hand side.

To characterise the neighbourhood that a function occupies for particular values of its inputs, it is necessary to specify constants. This is also the case when talking about how likely something is to occur. I follow convention in reserving the characters Delta and Epsilon from the Greek alphabet (Figure A.6) to denote positive and infinitesimal or specified constants, as is typical elsewhere. I capture the probability with which an event occurs, which is typically the probability with which a stochastic variable takes one or several possible values, using the probability operator. Where this probability is

conditional, I use the conditional probability operator. I summarise this syntax as

$$\begin{aligned}&\Pr\left(\text{Stochastic variable} = \text{Realisation of stochastic variable}\right)\\&\Pr\left(\text{Stochastic variable} = \text{Realisation of stochastic variable}|\text{Conditioning event}\right)\end{aligned} \tag{A.9}$$

depending upon whether it is an unconditional or conditional probability that is of interest, and where I write the conditioning event simply as the realisation of a stochastic variable (rather than as the full event that the stochastic variable takes that realisation) if this is sufficient to make the full event clear.

Since each of the following chapters are self-contained and standalone (with the exception of some of the appendices, that precede or follow certain chapters), I recycle notation between them. The reader should not, therefore, carry-over notation and terminology from one chapter to another. In several places, it is necessary for me to define functions that play a very limited role in the overall analysis; I denote such intermediate functions using characters from the Slavic or non-Slavic Cyrillic alphabets combined with Computer Modern regular serif font, omitting those characters from the Cyrillic alphabet that overlap with or appear similar to characters from the Latin alphabet (Figure A.7).

—| Chapter B |—

Modified Bessel functions of the first kind

Hereafter, where I refer to the modified Bessel function, I intend this to mean the modified Bessel function of the first kind.

—| Section 1 |—

Definition

The modified Bessel function is defined in NIST (2021, §10.25 and §10.32) and, its precursor, Abramowitz and Stegun (1964, §9.6), and I refer the reader to those references for further (detailed) information about this function. The modified Bessel function is defined, for all $a > 0$, as

$$\mathrm{I}_m\left(a\right) \doteq \left(a/2\right)^m \cdot \sum_{k=0}^{\infty} \frac{\left(a/2\right)^{2k}}{k! \cdot \Gamma\left(m+k+1\right)} = \frac{\left(a/2\right)^m}{\sqrt{\pi} \cdot \Gamma\left(m+1/2\right)} \cdot \int_0^{\pi} \exp\left(\pm a \cdot \cos\left(b\right)\right) \cdot \sin\left(b\right)^{2m} \cdot \mathrm{d}b \qquad \text{B.1}$$

or as

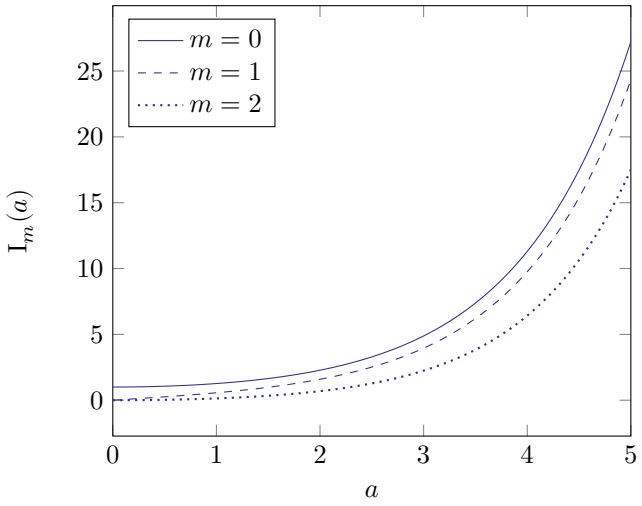
$$\mathrm{I}_m\left(a\right) \doteq \frac{1}{\pi} \cdot \int_0^{\pi} \exp\left(a \cdot \cos\left(b\right)\right) \cdot \cos\left(m \cdot b\right) \cdot \mathrm{d}b - \frac{\sin\left(m \cdot \pi\right)}{\pi} \cdot \int_0^{\infty} \exp\left(-m \cdot c - a \cdot \cosh\left(c\right)\right) \cdot \mathrm{d}c \qquad \text{B.2}$$

where m is said to be the order and a is said to be the argument. Various other definitions of the modified Bessel function exist, including as the purely imaginary solution to

$$a^2 \cdot \frac{\mathrm{d}^2 f\left(a\right)}{\mathrm{d}a^2} + a \cdot \frac{\mathrm{d}f\left(a\right)}{\mathrm{d}a} - \left(a^2 + m^2\right) \cdot f\left(a\right) = 0 \qquad \text{B.3}$$

which is a modification of Bessel's equation. This modification relates to the final term on the left-hand side of Equation B.3; Bessel's equation is generally written with the final term on the left-hand side entering additively. This is what distinguishes a Bessel function (that admits real solutions, and is a generalisation of a sine wave) from a modified Bessel function (that admits imaginary solutions, and is an increasing function). Bessel's equation and the modified Bessel function relate to Laplace's equation and harmonic functions (that describe the propagation of a wave along a taut string). The modified

FIGURE B.1
Modified Bessel functions of varying order



Modified Bessel functions of increasing order are plotted for several values of their argument.

Bessel function is a type of generalised hypergeometric function.

The modified Bessel function is plotted in Figure B.1 for various integer orders, although the function is equally well-defined for non-integer orders too, including negative ones.

—| SECTION 2 |—

RECCURENCE RELATIONS

The modified Bessel function satisfies the recurrence relation (NIST, 2021, §10.29)

$$\mathrm{I}_m(a) = \frac{a}{2m} \cdot \left(\mathrm{I}_{m-1}(a) - \mathrm{I}_{m+1}(a)\right) \tag{B.4}$$

that arises from

$$\mathrm{I}_{m-1}(a) - \frac{m}{a} \cdot \mathrm{I}_m(a) = \frac{\mathrm{d}}{\mathrm{d}a}\mathrm{I}_m(a) = \mathrm{I}_{m+1}(a) + \frac{m}{a} \cdot \mathrm{I}_m(a) \tag{B.5}$$

such that the modified Bessel function of a particular order can be computed recursively from the modified Bessel function of a lower order (indeed, this is how many computer programmes calculate the modified Bessel function).

AMOS (1974) introduces several further results that are not only useful for the characterisation of the modified Bessel function, but also for its computation. Adapting the notation of that paper to suit,

Amos (1974) shows that, for all $\kappa_q \geq \kappa_r > 0$,

$$\underline{\mathfrak{b}}\left(\kappa_q, \kappa_r, m\right) \leq \ln\left(\mathrm{I}_m\left(\kappa_q\right)\right) \leq \overline{\mathfrak{b}}\left(\kappa_q, \kappa_r, m\right) \tag{B.6}$$

where

$$\underline{\mathfrak{b}}\left(\kappa_q, \kappa_r, m\right) \doteq \ln\left(\mathrm{I}_m\left(\kappa_r\right)\right) + m \cdot \ln\left(\frac{\kappa_q}{\kappa_r}\right) + \underline{a}_m \cdot \ln\left(\frac{\underline{a}_m + \sqrt{\kappa_r^2 + \overline{a}_m^2}}{\underline{a}_m + \sqrt{\kappa_q^2 + \overline{a}_m^2}}\right) + \frac{\kappa_q^2 - \kappa_r^2}{\sqrt{\kappa_q^2 + \overline{a}_m^2} + \sqrt{\kappa_r^2 + \overline{a}_m^2}} \tag{B.7}$$

and

$$\overline{\mathfrak{b}}\left(\kappa_q, \kappa_r, m\right) \doteq \ln\left(\mathrm{I}_m\left(\kappa_r\right)\right) + m \cdot \ln\left(\frac{\kappa_q}{\kappa_r}\right) + \underline{a}_m \cdot \ln\left(\frac{\underline{a}_m + \sqrt{\kappa_r^2 + \underline{a}_m^2}}{\underline{a}_m + \sqrt{\kappa_q^2 + \underline{a}_m^2}}\right) + \frac{\kappa_q^2 - \kappa_r^2}{\sqrt{\kappa_q^2 + \underline{a}_m^2} + \sqrt{\kappa_r^2 + \underline{a}_m^2}} \tag{B.8}$$

given $\underline{a}_m \doteq m + 1/2$ and $\overline{a}_m \doteq m + 3/2$. I note that Equation B.6 is reversed when the arguments instead satisfy $\kappa_r > \kappa_q$. In the special case where $\kappa_r \to 0$, Amos (1974) shows that Equations B.7 and B.8 reduce to

$$\underline{\mathfrak{b}}\left(\kappa_q, 0, m\right) = \frac{1}{2} \cdot \ln\left(\frac{2}{\kappa_q}\right) - \ln\left(\Gamma\left(m+1\right)\right) + \underline{a}_m \cdot \ln\left(\frac{\kappa_q \cdot \left(\underline{a}_m + \overline{a}_m\right)/2}{\underline{a}_m + \sqrt{\kappa_q^2 + \overline{a}_m^2}}\right) + \frac{\kappa_q^2}{\overline{a}_m + \sqrt{\kappa_q^2 + \overline{a}_m^2}} \tag{B.9}$$

and

$$\overline{\mathfrak{b}}\left(\kappa_q, 0, m\right) = \frac{1}{2} \cdot \ln\left(\frac{2}{\kappa_q}\right) - \ln\left(\Gamma\left(m+1\right)\right) + \underline{a}_m \cdot \ln\left(\frac{\kappa_q \cdot \left(\underline{a}_m + \underline{a}_m\right)/2}{\underline{a}_m + \sqrt{\kappa_q^2 + \underline{a}_m^2}}\right) + \frac{\kappa_q^2}{\underline{a}_m + \sqrt{\kappa_q^2 + \underline{a}_m^2}} \tag{B.10}$$

respectively. It is self-evident that Equation B.10 exceeds Equation B.9 as is required for Equation B.6 to be non-empty. An implication of Equation B.6 is that

$$\frac{\kappa_q}{\underline{a}_m + \sqrt{\kappa_q^2 + \overline{a}_m^2}} \leq \frac{I_{m+1}\left(\kappa\right)}{\mathrm{I}_m\left(\kappa\right)} \leq \frac{\kappa_q}{\underline{a}_m + \sqrt{\kappa_q^2 + \underline{a}_m^2}} \tag{B.11}$$

which defines a subset of the unit interval as is evident from Equation B.4.

—| Section 3 |—

Limiting behaviour

The modified Bessel function satisfies the limiting behaviour (NIST, 2021, §10.30)

$$\frac{\mathrm{I}_m\left(a\right)}{a^m} \simeq \frac{1}{2^m \cdot \Gamma\left(v+1\right)} \tag{B.12}$$

which is proportional to the surface area of the hypersphere. Conversely, the modified Bessel function admits, for all $n \in \mathbb{N}_*$, the Poincaré asymptotic expansion (NIST, 2021, §2.1)

$$\mathrm{I}_m(a) = \frac{\exp(a)}{\sqrt{2\pi \cdot a}} \cdot \left(\sum_{j=0}^{n-1} (-1)^j \cdot \frac{\mathrm{Pochhammer}_j(m)}{a^j} + \mathrm{O}\left(\frac{1}{a^n}\right) \right) \tag{B.13}$$

which is derived from Hankel's expansion (NIST, 2021, §10.17 and §10.40) and which is defined in terms of a factorial sequence (NIST, 2021, §5.2—Pochhammer's symbol) that is polynomial in its argument. It is of no surprise that the modified Bessel function grows at an exponential rate given the appearance of Figure B.1, but this is easily verified from Equation B.13.

—| Chapter C |—

The f-divergence of a von Mises-Fisher distribution from some reference distributions

The von Mises-Fisher family of distributions is well-known in the field of directional statistics[1] but is foreign to economics and, as such, warrants some introduction. Also known as the Langevin family (Watson, 1984), the von Mises-Fisher family recognises those two titans of statistics, Sir Ronald Fisher and Richard von Mises, for their seminal contributions in considering Gaussianity on the circle (von Mises, 1918) and on the sphere (R. A. Fisher, 1953). Subsequent work has generalised the von Mises-Fisher family to $\mathbb{S}^i$, and has led to the definition of other related parametric distributions such as the Bingham family (Bingham, 1974) and the Fisher-Bingham or Kent family (Kent, 1982).

A von Mises-Fisher distribution assigns probability mass to the surface of the unit ball—the hypersphere. As such, the von Mises-Fisher family is relevant to situations where the researcher is interested in either the sampling of directional vectors—i.e., vectors of unit length—or in the clustering of some phenomenon on a circular object, such as occurs if data is periodic. Applications range from the study of sea turtle navigation (Hillen et al., 2017), to the study of perihelia of long-tailed comets (Mardia, 1975) and near-earth objects (Sei et al., 2013), as well as to the study of patient arrival data (Mardia, 1975). Sabelfeld (2018) even links the von Mises-Fisher family to the solving of high-dimensional diffusion-advection-reaction equations. The von Mises-Fisher family is a two-parameter family, summarised by a concentration parameter (or, simply, concentration), which we denote by $\kappa > 0$, and a mean direction, which we denote by $\mu \in \mathbb{S}^i$.

The main contribution of this chapter is to provide analytical expressions for the f-divergence of a von Mises-Fisher distribution from two relevant reference distributions given several common choices of function. We study the broad class of Rényi divergence of simple order as well as several other measures of (statistical) divergence that relate to the Rényi class—the χ-square distance, the squared-Hellinger distance and the Kullback-Leibler divergence. Each is, of course, a measure of the difference between

1 See Mardia and Jupp (2009) for a summary of important results in the field of directional statistics.

two probability distributions. Several well-known inequalities relate these measures to the total variation distance (Bretagnolle and Huber, 1978; Pinsker, 1964), which is often of interest. The reference distributions that we specify are another, distinct, von Mises-Fisher distribution on $\mathbb{S}^i$ and the uniform distribution on the hypersphere. We are unaware of such expressions being available elsewhere. Alongside these expressions, we characterise how the various measures of divergence that we consider change as a von Mises-Fisher distribution becomes increasingly concentrated and degenerate. In particular, we clarify the leading order terms of asymptotic expansions,[2] which relies on results in Amos (1974). These asymptotic expansions offer analytically tractable polynomial approximations of each of the measures of divergence that we consider in terms of the concentration parameter, with these approximations accurate when this parameter takes large values.

Obtaining analytical expressions of statistical divergence is useful for implementing minimum distance-type or penalised estimation methods, and also for characterising the statistical performance of these procedures. See, for instance, Kitagawa et al. (2022b), which builds upon the analytical expression of the Kullback-Leibler divergence that is derived in this chapter to estimate the randomised treatment assignment rule that minimises a penalised empirical welfare criterion. Polynomial approximation of statistical divergences for parameter values corresponding to high concentration are useful for characterising the convergence behaviour of numerous objects including the concentration rate of the posterior distribution and of the PAC-Bayes regret bounds that are formulated in Kitagawa et al. (2022b), and that are exploited in that paper.

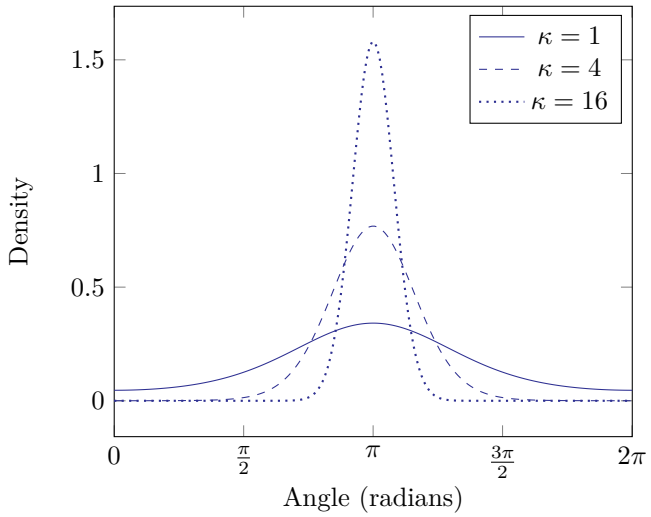
Analytical expressions for the moments and other known distributional features of the von Mises-Fisher family are, to varying extents, available elsewhere. These include statement of the first two moments of a von Mises-Fisher distribution (available in or adaptable from Mardia and Jupp, 2009, §9.3 and §9.6, respectively) and its associated Fisher information matrix (Hornik and Grün, 2013). Dhillon and Sra (2003), Hillen et al. (2017), and Hornik and Grün (2013) demonstrate three distinct approaches to obtaining expressions for these moments. These are integration by substitution following transformation to spherical coordinates, application of the divergence theorem,[3] and differentiation of the moment-generating function, respectively. We use several of the results in these papers directly.[4] In particular, knowledge of the first moment is essential to characterising the divergence of a von Mises-Fisher distribution from our chosen reference distributions.

2 In the appendices, we demonstrate how Hankel expansions can be performed to complement results given in Kitagawa et al. (2022b).

3 The divergence theorem relates the area of a surface integral to a volume integral.

4 We otherwise establish these results in Section 2.e for the benefit of the reader, should they be unfamiliar with spherical distributions and their associated concepts. Our use of the moment-generating function is very much a case of horses for courses, to use the British idiom: integration by susbstitution and the divergence theorem can also be used, but their application is more involved in this setting.

FIGURE C.1
von Mises-Fisher distributions on the circle



The density function of a von Mises-Fisher distribution on the circle with mean direction $(-1, 0)$—i.e., a polar orientation with reference angle equal to π radians—for several values of the concentration parameter. A spherical coordinate system is used.

Directional objects are common in economics, and the von Mises-Fisher family of distributions is relevant to many environments and several methods. For instance, consider the canonical binary choice model with latent random utility. The rational choice of the individual, which we denote by $t \in \{0, 1\}$, is determined according to the linear index equation

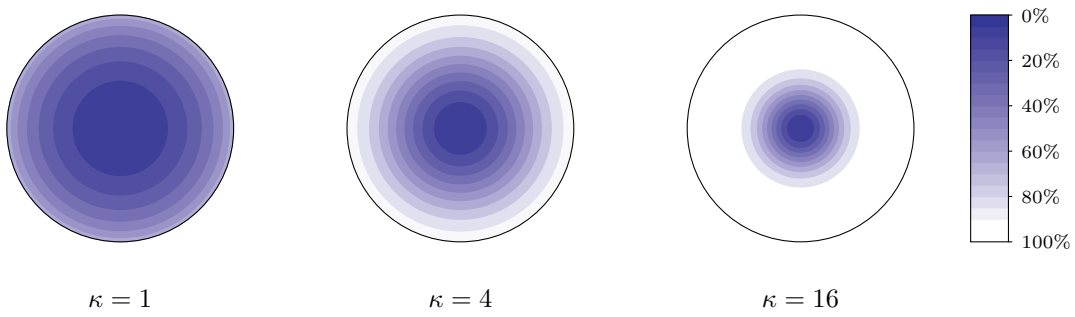
$$t = 1(\boldsymbol{x}^\top \gamma - u \geq 0) \tag{C.1}$$

where $\boldsymbol{x} \in \mathbb{R}^i$ and $u \in \mathbb{R}^i$ denote the individual's observable characteristics (including an intercept) and latent heterogeneity, respectively. The conditional zero-median restriction inherent in the semiparametric maximum score approach of MANSKI (1975, 1985) does not identify the scale of the utility coefficients $\gamma \in \mathbb{R}^i$. It is common to normalise the parameter space of γ to the collection of vectors satisfying $\|\gamma\|_2 = 1$—i.e., to the hypersphere. Similarly, in the context of statistical treatment choice (MANSKI, 2004a), KITAGAWA and TETENOV (2018a) considers individualised treatment assignment rules based upon a linear index,

$$t = 1(\boldsymbol{x}^\top \gamma \geq 0) \tag{C.2}$$

where $\boldsymbol{x}^\top \gamma \in \mathbb{R}$ is an eligibility score that aggregates the individual's observable characteristics and determines whether she should be assigned to treatment—i.e., if $\boldsymbol{x} \in \mathbb{R}^i$ maps to $t = 1$—or to non-treatment—i.e., if $\boldsymbol{x} \in \mathbb{R}^i$ maps to $t = 0$. Such assignment rules are invariant to multiplication of the eligibility score by a positive constant and can be uniquely indexed by a parameter vector on the

FIGURE C.2
von Mises-Fisher distributions on the sphere



Orthogonal projection of the sphere, oriented to the mean direction, for several values of the concentration parameter. Each contour describes a region in which the von Mises-Fisher distribution assigns 10% mass, with contours distinguished by their shading. As the value of the concentration parameter is increased, more mass is assigned to the vicinity of the pole.

hypersphere.

Optimising maximum score or an empirical welfare criterion is difficult, however, and complicates estimation of and inference on γ. This motivates a quasi-Bayesian approach as considered in CHERNOZHUKOV and HONG (2003a). The von Mises-Fisher family offers a parsimonious and convenient prior specification over γ within the quasi-Bayesian framework, with prior elicitation facilitated by knowledge of the moments of the distribution. In a related but different context, PAC-Bayesian analysis, which is widely studied in machine learning (ALQUIER et al., 2016; CATONI, 2007; GERMAIN et al., 2009; MCALLESTER, 2003, to name but a few relevant papers), considers exponentiated negative empirical risk as a quasi-likelihood, and forms a posterior distribution over prediction rules. The von Mises-Fisher family is then not only useful as a specified prior over directional parameters, but can also be used to approximate a posterior distribution over γ in the linear classification rule or linear index treatment assignment rule settings (KITAGAWA et al., 2022b).

Another context where the parameter space is isomorphic to the hypersphere is the class of underidentifying linear simultaneous equation models in which the imposed model restrictions identify structural parameters up to sets of orthonormal transformations. See, for instance, ARIAS et al. (2018), GIACOMINI and KITAGAWA (2021a), and UHLIG (2005) for a class of set-identified structural vector autoregressions in which the identified set of an impulse-response is spanned by the class of orthonormal matrices. The isomorphism of the hypersphere and the orthogonal group suggests that a spherical distribution—and a von Mises-Fisher distribution in particular—can be used as a prior distribution for the non-identified orthonormal matrices. When combined with prior elicitation of the reduced-form parameters, the moments of a von Mises-Fisher distribution can be used to translate a belief about the structural parameters into a prior distribution over the non-identified orthonormal matrices. Like the Gaussian family of distribu-

tions on the hyperplane from which it can be derived, the von Mises-Fisher family is highly restrictive[5] but, nonetheless, forms an interesting baseline case to study.

We are unaware of any paper that characterises the f-divergence of a von Mises-Fisher distribution as we do. For instance, Diethe (2015) similarly studies the Kullback-Leibler divergence of von Mises-Fisher distributions. Whereas we provide exact analytical expressions, Diethe (2015) either provides upper bounds on the Kullback-Leibler divergence, or else provides analytical expressions that rely on an approximation that is valid only when the von Mises-Fisher distribution is close to the uniform distribution over the hypersphere, something that we do not rely on. Where updating of the von Mises-Fisher distribution has been considered, this appears to have mainly centred on the likelihood function and its characterisation rather than on measures of divergence *per se*. We refer to Lin et al. (2017) and Mardia and El-Atoum (1976) as pertinent examples.

A von Mises-Fisher distribution constitutes a conjugate prior (Mardia and El-Atoum, 1976). Our choice of reference distributions—another, distinct, von Mises-Fisher distribution on $\mathbb{S}^i$ and the uniform distribution on the hypersphere—reflects both this and the prevalence of the uniform distribution in practice. Another commonly invoked choice that we do not consider is the Jeffreys prior, which is proportional to the square root of the determinant of the Fisher information matrix relative to the parametrisation employed.[6] Hornik and Grün (2013) derive the Fisher information matrix and its determinant and show that the Jeffreys prior is improper in this setting.

—| Section 1 |—

The probability density function and moments of the von Mises-Fisher family

Throughout this chapter, we exploit the fact that the von Mises-Fisher family is an exponential family and, accordingly, we adopt the terminology that is used in conjunction with that well-known class. Defining $m \doteq p/2 - 1$ for convenience and maintaining $\kappa \geq 0$ and $\|\mu\|_2 = 1$, we write the probability density function of a von Mises-Fisher random vector, which we denote by $\boldsymbol{V} \in \mathbb{S}^{p-1}$, for integer $p > 1$, as

$$f(\boldsymbol{v}; \kappa, \mu) \doteq \frac{\exp(\kappa \cdot \mu^\top \boldsymbol{v})}{\text{ℵK}(\kappa, m)} \tag{C.3}$$

5 The von Mises-Fisher family is akin to the class of Gaussian distributions that feature a diagonal variance matrix—i.e., statistically independent Gaussian random variables—with the same entry on each element of the diagonal.

6 Our parametrisation of the von Mises-Fisher family specifies a concentration and a mean direction—what Hornik and Grün (2013) call a polar coordinate parametrisation. Equally, one could combine the concentration and mean direction into a single parameter vector, η say, such that the concentration corresponds to $\|\eta\|_2$ and the mean direction corresponds to $\eta/\|\eta\|_2$. This is the parametrisation that Hornik and Grün (2013) specifies.

where we reiterate that κ and μ are the concentration and mean direction, respectively, and where, for all $\kappa \geq 0$,

$$\text{Ж}(\kappa, m) \doteq \int_{\mathbb{S}^{p-1}} \exp\left(\kappa \cdot \mu^\top \boldsymbol{v}\right) \cdot \mathrm{d}\boldsymbol{v} = \begin{cases} [2\pi]^{m+1} \cdot \mathrm{I}_m(\kappa) / \kappa^m & \text{if } \kappa > 0 \\ 2\pi^{m+1} / \Gamma(m+1) & \text{if } \kappa = 0 \end{cases} \tag{C.4}$$

thereby guaranteeing that the density function integrates to one.[7] We recall the definition of the modified Bessel function in Chapter B. We refer to the exponentiation—i.e., the numerator in Equation C.3—as the kernel of the density function, and to the normalising constant—i.e., the denominator in Equation C.3—as the partition function. We emphasise that the integral in Equation C.4 is over the hypersphere and it is this fact that makes derivation of statistical features of the von Mises-Fisher random vector difficult. Moreover, we note that our choice of parametrisation is but one way that a von Mises-Fisher distribution can be parametrised. Another parametrisation that is better suited to certain analyses of von Mises-Fisher distributions (in particular, derivation of their moments) is presented in Hornik and Grün (2013).

The von Mises-Fisher family is the hyperspherical analogue of the Gaussian family, which is informative as to its shape: the von Mises-Fisher family is unimodal and symmetric about its mean direction, with the concentration parameter determining the degeneracy (when $\kappa \to \infty$) and uniformity (when $\kappa \to 0$) of the distribution, and it assigns positive density to the entirety of the hypersphere. This relationship is shown by appropriately normalising the probability density function of statistically independent Gaussian random variables with variance $1/\kappa$ that are distributed on the hypersphere. Importantly, the von Mises-Fisher family is rotationally equivariant, which is the hyperspherical analogue of the translation equivariance property that is exhibited by the Gaussian family.

The (centred) moments of the von Mises-Fisher family can be obtained by differentiating its moment-generating function. Since the von Mises-Fisher family is an exponential family, the moment-generating function is equal to the log-partition function. Whilst the partition function of the von Mises-Fisher family has a closed-form expression, this is not necessarily true for other named directional families. For instance, we are unaware of any closed-form expressions for the partition functions of both the Kent and von Mises families of distributions beyond the bivariate case.[8] Differentiating the log-partition function,

7 See Dhillon and Sra (2003, §B.2) for proof of this statement, and also for derivation of the determinant of the Jacobian matrix in the generalised spherical coordinate transform of this integral.

8 Whilst numerical methods can be used to compute the moments of the Kent and von Mises families (see Best and N. I. Fisher, 1979; Kent et al., 2013 for suitable rejection sampling routines; also Mardia et al., 2008, which establishes that the conditional von Mises density is itself von Mises), the lack of closed-form expressions for these other named directional families certainly limits their tractability and usefulness in applications featuring a high-dimensional covariate or outcome vector.

we obtain

$$\mathrm{E}\left(\boldsymbol{V}\right) = \frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)} \cdot \mu \tag{C.5}$$

which exploits results in NIST (2021, §10.29). The higher-order moments of the von Mises-Fisher family can be derived via recursive differentiation.[9]

Equation C.5 is adapted from results that are available in Mardia and Jupp (2009), which also discusses several other important results, including the asymptotic and high-concentration behaviour of von Mises-Fisher random vectors and their tangent normal vectors. We emphasise that the expression for the variance that we present here is distinct from the circular variance, which is simply the distance of the mean resultant length from the surface—i.e., one minus the mean resultant length.

—| Section 2 |—

Measuring the f-divergence of an obtained distribution from a reference distribution

To facilitate consideration of the f-divergence of an obtained distribution from a reference distribution, we suppose that, for all $\boldsymbol{a} \subset \mathbb{S}^{p-1}$,

$$\begin{aligned} \Pr\left(\boldsymbol{Q} \in \boldsymbol{a}\right) &= \int_{\boldsymbol{a}} f\left(\boldsymbol{v}; \kappa_q, \mu_q\right) \cdot \mathrm{d}\boldsymbol{v} \\ \Pr\left(\boldsymbol{R} \in \boldsymbol{a}\right) &= \int_{\boldsymbol{a}} f\left(\boldsymbol{v}; \kappa_r, \mu_r\right) \cdot \mathrm{d}\boldsymbol{v} \end{aligned} \tag{C.8}$$

such that $\boldsymbol{Q} \in \mathbb{S}^{p-1}$ and $\boldsymbol{R} \in \mathbb{S}^{p-1}$ are two von Mises-Fisher random vectors with the corresponding mean direction and concentration parameters satisfying all of the usual properties that we maintain.[10] We use the notation above to distinguish the two von Mises-Fisher random vectors from the κ-concentrated μ-oriented $\boldsymbol{V}$ that is previously discussed.

We adopt the convention of referring to the probability distribution of the random vector $\boldsymbol{Q}$ as the ob-

9 For reference, the second moment of the von Mises-Fisher family is

$$\mathrm{Variance}\left(\boldsymbol{V}\right) = \frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)} \cdot \left[\frac{1}{\kappa} \cdot \mathbf{I}_p + \left[\frac{\mathrm{I}_{m+2}\left(\kappa\right)}{\mathrm{I}_{m+1}\left(\kappa\right)} - \frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)}\right] \cdot \mu\mu^{\top}\right] \tag{C.6}$$

where $\mathbf{I}_p$ is the identity matrix of subscripted dimension; this expression can otherwise be written as

$$\mathrm{Variance}\left(\boldsymbol{V}\right) = \frac{1}{\kappa} \cdot \frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)} \cdot \mathbf{I}_p + \left[1 - \frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)} \cdot \left[\frac{p}{\kappa} - \frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)}\right]\right] \cdot \mu\mu^{\top} \tag{C.7}$$

which exploits recurrence relations in NIST (2021, §10.29). As noted in Hornik and Grün (2013), given that the von Mises-Fisher family is an exponential family, its second moment coincides with its Fisher information matrix under the parametrisation considered in that paper. We derive the first and second moments of the von Mises-Fisher family in the appendices.

10 A necessary condition that we require is that the probability distribution of the random vector $\boldsymbol{Q}$ is absolutely continuous with respect to the probability distribution of the random vector $\boldsymbol{R}$ (Lattimore and Szepesvári, 2020, §2.7 and §14.5), thereby guaranteeing that each of the measures that we consider is well-defined. The von Mises-Fisher family clearly satisfies this requirement, assigning positive density to all points on the hypersphere.

tained distribution and to the probability distribution of the random vector $\boldsymbol{R}$ as the reference distribution, and of measuring the divergence of the obtained distribution from the reference distribution.

We reiterate that the uniform distribution on the hypersphere corresponds to the limiting case where the concentration of the von Mises-Fisher distribution is zero. As such, we emphasise that our framework is also compatible with the reference distribution being equal to the uniform distribution on the hypersphere.

For all of the measures that we consider, each measure approaches its maximum possible value when the obtained distribution is degenerate. We defer proof of each proposition in this section to the appendices.

With the exception of the total variation distance, each of the measures that we consider has a closed-form expression that depends only upon the parameters of the two distributions whose divergence they capture (and even then, most of the expressions that we obtain simply feature the concentration parameter). These expressions feature ratios of modified Bessel functions, whether of consecutive order or of the same order. Various statistical programming languages have inbuilt functionality to compute modified Bessel functions that perform well for comparatively small values of their argument (i.e., for low concentrations) but that can break for large values. Amos (1974) derives some inequalities (see the appendices) that are robust to large arguments and that can be used when inbuilt functionality breaks.

Whilst one or more of these measures may be of ultimate interest, it is perhaps more likely that they form only a constituent part of what the researcher is interested in (see, for instance, Kitagawa et al., 2022b). The rejection sampling algorithm of Wood (1994; see Tsagris et al., 2022, which implements the suggested algorithm, for further details) can be used to sample vectors from a von Mises-Fisher distribution as part of a routine to approximate related objects that do not have a closed-form expression.

Asymptotic calculations can be performed using Hankel series expansion (see NIST, 2021, §10.17 and §10.40, which rely upon §2.1 and the Poincáre expression therein). We include a demonstration in the appendices, applying a Hankel series expansion to the circular variance of the von Mises-Fisher family of distributions,[11] which is defined as one minus the ratio of modified Bessel functions—i.e., one minus the mean resultant length. We note that Hankel series expansion is appropriate when $\kappa \to \infty$, which is the limiting behaviour that we are interested in. We emphasise that the circular variance is strictly contained in the unit interval, which is asymptotically guaranteed by the fact that $p \geq 2$—i.e., the minimal hypersphere (the circle) is defined on the real plane.

11 See Mardia and Jupp (2009) for further, general, details about the circular variance.

—| Subsection 2.a |—

Rényi divergence

Following van Erven and Harremoës (2014), we define the Rényi divergence of simple order $0 < \alpha < 1$ or $1 < \alpha < \infty$ as

$$d_{\alpha}\left(\boldsymbol{Q}, \boldsymbol{R}\right) \doteq \frac{1}{\alpha - 1} \cdot \ln\left(\int_{\mathbb{S}^{p-1}} f\left(\boldsymbol{v}; \kappa_{q}, \mu_{q}\right)^{\alpha} \cdot f\left(\boldsymbol{v}; \kappa_{r}, \mu_{r}\right)^{1-\alpha} \cdot \mathrm{d}\boldsymbol{v}\right) \tag{C.9}$$

The Rényi divergence is the most general measure of f-divergence that we consider and describes a broad class that relates to the other measures that we study—namely, the χ-squared distance, the squared-Hellinger distance and the Kullback-Leibler divergence.

Proposition C.1. *Given $\boldsymbol{Q} \in \mathbb{S}^{p-1}$ and $\boldsymbol{R} \in \mathbb{S}^{p-1}$ are distributed as p-variate von Mises-Fisher random vectors with concentrations $\kappa_{q} > 0$ and $\kappa_{r} > 0$, respectively, and mean directions $\mu_{q} \in \mathbb{S}^{p-1}$ and $\mu_{r} \in \mathbb{S}^{p-1}$, respectively, and recalling the definition of m, the Rényi divergence of simple order $0 < \alpha < 1$ or $1 < \alpha < \infty$ is*

$$d_{\alpha}\left(\boldsymbol{Q}, \boldsymbol{R}\right) = \frac{m}{\alpha - 1} \cdot \ln\left(\frac{\kappa_{q}^{\alpha} \cdot \kappa_{r}^{1-\alpha}}{\kappa_{\alpha}}\right) + \frac{\alpha}{\alpha - 1} \cdot \ln\left(\frac{\mathrm{I}_{m}\left(\kappa_{\alpha}\right)}{\mathrm{I}_{m}\left(\kappa_{q}\right)}\right) - \ln\left(\frac{\mathrm{I}_{m}\left(\kappa_{\alpha}\right)}{\mathrm{I}_{m}\left(\kappa_{r}\right)}\right) \tag{C.10}$$

where

$$\kappa_{\alpha} \doteq \left\|\alpha \cdot \kappa_{q} \cdot \mu_{q} + (1 - \alpha) \cdot \kappa_{r} \cdot \mu_{r}\right\|_{2} \tag{C.11}$$

In the special case where $\kappa_{\alpha} = 0$,

$$d_{\alpha}\left(\boldsymbol{Q}, \boldsymbol{R}\right) = \frac{m}{\alpha - 1} \cdot \ln\left(\frac{\kappa_{q}^{\alpha} \cdot \kappa_{r}^{1-a}}{2}\right) + \frac{\alpha}{\alpha - 1} \cdot \ln\left(\frac{1/\Gamma\left(m + 1\right)}{\mathrm{I}_{m}\left(\kappa_{q}\right)}\right) - \ln\left(\frac{1/\Gamma\left(m + 1\right)}{\mathrm{I}_{m}\left(\kappa_{r}\right)}\right) \tag{C.12}$$

with

$$\kappa_{r} = \frac{\alpha}{|1 - \alpha|} \cdot \kappa_{q} \text{ and } \mu_{r} + \mathrm{Sign}\left(1 - \alpha\right) \cdot \mu_{q} = 0 \tag{C.13}$$

In the special case where $\boldsymbol{R}$ is, instead, uniformly distributed on the hypersphere,

$$d_{\alpha}\left(\boldsymbol{Q}, \boldsymbol{R}\right) = \frac{m}{\alpha - 1} \cdot \ln\left(\frac{2^{1-\alpha}}{\alpha \cdot \kappa_{q}^{1-\alpha}}\right) + \frac{\alpha}{\alpha - 1} \cdot \ln\left(\frac{\mathrm{I}_{m}\left(\alpha \cdot \kappa_{q}\right)}{\mathrm{I}_{m}\left(\kappa_{q}\right)}\right) - \ln\left(\frac{\mathrm{I}_{m}\left(\alpha \cdot \kappa_{q}\right)}{1/\Gamma\left(m + 1\right)}\right) \tag{C.14}$$

The Rényi divergence is an $\mathrm{O}\left(\ln\left(\kappa_{q}\right)\right)$ function.

—| Subsection 2.b |—

χ-square distance

We define the χ-square distance as

$$d_{\chi}\left(\boldsymbol{Q},\boldsymbol{R}\right) \doteq \int_{\mathbb{S}^{p-1}} \frac{\left[f\left(\boldsymbol{v};\kappa_q,\mu_q\right)-f\left(\boldsymbol{v};\kappa_r,\mu_r\right)\right]^2}{f\left(\boldsymbol{v};\kappa_r,\mu_r\right)}\cdot \mathrm{d}\boldsymbol{v} \tag{C.15}$$

$$= \int_{\mathbb{S}^{p-1}} \frac{f\left(\boldsymbol{v};\kappa_q,\mu_q\right)^2}{f\left(\boldsymbol{v};\kappa_r,\mu_r\right)}\cdot \mathrm{d}\boldsymbol{v} - 1 \tag{C.16}$$

with Equation C.16 being the more convenient definition to work with. We observe that the χ-square distance relates to the Rényi class of measures via the equality relation

$$d_{\alpha}\left(\boldsymbol{Q},\boldsymbol{R}\right)|_{\alpha=2} = \ln\left(1+d_{\chi}\left(\boldsymbol{Q},\boldsymbol{R}\right)\right) \tag{C.17}$$

as per van Erven and Harremoës (2014), which serves to illustrate the breadth of the Rényi class.

Proposition C.2. *Given $\boldsymbol{Q}\in\mathbb{S}^{p-1}$ and $\boldsymbol{R}\in\mathbb{S}^{p-1}$ are distributed as p-variate von Mises-Fisher random vectors with concentrations $\kappa_q>0$ and $\kappa_r>0$, respectively, and mean directions $\mu_q\in\mathbb{S}^{p-1}$ and $\mu_r\in\mathbb{S}^{p-1}$, respectively, and recalling the definition of m, the χ-square distance is*

$$d_{\chi}\left(\boldsymbol{Q},\boldsymbol{R}\right) = \frac{\kappa_q^{2m}\cdot \mathrm{I}_m\left(\kappa_{\chi}\right)\cdot \mathrm{I}_m\left(\kappa_r\right)}{\kappa_{\chi}^m\cdot\kappa_r^m\cdot \mathrm{I}_m\left(\kappa_q\right)^2} - 1 \tag{C.18}$$

where

$$\kappa_{\chi} \doteq \left\|2\kappa_q\cdot\mu_q-\kappa_r\cdot\mu_r\right\|_2 \tag{C.19}$$

In the special case where $\kappa_{\chi}=0$,

$$d_{\chi}\left(\boldsymbol{Q},\boldsymbol{R}\right) = \frac{\kappa_q^m\cdot \mathrm{I}_m\left(2\kappa_q\right)}{4^m\cdot\Gamma\left(m+1\right)\cdot \mathrm{I}_m\left(\kappa_q\right)^2} - 1 \tag{C.20}$$

with

$$\kappa_r = 2\kappa_q \text{ and } \mu_r-\mu_q=0 \tag{C.21}$$

In the special case where $\boldsymbol{R}$ is, instead, uniformly distributed on the hypersphere, the χ-square distance coincides with Equation C.20. The χ-square distance is an $\mathrm{O}\left(\kappa_q\right)$ function.

—| Subsection 2.c |—

Squared-Hellinger distance

We define the squared-Hellinger distance as

$$d_h\left(\boldsymbol{Q},\boldsymbol{R}\right)^2 \doteq \int_{\mathbb{S}^{p-1}} \left[\sqrt{f\left(\boldsymbol{v};\kappa_q,\mu_q\right)} - \sqrt{f\left(\boldsymbol{v};\kappa_r,\mu_r\right)}\right]^2 \cdot \mathrm{d}\boldsymbol{v} \tag{C.22}$$

$$= 2\left[1 - \int_{\mathbb{S}^{p-1}} \sqrt{f\left(\boldsymbol{v};\kappa_q,\mu_q\right)\cdot f\left(\boldsymbol{v};\kappa_r,\mu_r\right)} \cdot \mathrm{d}\boldsymbol{v}\right] \tag{C.23}$$

with Equation C.23 being the more convenient definition to work with. We observe that the squared-Hellinger distance relates to the Rényi class of measures via the equality relation

$$d_\alpha\left(\boldsymbol{Q},\boldsymbol{R}\right)\big|_{\alpha=1/2} = -2\ln\left(1 - \frac{d_h\left(\boldsymbol{Q},\boldsymbol{R}\right)^2}{2}\right) \tag{C.24}$$

as per van Erven and Harremoës (2014), which serves to illustrate the breadth of the Rényi class and is also interpretable as twice the negative logarithm of the Bhattacharyya coefficient (Bhattacharyya, 1943). The Bhattacharyya coefficient is an approximate measure of the amount of overlap between two probability distributions, such that when the obtained and reference distributions are close (i.e., they have a similar concentration and mean direction) then the squared-Hellinger distance is small in absolute value, as is to be expected.

Proposition C.3. *Given $\boldsymbol{Q} \in \mathbb{S}^{p-1}$ and $\boldsymbol{R} \in \mathbb{S}^{p-1}$ are distributed as p-variate von Mises-Fisher random vectors with concentrations $\kappa_q > 0$ and $\kappa_r > 0$, respectively, and mean directions $\mu_q \in \mathbb{S}^{p-1}$ and $\mu_r \in \mathbb{S}^{p-1}$, respectively, and recalling the definition of m, the squared-Hellinger distance is*

$$d_h\left(\boldsymbol{Q},\boldsymbol{R}\right)^2 = 2\left[1 - \sqrt{\frac{\kappa_q^m\cdot\kappa_r^m\cdot\mathrm{I}_m\left(\kappa_h\right)^2}{\kappa_h^{2m}\cdot\mathrm{I}_m\left(\kappa_q\right)\cdot\mathrm{I}_m\left(\kappa_r\right)}}\right] \tag{C.25}$$

where we define

$$\kappa_h \doteq \frac{1}{2}\cdot\left\|\kappa_q\cdot\mu_q + \kappa_r\cdot\mu_r\right\|_2 \tag{C.26}$$

In the special case where $\kappa_h = 0$,

$$d_h\left(\boldsymbol{Q},\boldsymbol{R}\right)^2 = 2\left[1 - \frac{\kappa_q^m}{2^m\cdot\mathrm{I}_m\left(\kappa_q\right)\cdot\Gamma\left(m+1\right)}\right] \tag{C.27}$$

with

$$\kappa_r = \kappa_q \text{ and } \mu_r + \mu_q = 0 \tag{C.28}$$

In the special case where $\boldsymbol{R}$ *is, instead, uniformly distributed on the hypersphere,*

$$d_h\left(\boldsymbol{Q}, \boldsymbol{R}\right)^2 = 2\left[1 - \sqrt{\frac{2^{3m} \cdot \mathrm{I}_m\left(\kappa_q/2\right)^2 \cdot \Gamma\left(m+1\right)}{\kappa_q^m \cdot \mathrm{I}_m\left(\kappa_q\right)}}\right] \tag{C.29}$$

The squared-Hellinger distance is an $\mathrm{O}\left(1 - 1/\kappa_q\right)$ *function.*

—| Subsection 2.d |—

Kullback-Leibler divergence

We define the Kullback-Leibler divergence as

$$d_\ell\left(\boldsymbol{Q}, \boldsymbol{R}\right) \doteq \int_{\mathbb{S}^p} \ln\left(\frac{f\left(\boldsymbol{v}; \kappa_q, \mu_q\right)}{f\left(\boldsymbol{v}; \kappa_r, \mu_r\right)}\right) \cdot f\left(\boldsymbol{v}; \kappa_q, \mu_q\right) \cdot \mathrm{d}\boldsymbol{v} \tag{C.30}$$

We observe that the Kullback-Leibler divergence is a limiting case of the Rényi divergence. That is,

$$\lim_{\alpha \to 1} d_\alpha\left(\boldsymbol{Q}, \boldsymbol{R}\right) = d_\ell\left(\boldsymbol{Q}, \boldsymbol{R}\right) \tag{C.31}$$

as per van Erven and Harremoës (2014), which serves to illustrate the breadth of the Rényi class.

Proposition C.4. *Given* $\boldsymbol{Q} \in \mathbb{S}^{p-1}$ *and* $\boldsymbol{R} \in \mathbb{S}^{p-1}$ *are distributed as p-variate von Mises-Fisher random vectors with concentrations* $\kappa_q > 0$ *and* $\kappa_r > 0$*, respectively, and mean directions* $\mu_q \in \mathbb{S}^{p-1}$ *and* $\mu_r \in \mathbb{S}^{p-1}$*, respectively, and recalling that* $m \doteq p/2 - 1$*, the Kullback-Leibler divergence is*

$$d_\ell\left(\boldsymbol{Q}, \boldsymbol{R}\right) = m \cdot \ln\left(\frac{\kappa_q}{\kappa_r}\right) - \ln\left(\frac{\mathrm{I}_m\left(\kappa_q\right)}{\mathrm{I}_m\left(\kappa_r\right)}\right) + \frac{I_{m+1}\left(\kappa_q\right)}{\mathrm{I}_m\left(\kappa_q\right)} \cdot \left[\kappa_q \cdot \mu_q - \kappa_r \cdot \mu_r\right]^\top \mu_q \tag{C.32}$$

In the special case where $\boldsymbol{R}$ *is, instead, uniformly distributed on the hypersphere,*

$$d_\ell\left(\boldsymbol{Q}, \boldsymbol{R}\right) = m \cdot \ln\left(\frac{\kappa_q}{2}\right) - \ln\left(\mathrm{I}_m\left(\kappa_q\right)\right) - \ln\left(\Gamma\left(m+1\right)\right) + \frac{I_{m+1}\left(\kappa_q\right)}{\mathrm{I}_m\left(\kappa_q\right)} \cdot \kappa_q \tag{C.33}$$

The Kullback-Leibler divergence is an $\mathrm{O}\left(\ln\left(\kappa_q\right)\right)$ *function.*

We note that the final terms of Equations C.32 and C.33 are proportional to the first moment of a von Mises-Fisher distribution (specifically, the first moment of the obtained distribution). In particular, we

note that the final term of Equation C.32 can otherwise be written as

$$\frac{I_{m+1}\left(\kappa_q\right)}{\mathrm{I}_m\left(\kappa_q\right)} \cdot \left[\kappa_q \cdot \mu_q - \kappa_r \cdot \mu_r\right]^\top \mu_q = \frac{I_{m+1}\left(\kappa_q\right)}{\mathrm{I}_m\left(\kappa_q\right)} \cdot \left[\kappa_q - \kappa_r \cdot \mu_r^\top \mu_q\right] \tag{C.34}$$

which we contrast with the corresponding term in Equation C.33. The difference between Equations C.33 and C.34 arises because the mean direction does not enter the probability density function of the uniform distribution on the hypersphere. The intuition here is that, in the special case where the reference distribution is the uniform distribution on the hypersphere, the mean direction can always be taken to be equal to the mean direction.

—| Subsection 2.e |—

Total variation distance

We define the total variation distance, for all measurable $\boldsymbol{a} \subset \mathbb{S}^{p-1}$, as

$$d_t\left(\boldsymbol{Q}, \boldsymbol{R}\right) \doteq \sup_{\boldsymbol{a}} \left| \int_{\boldsymbol{a}} f\left(\boldsymbol{v}; \kappa_q, \mu_q\right) - f\left(\boldsymbol{v}; \kappa_r, \mu_r\right) \cdot \mathrm{d}\boldsymbol{v} \right| \tag{C.35}$$

which is well-defined if the underlying probability space is endowed with the Borel σ-algebra—something that we, implicitly, maintain throughout all parts of our analysis.

Corollary C.1. *Given $\boldsymbol{Q} \in \mathbb{S}^{p-1}$ and $\boldsymbol{R} \in \mathbb{S}^{p-1}$ are distributed as p-variate von Mises-Fisher random vectors with concentrations $\kappa_q > 0$ and $\kappa_r > 0$, respectively, and mean directions $\mu_q \in \mathbb{S}^{p-1}$ and $\mu_r \in \mathbb{S}^{p-1}$, respectively, and recalling that $m \doteq p/2 - 1$, the total variation distance satisfies the inequality relations*

$$d_t\left(\boldsymbol{Q}, \boldsymbol{R}\right)^2 \leq d_h\left(\boldsymbol{Q}, \boldsymbol{R}\right)^2 \leq d_\ell\left(\boldsymbol{Q}, \boldsymbol{R}\right) \leq d_\chi\left(\boldsymbol{Q}, \boldsymbol{R}\right) \tag{C.36}$$

$$d_t\left(\boldsymbol{Q}, \boldsymbol{R}\right) \leq \sqrt{1 - \exp\left(-d_\ell\left(\boldsymbol{Q}, \boldsymbol{R}\right)\right)} \leq 1 - \frac{d_\ell\left(\boldsymbol{Q}, \boldsymbol{R}\right)}{2} \tag{C.37}$$

and, for all orders $0 < \alpha \leq 1$,

$$d_t\left(\boldsymbol{Q}, \boldsymbol{R}\right)^2 \leq \frac{\alpha \cdot d_\alpha\left(\boldsymbol{Q}, \boldsymbol{R}\right)}{2} \tag{C.38}$$

where each measure of divergence is as defined in Section 2.

The inequalities in Corollary C.1 are well-known (see, for instance, Lattimore and Szepesvári, 2020, §14.3 and references therein, and van Erven and Harremoës, 2014). The total variation distance is difficult to characterise in this setting due to the need to integrate over a region of the hypersphere. The inequality relations stated in Corollary C.1 provide a means to bound the total variation distance

from above using measures that are more easily characterised. In particular, Equation C.38 is a generalisation of Pinsker's inequality, yielding the classic statement of that inequality at the limit (see Equation C.31).

—| Appendix C.1 |—

Proofs

↪ *Proof of Equation C.5.* We reiterate that the von Mises-Fisher family is an exponential family and, as such, its moments can be obtained via differentiation of the log-partition function, with the first moment equal to the first derivative, the second moment equal to the second derivative, and so forth. To facilitate our analysis of the log-partition function and its derivatives, we rewrite the probability density function of the von Mises-Fisher family in terms of a single parameter vector. We let

$$\begin{aligned} \kappa &\doteq \|\eta\|_2 \\ \mu &\doteq \eta / \|\eta\|_2 \end{aligned} \tag{C.39}$$

where $\eta \in \mathbb{R}^p$. The log-partition function is equal to

$$\ln\left(\text{Ж}\text{K}\left(\kappa, m\right)\right) = \ln\left(\int_{\mathbb{S}^{p-1}} \exp\left(\kappa \cdot \mu^\top \boldsymbol{v}\right) \cdot \mathrm{d}\boldsymbol{v}\right) = -m \cdot \ln\left(\kappa\right) + \left[m+1\right] \cdot \ln\left(2\pi\right) + \ln\left(\mathrm{I}_m\left(\kappa\right)\right) \tag{C.40}$$

The moments of the von Mises-Fisher family of distributions are obtained by recursively differentiating this function with respect to the new parameter vector.

We now present some derivatives that are useful for the construction of the first, second and higher-order moments. First,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\eta}\kappa &= \mu \\ \frac{\mathrm{d}}{\mathrm{d}\eta^\top}\mu &= \frac{1}{\kappa} \cdot \left[\mathbf{I}_p - \mu\mu^\top\right] \end{aligned} \tag{C.41}$$

Second, for all $n \in \mathbb{N}_*$,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\eta^\top}\frac{\mathrm{I}_{m+n}\left(\kappa\right)}{\mathrm{I}_m\left(\kappa\right)} &= \frac{\mathrm{I}_m \cdot \left(\kappa\right) \cdot \mathrm{I}'_{m+n}\left(\kappa\right) - \mathrm{I}_{m+n}\left(\kappa\right) \cdot \mathrm{I}'_m\left(\kappa\right)}{\mathrm{I}_m\left(\kappa\right)^2} \cdot \mu^\top \\ &= \left[\frac{\mathrm{I}_{m+n+1}\left(\kappa\right)}{\mathrm{I}_m\left(\kappa\right)} + \frac{n \cdot \mathrm{I}_{m+n}\left(\kappa\right)}{\kappa \cdot \mathrm{I}_m\left(\kappa\right)} - \frac{\mathrm{I}_{m+n}\left(\kappa\right)}{\mathrm{I}_m\left(\kappa\right)} \cdot \frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_m\left(\kappa\right)}\right] \cdot \mu^\top \\ &= \left[\prod_{j=0}^{n-1} \frac{\mathrm{I}_{m+j+1}\left(\kappa\right)}{\mathrm{I}_{m+j}\left(\kappa\right)}\right] \cdot \left[\frac{\mathrm{I}_{m+n+1}\left(\kappa\right)}{\mathrm{I}_{m+n}\left(\kappa\right)} + \frac{n}{\kappa} - \frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_m\left(\kappa\right)}\right] \cdot \mu^\top \end{aligned} \tag{C.42}$$

which exploits Equation B.4 and the telescoping property of the ratios. We now differentiate Equa-

tion C.40 (i.e., we take the first derivative of the log-partition function), which yields

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\ln\left(\text{ƛK}\left(\kappa,m\right)\right)=\frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)}\cdot\mu \tag{C.43}$$

where we use Equation C.41 and Equation B.4. This proves the first result of the corollary. We then differentiate Equation C.43 (i.e., we take the second derivative of the log-partition function), which yields

$$\frac{\mathrm{d}}{\mathrm{d}\eta^{\top}}\frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)}\cdot\mu=\frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)}\cdot\left[\frac{1}{\kappa}\cdot\mathbf{I}_{p}+\left[\frac{\mathrm{I}_{m+2}\left(\kappa\right)}{\mathrm{I}_{m+1}\left(\kappa\right)}-\frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)}\right]\cdot\mu\mu^{\top}\right] \tag{C.44}$$

where we use Equations B.4, C.41 and C.42. Substituting

$$\frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)}\cdot\frac{\mathrm{I}_{m+2}\left(\kappa\right)}{\mathrm{I}_{m+1}\left(\kappa\right)}=1-\frac{p}{\kappa}\cdot\frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)} \tag{C.45}$$

which is valid by Equation B.4, we obtain an alternative expression for the variance. This proves the second result of the corollary. The higher-order moments of the von Mises-Fisher family can be obtained via recursive differentiation of these expressions, with application of the product rule of differentiation and the derivatives that are presented in Equations C.41 and C.42 then sufficient to construct these moments. ■

To faciliate characterisation of the limiting behaviour of the various measures of divergence that we consider, we apply Equation B.6.

Corollary C.2. *Given that* $\kappa>1$*, then for* $\underline{\text{Ɒ}}\left(\kappa,0,m\right)$ *and* $\overline{\text{Ɒ}}\left(\kappa,0,m\right)$ *defined as in Equations B.9 and B.10, with* $m\geq0$

$$\underline{\text{Ɒ}}\left(\kappa,0,m\right)\geq\kappa-\frac{1}{2}\cdot\ln\left(\kappa\right)-m\cdot\ln\left(2\right)-\ln\left(\Gamma\left(m+1\right)\right)-\overline{a}_{m} \tag{C.46}$$

and

$$\overline{\text{Ɒ}}\left(\kappa,0,m\right)\leq\kappa-\frac{1}{2}\cdot\ln\left(\kappa\right)-m\cdot\ln\left(2\right)-\ln\left(\Gamma\left(m+1\right)\right)+\underline{a}_{m}\cdot\ln\left(2\underline{a}_{m}\right) \tag{C.47}$$

such that $\ln\left(\mathrm{I}_{m}\left(\kappa\right)\right)$ *is an* $\mathrm{O}\left(\kappa-\ln\left(\kappa\right)/2\right)$ *function.*

Corollary C.2 follows from Equations B.9 and B.10 and the simple exploitation of the properties of increasing concave functions. Moreover, that we choose to state that the logarithm of the modified Bessel function is an $\mathrm{O}\left(\kappa-\ln\left(\kappa\right)/2\right)$ function rather than a linear function is for practical reasons. Specifically, we rely on Corollary C.2 to prove many of the statements in Section 1, for which we find

that the linear component typically cancels. We observe that Corollary C.2 aligns with results elsewhere NIST (2021, §10.30).

↪ *Proof of Corollary C.2.* We rewrite Equations B.9 and B.10 using the formula for the difference of two squares as

$$\underline{\mathfrak{D}}(\kappa,0,m) = \frac{1}{2}\cdot\ln\left(\frac{2}{\kappa}\right) - \ln\left(\Gamma\left(m+1\right)\right) + \underline{a}_m\cdot\ln\left(\frac{\kappa\cdot\left[\underline{a}_m+\overline{a}_m\right]/2}{\underline{a}_m+\sqrt{\kappa^2+\overline{a}_m^2}}\right) + \sqrt{\kappa^2+\overline{a}_m^2} - \overline{a}_m \tag{C.48}$$

$$\geq \frac{1}{2}\cdot\ln\left(\frac{2}{\kappa}\right) - \ln\left(\Gamma\left(m+1\right)\right) + \underline{a}_m\cdot\ln\left(\frac{\kappa\cdot\left[\underline{a}_m+\overline{a}_m\right]/2}{\underline{a}_m+\kappa+\overline{a}_m}\right) + \kappa - \overline{a}_m \tag{C.49}$$

$$= \frac{1}{2}\cdot\ln\left(\frac{2}{\kappa}\right) - \ln\left(\Gamma\left(m+1\right)\right) + \underline{a}_m\cdot\ln\left(\kappa\right) + \underline{a}_m\cdot\ln\left(\frac{\left[\underline{a}_m+\overline{a}_m\right]/2}{\underline{a}_m+\overline{a}_m+\kappa}\right) + \kappa - \overline{a}_m \tag{C.50}$$

$$\geq \frac{1}{2}\cdot\ln\left(\frac{2}{\kappa}\right) - \ln\left(\Gamma\left(m+1\right)\right) + \left(\underline{a}_m-\underline{a}_m\right)\cdot\ln\left(\kappa\right) + \underline{a}_m\cdot\ln\left(\frac{\underline{a}_m+\overline{a}_m}{2\left[\underline{a}_m+\overline{a}_m\right]}\right) + \kappa - \overline{a}_m \tag{C.51}$$

$$= \kappa - \frac{1}{2}\cdot\ln\left(\kappa\right) - m\cdot\ln\left(2\right) - \ln\left(\Gamma\left(m+1\right)\right) - \overline{a}_m \tag{C.52}$$

and

$$\overline{\mathfrak{D}}(\kappa,0,m) = \frac{1}{2}\cdot\ln\left(\frac{2}{\kappa}\right) - \ln\left(\Gamma\left(m+1\right)\right) + \underline{a}_m\cdot\ln\left(\frac{\kappa\cdot\left[\underline{a}_m+\underline{a}_m\right]/2}{\underline{a}_m+\sqrt{\kappa^2+\underline{a}_m^2}}\right) + \sqrt{\kappa^2+\underline{a}_m^2} - \underline{a}_m \tag{C.53}$$

$$\leq \frac{1}{2}\cdot\ln\left(\frac{2}{\kappa}\right) - \ln\left(\Gamma\left(m+1\right)\right) + \underline{a}_m\cdot\ln\left(\frac{2\underline{a}_m\cdot\kappa}{2\kappa}\right) + \kappa + \underline{a}_m - \underline{a}_m \tag{C.54}$$

$$= \kappa - \frac{1}{2}\cdot\ln\left(\kappa\right) - m\cdot\ln\left(2\right) - \ln\left(\Gamma\left(m+1\right)\right) + \underline{a}_m\cdot\ln\left(2\underline{a}_m\right) \tag{C.55}$$

respectively. Here, we exploit the concavity of the logarithmic and square root functions, and set some terms equal to zero where it is useful to do so. This establishes the veracity of the first part of Corollary C.2. We now need to show that the logarithm of the modified Bessel function, which we know lies between these two bounds for all $\kappa > 0$, is an $\mathrm{O}\left(\kappa - \ln\left(\kappa\right)/2\right)$ function. To do so, we show that the maximum of the absolute value of Equations C.52 and C.55 is bounded from above by a function that has the required order. Specifically, for all $\kappa > 0$,

$$\left|\ln\left(\mathrm{I}_m\left(\kappa\right)\right)\right| \leq \kappa - \frac{1}{2}\cdot\ln\left(\kappa\right) + m\cdot\ln\left(2\right) + \ln\left(\Gamma\left(m+1\right)\right) + \max\left(\overline{a}_m, \underline{a}_m\cdot\ln\left(2\underline{a}_m\right)\right) \tag{C.56}$$

$$\leq \left[\kappa - \frac{1}{2}\cdot\ln\left(\kappa\right)\right]\cdot\left[1 + \frac{m\cdot\ln\left(2\right) + \ln\left(\Gamma\left(m+1\right)\right) + \max\left(\overline{a}_m, \underline{a}_m\cdot\ln\left(2\underline{a}_m\right)\right)}{\left(1+\ln\left(2\right)\right)/2}\right] \tag{C.57}$$

where the denominator in the second term of Equation C.57 is the minimum value that the first term of Equation C.57 can attain, which establishes the result. ■

Each of the various measures of divergence that we consider involves integration of the kernel for some

value of the concentration parameter. The implication of this property is that each measure can be expressed in terms of logarithmic ratios of the form

$$\ln\left(\frac{\text{Ж}\left(\kappa_d, m\right)}{\text{Ж}\left(\kappa, m\right)}\right) = \begin{cases} \ln\left(\mathrm{I}_m\left(\kappa_d\right)\right) - \ln\left(\mathrm{I}_m\left(\kappa\right)\right) - m \cdot \left[\ln\left(\kappa_d\right) - \ln\left(\kappa\right)\right] & \text{if } \kappa > 0 \\ \ln\left(\mathrm{I}_m\left(\kappa_d\right)\right) + \ln\left(\Gamma\left(m+1\right)\right) - m \cdot \left[\ln\left(\kappa_d\right) - \ln\left(2\right)\right] & \text{if } \kappa = 0 \end{cases} \tag{C.58}$$

where $\kappa_d > 0$ and $\kappa \geq 0$ are placeholders for either the obtained concentration, the reference concentration, or their weighted average.

↪ *Proof of Proposition C.1.* We focus on the integrand in the definition of the Rényi divergence, and note that

$$f\left(\boldsymbol{v}; \kappa_q, \mu_q\right)^{\alpha} \cdot f\left(\boldsymbol{v}; \kappa_r, \mu_r\right)^{1-\alpha} = \frac{\exp\left(\left(\alpha \cdot \kappa_q \cdot \mu_q + \left[1-\alpha\right] \cdot \kappa_r \cdot \mu_r\right)^{\top} \boldsymbol{v}\right)}{\text{Ж}\left(\kappa_q, m\right)^{\alpha} \cdot \text{Ж}\left(\kappa_r, m\right)^{1-\alpha}} \tag{C.59}$$

We then define

$$\begin{aligned} \kappa_\alpha &\doteq \left\|\alpha \cdot \kappa_q \cdot \mu_q + \left[1-\alpha\right] \cdot \kappa_r \cdot \mu_r\right\|_2 \\ \mu_\alpha &\doteq \left[\alpha \cdot \kappa_q \cdot \mu_q + \left[1-\alpha\right] \cdot \kappa_r \cdot \mu_r\right] / \left\|\alpha \cdot \kappa_q \cdot \mu_q + \left[1-\alpha\right] \cdot \kappa_r \cdot \mu_r\right\|_2 \end{aligned} \tag{C.60}$$

such that the resulting expression respects the conventions that we have adopted for the concentration and mean direction—i.e., that the concentration is non-negative and that the mean direction is a unit vector. We substitute Equation C.60 into Equation C.59 and integrate over the hypersphere, finding that

$$\frac{1}{\alpha-1} \cdot \ln\left(\int_{\mathbb{S}^{p-1}} \frac{\exp\left(\kappa_\alpha \cdot \mu_\alpha^{\top} \boldsymbol{v}\right)}{\text{Ж}\left(\kappa_q, m\right)^{\alpha} \cdot \text{Ж}\left(\kappa_r, m\right)^{1-\alpha}} \cdot \mathrm{d}\boldsymbol{v}\right) = \frac{\alpha}{\alpha-1} \cdot \ln\left(\frac{\text{Ж}\left(\kappa_\alpha, m\right)}{\text{Ж}\left(\kappa_q, m\right)}\right) - \ln\left(\frac{\text{Ж}\left(\kappa_\alpha, m\right)}{\text{Ж}\left(\kappa_r, m\right)}\right) \tag{C.61}$$

Hence, applying Equation C.58 to Equation C.61 implies that

$$d_\alpha\left(\boldsymbol{Q}, \boldsymbol{R}\right) = \frac{\alpha}{\alpha-1} \cdot \ln\left(\frac{\kappa_q^m \cdot \mathrm{I}_m\left(\kappa_\alpha\right)}{\kappa_\alpha^m \cdot \mathrm{I}_m\left(\kappa_q\right)}\right) - \ln\left(\frac{\kappa_r^m \cdot \mathrm{I}_m\left(\kappa_\alpha\right)}{\kappa_\alpha^m \cdot \mathrm{I}_m\left(\kappa_r\right)}\right) \tag{C.62}$$

which yields Equation C.10 upon rearrangement.

In the special case where $\kappa_\alpha = 0$, the numerator on the right-hand side of Equation C.59 is one, with the corresponding integral over the hypersphere equal to the surface area of the unit ball. The formula

for the surface area of the unit ball is well-known. Hence, Equation C.61 reduces to

$$d_{\alpha}\left(\boldsymbol{Q},\boldsymbol{R}\right)=\frac{m}{\alpha-1}\cdot\ln\left(\frac{\kappa_{q}^{\alpha}\cdot\kappa_{r}^{1-\alpha}}{2}\right)+\frac{\alpha}{\alpha-1}\cdot\ln\left(\frac{1/\Gamma\left(m+1\right)}{\mathrm{I}_{m}\left(\kappa_{q}\right)}\right)-\ln\left(\frac{1/\Gamma\left(m+1\right)}{\mathrm{I}_{m}\left(\kappa_{r}\right)}\right) \tag{C.63}$$

We reiterate that this special case occurs only when Equation C.13 holds, which facilitates the restatement of the Renyí divergence in terms of one of the two concentration parameters. We choose to state the Renyí divergence in terms of both concentration parameters because the expression that we obtain is relatively simple.

In the special case where the reference distribution is the uniform distribution on the hypersphere, Equation C.61 reduces to

$$d_{\alpha}\left(\boldsymbol{Q},\boldsymbol{R}\right)=\frac{\alpha}{\alpha-1}\cdot\ln\left(\frac{\kappa_{q}^{m}\cdot\mathrm{I}_{m}\left(\kappa_{\alpha}\right)}{\kappa_{\alpha}^{m}\cdot\mathrm{I}_{m}\left(\kappa_{q}\right)}\right)-\ln\left(\frac{2^{m}\cdot\mathrm{I}_{m}\left(\kappa_{\alpha}\right)}{\kappa_{\alpha}^{m}\cdot1/\Gamma\left(m+1\right)}\right) \tag{C.64}$$

We then recall that $\kappa_{\alpha}=\alpha\cdot\kappa_{q}$ whenever $\kappa_{r}=0$, such that

$$d_{\alpha}\left(\boldsymbol{Q},\boldsymbol{R}\right)=\frac{\alpha}{\alpha-1}\cdot\ln\left(\frac{\mathrm{I}_{m}\left(\alpha\cdot\kappa_{q}\right)}{\mathrm{I}_{m}\left(\kappa_{q}\right)}\right)-\ln\left(\frac{2^{m}\cdot\mathrm{I}_{m}\left(\alpha\cdot\kappa_{q}\right)}{\kappa_{q}^{m}\cdot1/\Gamma\left(m+1\right)}\right)-\frac{m}{\alpha-1}\cdot\ln\left(\alpha\right) \tag{C.65}$$

from which we obtain Equation C.14 upon rearrangement.

We now turn our attention to the limiting behaviour of the Rényi divergence with respect to an increase in κ_{q} given that $\kappa_{q}>\kappa_{r}$ and κ_{q} is sufficiently large such that $\min\left(\kappa_{\alpha},\kappa_{q}\right)>1$, holding $\kappa_{r}>0$ fixed. We recall that

$$\underline{\mathfrak{Tb}}\left(\kappa_{q},\kappa_{r},m\right)\leq\ln\left(\mathrm{I}_{m}\left(\kappa_{q}\right)\right)\leq\overline{\mathfrak{Tb}}\left(\kappa_{q},\kappa_{r},m\right) \tag{C.66}$$

where

$$\underline{\mathfrak{Tb}}\left(\kappa_{q},\kappa_{r},m\right)\doteq\ln\left(\mathrm{I}_{m}\left(\kappa_{r}\right)\right)+m\cdot\ln\left(\frac{\kappa_{q}}{\kappa_{r}}\right)+\underline{\mathfrak{Jb}}\left(\kappa_{q},\kappa_{r},m\right) \tag{C.67}$$

and

$$\overline{\mathfrak{Tb}}\left(\kappa_{q},\kappa_{r},m\right)\doteq\ln\left(\mathrm{I}_{m}\left(\kappa_{r}\right)\right)+m\cdot\ln\left(\frac{\kappa_{q}}{\kappa_{r}}\right)+\overline{\mathfrak{Jb}}\left(\kappa_{q},\kappa_{r},m\right) \tag{C.68}$$

where

$$\underline{\text{Љ}}\left(\kappa_q, \kappa_r, m\right) \doteq \underline{a}_m \cdot \ln\left(\frac{\underline{a}_m + \sqrt{\kappa_r^2 + \overline{a}_m^2}}{\underline{a}_m + \sqrt{\kappa_q^2 + \overline{a}_m^2}}\right) + \sqrt{\kappa_q^2 + \overline{a}_m^2} - \sqrt{\kappa_r^2 + \overline{a}_m^2} \tag{C.69}$$

and

$$\overline{\text{Љ}}\left(\kappa_q, \kappa_r, m\right) \doteq \underline{a}_m \cdot \ln\left(\frac{\underline{a}_m + \sqrt{\kappa_r^2 + \underline{a}_m^2}}{\underline{a}_m + \sqrt{\kappa_q^2 + \underline{a}_m^2}}\right) + \sqrt{\kappa_q^2 + \underline{a}_m^2} - \sqrt{\kappa_r^2 + \underline{a}_m^2} \tag{C.70}$$

given $\underline{a}_m \doteq m + 1/2$ and $\overline{a}_m \doteq m + 3/2$. We further recall that the Rényi divergence is defined as

$$d_\alpha\left(\boldsymbol{Q}, \boldsymbol{R}\right) = \frac{m}{\alpha - 1} \cdot \ln\left(\frac{\kappa_q^\alpha \cdot \kappa_r^{1-\alpha}}{\kappa_\alpha}\right) + \frac{\alpha}{\alpha - 1} \cdot \ln\left(\frac{\mathrm{I}_m\left(\kappa_\alpha\right)}{\mathrm{I}_m\left(\kappa_q\right)}\right) - \ln\left(\frac{\mathrm{I}_m\left(\kappa_\alpha\right)}{\mathrm{I}_m\left(\kappa_r\right)}\right) \tag{C.71}$$

and so satisfies

$$\frac{1}{\alpha - 1} \cdot \text{Ⓐ} + \frac{\alpha}{\alpha - 1} \cdot \text{Ⓒ} + \frac{1 - \alpha}{\alpha - 1} \cdot \text{Ⓓ} \leq d_\alpha\left(\boldsymbol{Q}, \boldsymbol{R}\right) \leq \frac{1}{\alpha - 1} \cdot \text{Ⓑ} - \frac{\alpha}{\alpha - 1} \cdot \text{Ⓒ} - \frac{1 - \alpha}{\alpha - 1} \cdot \text{Ⓓ} \tag{C.72}$$

where

$$\begin{aligned}
\text{Ⓐ} &\doteq \alpha \cdot \underline{\text{Љ}}\left(\kappa_\alpha, \kappa_q, m\right) + \left[1 - \alpha\right] \cdot \underline{\text{Љ}}\left(\kappa_\alpha, \kappa_r, m\right) \\
\text{Ⓑ} &\doteq \alpha \cdot \overline{\text{Љ}}\left(\kappa_\alpha, \kappa_q, m\right) + \left[1 - \alpha\right] \cdot \overline{\text{Љ}}\left(\kappa_\alpha, \kappa_r, m\right) \\
\text{Ⓒ} &\doteq 1\left(\text{sign}\left(\alpha - 1\right) \neq \text{sign}\left(\kappa_\alpha - \kappa_q\right)\right) \cdot \left(\overline{\text{Љ}} - \underline{\text{Љ}}\right) \circ \left(\kappa_\alpha, \kappa_q, m\right) \\
\text{Ⓓ} &\doteq 1\left(\kappa_\alpha > \kappa_r\right) \cdot \left(\overline{\text{Љ}} - \underline{\text{Љ}}\right) \circ \left(\kappa_\alpha, \kappa_r, m\right)
\end{aligned} \tag{C.73}$$

as per Equation C.66. It immediately follows (by considering each possible ordering of κ and the magnitude of α relative to one) that

$$\left|d_\alpha\left(\boldsymbol{Q}, \boldsymbol{R}\right)\right| \leq \frac{1}{\left|\alpha - 1\right|} \cdot \left[\underline{a}_m \cdot \left|\ln\left(\frac{\kappa_q^\alpha \cdot \kappa_r^{1-\alpha}}{\kappa_\alpha}\right)\right| + \left|\kappa_\alpha - \alpha \cdot \kappa_q - \left[1 - \alpha\right] \cdot \kappa_r\right| + \overline{a}_m\right] \tag{C.74}$$

which then reduces to

$$\left|d_\alpha\left(\boldsymbol{Q}, \boldsymbol{R}\right)\right| \leq \frac{1}{\left|\alpha - 1\right|} \cdot \left[\underline{a}_m \cdot \left|\ln\left(\frac{\kappa_q^\alpha \cdot \kappa_r^{1-\alpha}}{\kappa_\alpha}\right)\right| + \overline{a}_m\right] + 2\kappa_r \tag{C.75}$$

by virtue of the fact that

$$\alpha \cdot \kappa_q - \left|1 - \alpha\right| \cdot \kappa_r \leq \kappa_\alpha \leq \alpha \cdot \kappa_q + \left|1 - \alpha\right| \cdot \kappa_r \tag{C.76}$$

as per its definition. Moreover, given that

$$\ln\left(\frac{\kappa_q^{\alpha}\cdot\kappa_r^{1-\alpha}}{\kappa_\alpha}\right)=-\ln\left(\left\|\alpha\cdot\left(\frac{\kappa_q}{\kappa_r}\right)^{1-\alpha}\cdot\mu_q+[1-\alpha]\cdot\left(\frac{\kappa_r}{\kappa_q}\right)^{\alpha}\cdot\mu_r\right\|_2\right) \tag{C.77}$$

it is then self-evident that the limiting behaviour of the Rényi divergence is then wholly determined by the first term on the right-hand side of Equation C.75, which is $\mathrm{O}\left(\ln\left(\kappa_q\right)\right)$ function—and so too the Rényi divergence.[12]

It is straightforward to determine that this result extends to the special case where the reference distribution is the uniform distribution on the hypersphere, since the final two logarithmic terms in Equation C.14 can easily be shown to be of the same order as the first logarithmic term using Corollary C.2. ■

↪ *Proof of Proposition C.2.* Proposition C.2 is implied by Proposition C.1 and Equation C.17. ■

↪ *Proof of Proposition C.3.* Proposition C.3 is implied by Proposition C.1 and Equation C.24. ■

↪ *Proof of Proposition C.4.* We focus on the integrand in the definition of the Kullback-Leibler divergence and note that

$$\ln\left(\frac{f\left(\boldsymbol{v};\kappa_q,\mu_q\right)}{f\left(\boldsymbol{v};\kappa_r,\mu_r\right)}\right)\cdot f\left(\boldsymbol{v};\kappa_q,\mu_q\right)=\left(\left(\kappa_q\cdot\mu_q-\kappa_r\cdot\mu_r\right)^{\top}\boldsymbol{v}+\ln\left(\frac{\text{Ж}\left(\kappa_r,m\right)}{\text{Ж}\left(\kappa_q,m\right)}\right)\right)\cdot f\left(\boldsymbol{v};\kappa_q,\mu_q\right) \tag{C.78}$$

We integrate each term on the right-hand side of Equation C.78, finding that

$$\int_{\mathbb{S}^{p-1}}\left(\kappa_q\cdot\mu_q-\kappa_r\cdot\mu_r\right)^{\top}\boldsymbol{v}\cdot f\left(\boldsymbol{v};\kappa_q,\mu_q\right)\cdot\mathrm{d}\boldsymbol{v}=\left(\kappa_q\cdot\mu_q-\kappa_r\cdot\mu_r\right)^{\top}\mathrm{E}\left(\boldsymbol{Q}\right) \tag{C.79}$$

and

$$\int_{\mathbb{S}^{p-1}}\ln\left(\frac{\text{Ж}\left(\kappa_r,m\right)}{\text{Ж}\left(\kappa_q,m\right)}\right)\cdot f\left(\boldsymbol{v};\kappa_q,\mu_q\right)\cdot\mathrm{d}\boldsymbol{v}=\ln\left(\frac{\kappa_q^m\cdot\mathrm{I}_m\left(\kappa_r\right)}{\kappa_r^m\cdot\mathrm{I}_m\left(\kappa_q\right)}\right) \tag{C.80}$$

respectively, which relies on the fact that the density function integrates to one and which we obtain from Equation C.58. Adding Equations C.79 and C.80 and substituting the results of Equation C.5—specifically, the result pertaining to the first moment—we obtain Equation C.32.

In the special case where the reference distribution is the uniform distribution on the hypersphere,

12 Equation C.77 can be negative and decreasing for some values of κ_q, with the overall value of the function then resembling a check function. Over the range of values for which $\kappa_\alpha > 1$, however, this function is purely increasing in κ_q. From Equation C.76, it is clear that $\kappa_\alpha > 1$ is satisfied if $\alpha\cdot\kappa_q > 1+|1-\alpha|\cdot\kappa_r$, regardless of the respective mean directions of the obtained and reference distributions.

Equation C.79 simplifies to

$$\int_{\mathbb{S}^{p-1}} \kappa_q \cdot \mu_q^\top \boldsymbol{v} \cdot f\left(\boldsymbol{v}; \kappa_q, \mu_q\right) \cdot \mathrm{d}\boldsymbol{v} = \kappa_q \cdot \mu_q^\top \mathrm{E}\left(\boldsymbol{Q}\right) \tag{C.81}$$

and Equation C.80 simplifies to

$$\int_{\mathbb{S}^{p-1}} \ln\left(\frac{2\pi^{m+1}}{\text{Ж}\left(\kappa_q, m\right) \cdot \Gamma\left(m+1\right)}\right) \cdot f\left(\boldsymbol{v}; \kappa_q, \mu_q\right) \cdot \mathrm{d}\boldsymbol{v} = \ln\left(\frac{\left(\kappa_q/2\right)^m}{\mathrm{I}_m\left(\kappa_q\right) \cdot \Gamma\left(m+1\right)}\right) \tag{C.82}$$

which we obtain from Equation C.58. Adding Equations C.81 and C.82 and substituting the results of Equation C.5—specifically, the result pertaining to the first moment—we obtain Equation C.33. In particular, we rely on the fact that the inner product of a single mean direction is one.

We now turn our attention to the limiting behaviour of the Kullback-Leibler divergence with respect to an increase in κ_q given that $\kappa_q \geq \min\left(1, \kappa_r\right)$, holding $\kappa_r > 0$ fixed. Given Equation C.31 and Proposition C.1, it is perhaps unsurprising that we find that the Kullback-Leibler divergence is also an $\mathrm{O}\left(\ln\left(\kappa_q\right)\right)$ function. To show this, however, requires a slight modification of our approach as compared to the proof of Proposition C.1. Whilst the main idea remains the same (replace the logarithms of modified Bessel functions with their lower or upper bounds and characterise the limiting behaviour), the Kullback-Leibler divergence is also a function of ratios of modified Bessel functions. We must take this into account when deriving bounds. We rely extensively on Corollary C.2, and recall that the Kullback-Leibler divergence is defined as

$$d_\ell\left(\boldsymbol{Q}, \boldsymbol{R}\right) = m \cdot \ln\left(\frac{\kappa_q}{\kappa_r}\right) - \ln\left(\frac{\mathrm{I}_m\left(\kappa_q\right)}{\mathrm{I}_m\left(\kappa_r\right)}\right) + \frac{\mathrm{I}_{m+1}\left(\kappa_q\right)}{\mathrm{I}_m\left(\kappa_q\right)} \cdot \left(\kappa_q - \kappa_r \cdot \mu_r^\top \mu_q\right) \tag{C.83}$$

We bound the Kullback-Leibler divergence from below, as

$$d_\ell\left(\boldsymbol{Q}, \boldsymbol{R}\right) \geq m \cdot \ln\left(\frac{\kappa_q}{\kappa_r}\right) - \overline{\text{Ъ}}\left(\kappa_q, \kappa_r, m\right) + \ln\left(\mathrm{I}_m\left(\kappa_r\right)\right) + \frac{\kappa_q^2}{\underline{a}_m + \sqrt{\kappa_q^2 + \overline{a}_m^2}} - \frac{\kappa_q \cdot \kappa_r}{\underline{a}_m + \sqrt{\kappa_q^2 + \overline{a}_m^2}} \tag{C.84}$$

$$= \underline{a}_m \cdot \ln\left(\frac{\underline{a}_m + \sqrt{\kappa_q^2 + \underline{a}_m^2}}{\underline{a}_m + \sqrt{\kappa_r^2 + \underline{a}_m^2}}\right) - \underline{\text{Ц}}_\ell\left(\boldsymbol{\kappa}, m\right) \tag{C.85}$$

where

$$\underline{\text{Ц}}_\ell\left(\boldsymbol{\kappa}, m\right) \doteq \sqrt{\kappa_q^2 + \underline{a}_m^2} - \sqrt{\kappa_r^2 + \underline{a}_m^2} - \frac{\kappa_q^2}{\underline{a}_m + \sqrt{\kappa_q^2 + \overline{a}_m^2}} + \frac{\kappa_q \cdot \kappa_r}{\underline{a}_m + \sqrt{\kappa_q^2 + \overline{a}_m^2}} \tag{C.86}$$

$$\leq \sqrt{\kappa_q^2 + \overline{a}_m^2} - \sqrt{\kappa_r^2 + \underline{a}_m^2} - \sqrt{\kappa_q^2 + \overline{a}_m^2} + \overline{a}_m + \kappa_r \tag{C.87}$$

To move from Equation C.86 to Equation C.87, we replace $\underline{a}_m$ with $\overline{a}_m$ where it is appropriate to do so, apply the formula for the difference of two squares (the other square being zero) to the penultimate term, and set the dimensional constants in the denominator of the final term equal to zero. It then follows that

$$d_\ell\left(\boldsymbol{Q},\boldsymbol{R}\right) \geq \underline{a}_m \cdot \ln\left(\frac{\underline{a}_m + \sqrt{\kappa_q^2 + \underline{a}_m^2}}{\underline{a}_m + \sqrt{\kappa_r^2 + \underline{a}_m^2}}\right) - \kappa_r - \overline{a}_m + \sqrt{\kappa_r^2 + \underline{a}_m^2} \tag{C.88}$$

We bound the Kullback-Leibler divergence from above, as

$$d_\ell\left(\boldsymbol{Q},\boldsymbol{R}\right) \leq m \cdot \ln\left(\frac{\kappa_q}{\kappa_r}\right) - \underline{\mathrm{E}}\left(\kappa_q,\kappa_r,m\right) + \ln\left(\mathrm{I}_m\left(\kappa_r\right)\right) + \frac{\kappa_q^2}{\underline{a}_m + \sqrt{\kappa_q^2 + \underline{a}_m^2}} \tag{C.89}$$

$$= \underline{a}_m \cdot \ln\left(\frac{\underline{a}_m + \sqrt{\kappa_q^2 + \overline{a}_m^2}}{\underline{a}_m + \sqrt{\kappa_r^2 + \overline{a}_m^2}}\right) - \overline{\mathrm{H}}_\ell\left(\boldsymbol{\kappa},m\right) \tag{C.90}$$

where

$$\overline{\mathrm{H}}_\ell\left(\boldsymbol{\kappa},m\right) \doteq \sqrt{\kappa_q^2 + \overline{a}_m^2} - \sqrt{\kappa_r^2 + \overline{a}_m^2} - \frac{\kappa_q^2}{\underline{a}_m + \sqrt{\kappa_q^2 + \underline{a}_m^2}} \tag{C.91}$$

$$\geq \sqrt{\kappa_q^2 + \underline{a}_m^2} - \sqrt{\kappa_r^2 + \overline{a}_m^2} - \sqrt{\kappa_q^2 + \underline{a}_m^2} + \underline{a}_m \tag{C.92}$$

To move from Equation C.91 to Equation C.92, we replace $\overline{a}_m$ with $\underline{a}_m$ where it is appropriate to do so and apply the formula for the difference of two squares (the other square being zero) to the final term. It then follows that

$$d_\ell\left(\boldsymbol{Q},\boldsymbol{R}\right) \leq \underline{a}_m \cdot \ln\left(\frac{\underline{a}_m + \sqrt{\kappa_q^2 + \overline{a}_m^2}}{\underline{a}_m + \sqrt{\kappa_r^2 + \overline{a}_m^2}}\right) - \underline{a}_m + \sqrt{\kappa_r^2 + \overline{a}_m^2} \tag{C.93}$$

Together, Equations C.88 and C.93 imply that

$$\left|d_\ell\left(\kappa_q,\kappa_r\right)\right| \leq \underline{a}_m \cdot \ln\left(\frac{\underline{a}_m + \sqrt{\kappa_q^2 + \underline{a}_m^2}}{\underline{a}_m + \sqrt{\kappa_r^2 + \underline{a}_m^2}}\right) + \left|\sqrt{\kappa_r^2 + \underline{a}_m^2} - \overline{a}_m - \kappa_r\right| \tag{C.94}$$

and so it is trivial to show that the Kullback-Leibler divergence is an $\mathrm{O}\left(\ln\left(\kappa_q\right)\right)$ function. ■

—| Appendix C.2 |—

Hankel expansion of the circular variance

One quantity that is often of interest is the circular variance of the von Mises-Fisher family of distributions, which is defined as one minus the ratio of modified Bessel functions—i.e., one minus the mean resultant length. Kitagawa et al. (2022b) demonstrates that the circular variance is an $\mathrm{O}\left(1/\kappa\right)$ function. Here, we show that the same result can be obtained via Hankel series expansion (see Equation B.13 for a definition). We note that Hankel series expansion is appropriate when $\kappa \to \infty$, which is the limiting behaviour that we are interested in. We begin by writing the circular variance in the alternative form

$$1 - \frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)} = 1 - \frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)} = \frac{\mathrm{I}_{m}\left(\kappa\right) - \mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)} \tag{C.95}$$

to which we apply the expansion. Recalling Equation B.13 and setting $n = 2$ for simplicity,

$$1 - \frac{\mathrm{I}_{m+1}\left(\kappa\right)}{\mathrm{I}_{m}\left(\kappa\right)} = \frac{\left[\mathrm{Pochhammer}_{1}\left(m+1\right) - \mathrm{Pochhammer}_{1}\left(m\right)\right]/\kappa + \mathrm{O}\left(1/\kappa^{2}\right)}{1 - \mathrm{Pochhammer}_{1}\left(m\right)/\kappa + \mathrm{O}\left(1/\kappa^{2}\right)} \tag{C.96}$$

$$= \frac{\mathrm{Pochhammer}_{1}\left(m+1\right) - \mathrm{Pochhammer}_{1}\left(m\right) + \mathrm{O}\left(1/\kappa\right)}{\kappa - \mathrm{Pochhammer}_{1}\left(m\right) + \mathrm{O}\left(1/\kappa\right)} \tag{C.97}$$

which is, trivially, an $\mathrm{O}\left(1/\kappa\right)$ function. More accurate approximations can be attained by increasing n, but a simple approximation suffices to establish the rate of the circular variance.

—| Chapter D |—

Stochastic treatment choice with empirical welfare updating

The principal goal of programme evaluation is to inform the social planner as to which individuals within a target population should receive a given treatment. When treatment effects are heterogeneous in individuals' observable characteristics, the social planner can improve social welfare by implementing an individualised treatment assignment rule based upon these characteristics. The literature on statistical treatment choice initiated by Manski (2004b) studies how to estimate assignment rules based upon a finite sample and how to assess their welfare performance. Given an experimental or observational sample, existing approaches—including those proposed in Athey and Wager (2021), Hirano and Porter (2009), Kitagawa and Tetenov (2018b), and Manski (2004b)—yield deterministic assignment rules, which are functions mapping an individual's observable characteristics to a recommended treatment. That is, individuals who share the same observable characteristics are all assigned the same treatment. Such assignment rules are sharp and address the question of who should be treated? We adopt a broader view of the treatment choice problem by considering stochastic assignment rules that map individual observable characteristics to a probability distribution over the different treatment arms, instead addressing the question of with what probability should an individual be treated?

In static treatment choice problems with outcome distributions that exhibit the monotone likelihood ratio property, deterministic assignment rules form a class of admissible policies (Karlin and H. Rubin, 1956; Tetenov, 2012), such that restricting attention to this class is without loss (of welfare). Once we allow the class of outcome distributions to be unconstrained though, there is little theoretical justification for focusing on deterministic rules. In comparison to stochastic assignment rules, deterministic assignment rules have the following three potentially undesirable features. First, deterministic assignment rules cannot incorporate confidence or uncertainty about which treatment is best for each individual, with individuals typically assigned treatment if conventional point estimates suggest that treatment has a positive effect on average. The strength of evidence in support of this conclusion, usually presented in the form of confidence intervals or p-values, is not generally acted upon; what matters is whether empirical evidence supports a positive point estimate, and not whether it is sufficient or insufficient

to reject a non-positive effect. Such a sharp dichotomy of assignment is naturally overconfident in its prescription, and there is no theoretical basis for the incorporation of confidence intervals or p-values into frequentist-based decision-making. Stochastic assignment rules can represent such uncertainty, which arises due to the (finite sample) nature of experimental data or model misspecification, through their probability weighting of treatments. Second, stochastic assignment rules facilitate future evaluation, since implementing a stochastic assignment rule can generate a new experimental sample in which treatment is randomised conditional on individual observable characteristics. Third, unlike deterministic assignment rules for which the probability that a treatment is assigned changes discontinuously at some threshold, stochastic assignment rules feature assignment probabilities that smoothly change with respect to individual characteristics. Such a feature is desirable if a fairness criterion requiring that individuals with similar characteristics have similar probabilities of treatment (Dwork et al., 2012) is enforced.

This chapter proposes novel and general methods for obtaining stochastic individualised assignment rules based on randomised control trial data. Assuming that the social planner assigns individuals to a binary treatment, with her goal being to maximise additive (utilitarian) social welfare as in Manski (2004b), we exploit an empirical analogue of the social welfare criterion to generate individualised assignment probabilities. Specifically, we start with a prior distribution over a collection of deterministic assignment rules that we denote by $\mathfrak{G}$, each g of which partitions the space of individual observable characteristics into a group of characteristics and its complement. An individual is assigned treatment if their characteristics are such that $g(\boldsymbol{X}) = 1$, and is not assigned treatment if $g(\boldsymbol{X}) = 0$. We then update the prior distribution based upon an empirical analogue of the social welfare criterion to obtain a posterior distribution over $\mathfrak{G}$. To generate a stochastic assignment, we draw a $g \in \mathfrak{G}$ according to the posterior distribution over this collection, and implement the policy prescribed by g. In this way, the probability that an individual is treated is equal to the posterior probability that the g that is drawn is such that $g(\boldsymbol{X}) = 1$ (i.e., the individual is assigned treatment under several possible g, with their overall probability of being assigned treatment equal to the probability of these g being drawn).

One of the main contributions of this chapter is that we derive an optimal updating procedure for obtaining the posterior distribution over $\mathfrak{G}$. This procedure minimises an upper bound on welfare regret and yields an exponential tilting of the prior over $\mathfrak{G}$, where the tilting depends upon the empirical welfare criterion. This novel updating formula resembles the quasi-posterior distribution that appears in the Laplace-type estimation that is studied by Chernozhukov and Hong (2003b), but differs in that the constant factor in the exponential tilting term is determined endogenously by the Lagrange multiplier of the optimisation.

Despite our analytical characterisation of the optimal posterior distribution, computation of this distribution or sampling of g from it is not straightforward. We therefore consider a variational approximation of the optimal posterior distribution by a parametric distribution. In particular, as a specification of $\mathfrak{G}$, we consider the class of Linear Eligibility Score (LES) rules that assign treatment if $x^\top\gamma$ (the linear score) exceeds some threshold γ_0 (eligibility). Building upon the LES class, we exploit the invariance of the welfare criterion to the ratio of γ to γ_0 (scale invariance) and approximate the optimal posterior distribution using a multivariate von Mises-Fisher distribution.

For the practice of reporting and communicating individualised allocations of treatment, our approach of obtaining a posterior distribution over policies is useful for generating some quantities that existing methods yielding deterministic assignment rules cannot produce. First, the posterior probability that $g(\boldsymbol{X}) = 1$ offers a personalised probabilistic assessment that individuals with these characteristics favour treatment over no treatment. Reporting such a probability offers a novel alternative to the common practice of using the p-values of hypothesis testing to express confidence in a positive treatment effect, something which does not easily translate to a recommendation about what the social planner should do. Second, viewing the posterior over g as an inferential tool for the welfare-optimality of (deterministic) assignment policies within $\mathfrak{G}$, we can obtain a credible region for the optimal policy by, for instance, selecting its highest posterior density region. This approach to obtaining confidence sets for the optimal treatment assignment policy is an alternative to the frequentist approach that is studied in Rai (2019). Third, analogous to the practice of using a Bayesian posterior with a noninformative prior as a visual summary of the likelihood function, we can use our variationally approximated posterior over g as a visual summary of the exponentiated empirical welfare criterion function.

To demonstrate how to implement our approach and what it delivers in practice, we apply our methods to the JTPA Study sample that is studied by Bloom et al. (1997). Given observations of prior earnings and years of education, we ask with what probability should an individual be treated? Restricting attention to linear assignment rules and a variational approximation of the optimal posterior distribution by a multivariate von Mises-Fisher distribution, we estimate a stochastic assignment rule that is more likely to allocate JTPA assistance to individuals with high prior earnings and fewer years of education. Kitagawa and Tetenov (2018b), which similarly considers the JTPA Study sample, and estimates a deterministic assignment rule, serves as a useful benchmark for comparison. Aside from the obvious difference that we estimate a stochastic rule (i.e., every individual has a non-trivial probability to be allocated JTPA assistance under our rule), our estimated rule allocates JTPA assistance to a smaller fraction of the population than the deterministic rule of Kitagawa and Tetenov (2018b), which targets individuals with low prior earnings and few years of education for treatment. This difference reflects the shape of

the empirical welfare criterion, which can be captured by our approach but is missed by deterministic policies that are obtained as the mode of the empirical welfare criterion.

This chapter contributes to the growing literature on statistical treatment choice initiated by Manski (2004b). Exact minimax regret assignment rules are studied in Ishihara and Kitagawa (2021), Schlag (2006), Stoye (2009), Stoye (2012), Tetenov (2012), and Yata (2021). Hirano and Porter (2009) analyses asymptotically-optimal assignment rules in limit experiments, and Bhattacharya and Dupas (2012) considers capacity constrained policies. Kitagawa and Tetenov (2018b) proposes Empirical Welfare Maximisation (EWM) methods for individualised assignment, which maximise a sample analogue of the social welfare function over a constrained class of policies. Similar approaches have been studied in the literature on machine learning and personalised medicine, as in Beygelzimer and Langford (2009), Swaminathan and Joachims (2015), Zadrozny (2003), Zhang et al. (2012), and Zhao et al. (2012). Recent advances in learning individualised assignment policies include Adjaho and Christensen (2022), Athey and Wager (2021), D'Adamo (2021), Han (2022), Kido (2022), Kitagawa and Tetenov (2021), Kitagawa et al. (2021), Liu (2022), Mbakop and Tabord-Meehan (2021), Nie et al. (2021), Sakaguchi (2019), Sasaki and Ura (2020), Sun (2021), and Viviano (2021), to list but a few works. The assignment rules estimated in these works are all deterministic.

There are some earlier works that investigate the decision-theoretic justification for stochastic (fractional) assignment and the welfare performance of these rules, with Manski (2009) providing a detailed review of settings where stochastic rules are preferable. When the welfare criterion is only partially identified, minimax regret-optimal rules are stochastic given knowledge of the identified set (Manski, 2000a, 2005, 2007a,b), which remains true even after taking into account uncertainty of estimates of the bounds (Manski, 2022; Stoye, 2012; Yata, 2021). As shown by Manski and Tetenov (2007) and Manski, 2009, stochastic assignment rules can also be justified by a nonlinear welfare criterion in a point-identified setting. Kitagawa et al. (2022a) shows that, for a wide class of nonlinear welfare regret criteria, admissible assignment rules are stochastic (fractional). In particular, Kitagawa et al. (2022a) shows that the minimax squared-regret rule is stochastic, with the probability of assignment equal to the posterior probability of a positive treatment effect under the least-favorable null. Kitagawa et al. (2022a) proposes using this probability as a measure of the strength of evidence for a positive treatment effect, replacing the commonly used p-value of a hypothesis test. In contrast, this chapter obtains the probability of assignment from a posterior probability distribution over assignment rules, rather than over the treatment effect parameters, with a quasi-likelihood built upon the empirical welfare criterion. Kock et al. (2022) obtains a stochastic assignment rule in a setting where the oracle optimal rule is fractional due to nonlinearity in the social planner's chosen welfare criterion. In contrast to the

static treatment assignment problem, dynamic treatment assignment problems analysed in the multi-arm bandit literature often consider stochastic assignments that balance the exploitation versus exploration trade-off, such as the posterior probability matching algorithm of Thompson (1933) does. Thompson sampling algorithms build upon the standard Bayesian posterior distribution for treatment effects such that the allocation algorithm crucially relies on a parametric specification of the data generating process, which our approach does not require.

Chamberlain (2011) and Dehejia (2005) approach the treatment choice problem from a Bayesian perspective. In their framework, the potential outcome distributions are parametric, and it is over the parameters of these distributions that a prior is imposed. For the standard mean welfare criterion, the Bayes optimal allocation rule is deterministic. Our approach differs from these works in that we do not assume a prior distribution over the data generating process. We instead impose few restrictions on the data generating process, and form prior and posterior distributions over the parameters that index assignment rules. Our approach can be advantageous when compared to Chamberlain (2011) if the social planner is concerned about potential misspecification of the likelihood. If the likelihood is misspecified, the resulting Bayes-optimal assignment rule can be suboptimal even for large samples. In contrast, our approach is guaranteed to yield a distribution over policies that is guaranteed to concentrate on welfare-optimal policies without requiring a specification for the data generating process.

Our approach is also related to that of Bissiri et al. (2016) and Csaba and Szoke (2020), where loss function-driven (quasi-Bayes) updating rules are proposed. Rather than follow their approach by adopting exponentiated loss as a quasi-likelihood and solving the quasi-Bayesian decision problem, we obtain an optimal learning rule by minimising a high probability upper bound on welfare regret. This way of establishing optimality is similar to the structural risk minimisation approach of Vapnik (1998) and the Probably Approximately Correct (henceforth, PAC) analysis proposed by Valiant (1984), which was extended to the study of randomised predictors in McAllester (1999), and Shawe-Taylor and Williamson (1997), constituting the development of PAC-Bayes theory. For classification and regression problems, various PAC-Bayes bounds on prediction generalisation errors are obtained in Bégin et al. (2014, 2016), Catoni (2007), Derbeko et al. (2004), Germain et al. (2009), McAllester (2003), Pentina and Lampert (2015), and Seeger (2002), and can accommodate quasi-Bayesian procedures similar to ours. See Guedj (2019) for a recent review of this literature. To our knowledge, the PAC-Bayes bounds that we derive for treatment choice are new to the literature and offer a contribution of independent interest. We also note that Pellatt (2022) makes use of PAC-Bayes theory to analyse the treatment allocation problem with stochastic assignment rules under a budget (or resource) constraint.

Although the treatment choice problem is distinct from prediction problems—as is discussed in Kitagawa and Tetenov (2018b)—the EWM approach for treatment choice is closely related to the cost-sensitive binary classification problem, as first pointed out by Zadrozny (2003). The PAC-Bayes classification analysis with variational posterior approximation that is proposed by Alquier et al. (2016) is, therefore, closely related to our analysis. There are, however, important differences with Alquier et al. (2016). First, we make use of the approach proposed by Bégin et al. (2016), which allows for the construction of a variety of different bounds via a general convex function. This introduces the complication that classification is not standard (i.e., cost is homogeneous) and is instead cost-sensitive. Introducing heterogeneity in the cost of misclassification leads to a non-trivial challenge in deriving the PAC bounds. To address these complications we leverage results in Maurer (2004) for continuous loss functions over the unit interval. Second, Alquier et al. (2016) considers approximating the optimal posterior distribution by a Gaussian distribution. We exploit the scale invariance property of the welfare criterion and approximate the optimal posterior distribution by a multivariate von Mises-Fisher distribution over the hypersphere.

—| Section 1 |—

Framework

We suppose that the social planner (or the econometrician acting on her behalf) observes experimental data that comprises a probability distribution over $\{Y, T, \boldsymbol{X}\}$ that we denote by P^n and that is constructed from n independent and identically distributed draws. Here, $Y \in \mathbb{R}$ denotes response, which is some measured outcome of interest; $T \in \{0, 1\}$ denotes treatment, which is an indicator for whether an individual is treated or not (i.e., which group—the experimental group or the control group—an individual is a member of in the experimental data); and $\boldsymbol{X} \in \mathbb{R}^i$ denotes covariates, which are some measured individual characteristics. The population from which the experimental data is drawn comprises a probability distribution over $\{Y_0, Y_1, T, \boldsymbol{X}\}$ that we denote by P. Here, Y_1 and Y_0 are potential outcomes and relate to Y via the relationship

$$y = y_1 \cdot t + y_0 \cdot [1 - t] \tag{D.1}$$

We refer to P^n as the empirical distribution and to P (together with Equation D.1) as the data-generating process, and assume that $\boldsymbol{X}$ consists only of those characteristics that the social planner can use to discriminate between individuals in the target population, with budgetary, ethical or legal considerations precluding the use of other characteristics.

Throughout our analysis, we maintain several assumptions. We follow Manski (2004b) and the subse-

quent literature in supposing that the social welfare criterion is that of a utilitarian social planner who aims to maximise the average level of individual outcomes. We note that other criteria could also be implemented, such as inequality-averse social welfare and Gini social welfare.[1]

Assumption D.1 (External validity)**.** *The population to which policy is to be applied—the target population—has the same distribution over* $\{Y_1, Y_0, \boldsymbol{X}\}$ *as the marginal distribution of* $\{Y_1, Y_0, \boldsymbol{X}\}$ *that is obtained from the data generating process.*

Assumption D.2 (Unconfoundedness)**.** *The data generating process satisfies* $\{Y_1, Y_0\} \perp\!\!\!\perp T|\boldsymbol{X}$*.*

Assumptions D.1 and D.2 are satisfied, for instance, if the experimental data is extracted directly from the target population and the treatment is, conditional on the covariates, randomly assigned,[2] independently of the potential outcomes (Rosenbaum and D. B. Rubin, 1983).

Assumption D.3 (Bounded outcomes)**.** *There exists a constant* $0 < c_y < \infty$ *such that* $0 \leq Y \leq c_y$*.*

Assumption D.4 (Strict overlap)**.** *There exists a constant* $0 < \psi < 1/2$ *such that the propensity score satisfies* $\psi \leq e(x) \leq 1 - \psi$*, where* $e(x) \doteq \mathrm{E}_P(T|\boldsymbol{X})$*.*

As is discussed in Kitagawa and Tetenov (2018b) and Swaminathan and Joachims (2015), policies that maximise an empirical welfare criterion are not invariant to positive affine transformations of outcomes, which is the case for the empirical welfare criterion that we consider in this chapter. Given Assumptions D.3 and D.4, we can define

$$H \doteq \frac{\psi \cdot Y/c_y}{e(\boldsymbol{X}) \cdot T + [1 - e(\boldsymbol{X})] \cdot [1 - T]} \tag{D.2}$$

which is confined to the unit interval, and which we interpret as weights and that are motivated by an unbiased estimator of the (scaled) expected potential outcomes. We solve the planner's problem using these transformed outcomes. This transformation of outcomes does not affect the welfare ranking over assignment policies, both in the population and according to in-sample welfare criteria, yet facilitates our analysis.

Adopting an additive utilitarian perspective, the average level of social welfare attained by g is propor-

1 For example, Kasy (2016) and Kitagawa and Tetenov (2021) study a setting where the social welfare function is a weighted average of the outcomes with rank-dependent weights, which includes the Gini social welfare function as a special case.

2 Kitagawa and Tetenov (2018b) considers a setting where the marginal distribution of $\boldsymbol{X}$ differs between the population of interest and the data generating process. Adjaho and Christensen (2022) and Kido (2022) study settings that differ also in terms of the distribution of potential outcomes.

tional to

$$W(g) \doteq \mathrm{E}_P\left(Y_1 \cdot 1\left(g\left(\boldsymbol{X}\right)=1\right)+Y_0 \cdot 1\left(g\left(\boldsymbol{X}\right)=0\right)\right) \tag{D.3}$$

Given Assumptions D.1, D.2 and D.4 and that $Y = Y_0 + T \cdot (Y_1 - Y_0)$, we can re-write Equation D.3 as

$$W(g) = \mathrm{E}_P\left(\frac{Y \cdot T}{e\left(\boldsymbol{X}\right)} \cdot 1\left(g\left(\boldsymbol{X}\right)=1\right)+\frac{Y \cdot (1-T)}{1-e\left(\boldsymbol{X}\right)} \cdot 1\left(g\left(\boldsymbol{X}\right)=0\right)\right) \tag{D.4}$$

$$= \mathrm{E}_P\left(\frac{Y \cdot [1-T]}{1-e\left(\boldsymbol{X}\right)}\right)+\mathrm{E}_P\left(\left[\frac{Y \cdot T}{e\left(\boldsymbol{X}\right)}-\frac{Y \cdot [1-T]}{1-e\left(\boldsymbol{X}\right)}\right] \cdot g\left(\boldsymbol{X}\right)\right) \tag{D.5}$$

Accordingly, the sample analogue of Equation D.3 can be written as

$$W_n(g) \doteq \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i \cdot t_i}{e\left(\boldsymbol{x}_i\right)} \cdot 1\left(g((\boldsymbol{x}_i)=1\right)+\frac{y_i \cdot [1-t_i]}{1-e\left(\boldsymbol{x}_i\right)} \cdot 1\left(g\left(\boldsymbol{x}_i\right)=0\right)\right] \tag{D.6}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{y_i \cdot [1-t_i]}{1-e\left(\boldsymbol{x}_i\right)}+\frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i \cdot t_i}{e\left(\boldsymbol{x}_i\right)}-\frac{y_i \cdot [1-t_i]}{1-e\left(\boldsymbol{x}_i\right)}\right] \cdot g\left(\boldsymbol{x}_i\right) \tag{D.7}$$

where $W_n(g)$ is an unbiased estimator for the true level of welfare that arises from the implementation of a particular g. Given the additive social welfare criterion, the maximal welfare level can be attained by a deterministic policy. Hence, as far as the population welfare maximisation problem is concerned, the social planner wants to select the g that maximises $W(g)$.

Each g can be associated with the set of characteristics that it assigns treatment to. We refer to g as a deterministic asssignment rule, or simply as an assignment rule, with $\mathfrak{G}$ constituting the class of assignment rules. We can write Equation D.7 as

$$W_n(g) = \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \left[\frac{t_i}{e\left(\boldsymbol{x}_i\right)} \cdot g\left(\boldsymbol{x}_i\right)+\frac{1-t_i}{1-e\left(\boldsymbol{x}_i\right)} \cdot \left[1-g\left(\boldsymbol{x}_i\right)\right]\right] \tag{D.8}$$

$$= \frac{1}{n}\sum_{i=1}^{n} h_i \cdot \frac{c_y}{\psi} \cdot \left[t_i \cdot g\left(\boldsymbol{x}_i\right)+\left[1-t_i\right] \cdot \left[1-g\left(\boldsymbol{x}_i\right)\right]\right] \tag{D.9}$$

$$= \frac{1}{n}\sum_{i=1}^{n} h_i \cdot \frac{c_y}{\psi} \cdot 1\left(g\left(\boldsymbol{x}_i\right)=t_i\right) \tag{D.10}$$

$$= \frac{1}{n}\sum_{i=1}^{n} h_i \cdot \frac{c_y}{\psi}-\frac{1}{n}\sum_{i=1}^{n} h_i \cdot \frac{c_y}{\psi} \cdot 1\left(g\left(\boldsymbol{x}_i\right) \neq t_i\right) \tag{D.11}$$

where h_i is the realisation of H, as defined in Equation D.2, for a given individual. We observe from Equation D.11 that $W_n(g)$ depends upon g only through its second term, such that

$$\underset{g}{\operatorname{argmax}}\, W_n(g) = \underset{g}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n} h_i \cdot 1\left(g\left(\boldsymbol{x}_i\right) \neq t_i\right) \tag{D.12}$$

Accordingly, we define

$$R(g) \doteq \mathrm{E}_P\left(H \cdot 1\left(g\left(\boldsymbol{X}\right) \neq T\right)\right) \tag{D.13}$$

which we term the welfare risk of g, and its empirical analogue

$$R_S(g) \doteq \frac{1}{n}\sum_{i=1}^{n} h_i \cdot 1\left(g\left(\boldsymbol{x}_i\right) \neq t_i\right) \tag{D.14}$$

which we term the empirical welfare risk of g. In view of Equation D.12, the social planner's objective is to minimise Equation D.13 with respect to g, via Equation D.14, following the empirical risk minimisation principle of Vapnik (1998).

One special set of policies that we draw particular attention to is the LES class that is defined in Kitagawa and Tetenov (2018b), and that we denote by $\mathfrak{L}$. Assignment rules in this class are indexed by a finite-dimensional parameter vector γ and a threshold γ_0, and are associated with a binary function ℓ_β that satisfies

$$\ell_\beta\left(\boldsymbol{X}\right) \doteq 1\left(\boldsymbol{X}^\top\gamma \geq \gamma_0\right) \tag{D.15}$$

where we take β to include both γ and γ_0 (i.e., β is an m-dimensional vector). Each LES rule induces a partitioning of the covariate space into two half-spaces, such that individuals in the upper contour set receive treatment and individuals in the lower contour set do not. By restricting β to the unit hypersphere, we guarantee that each policy is associated with a unique β. In what follows, we exploit the interchangeability of β and the LES rule that it indexes, adopting β as the argument of the loss functions that we consider. For instance, and again with some abuse of notation, whenever we focus on the LES class of decision rules we write

$$R(\beta) = \mathrm{E}_P\left(H \cdot 1\left(\ell_\beta\left(\boldsymbol{X}\right) \neq T\right)\right) \tag{D.16}$$

and

$$R_S(\beta) = \frac{1}{n}\sum_{i=1}^{n} h_i \cdot 1\left(\ell_\beta\left(\boldsymbol{x}_i\right) \neq t_i\right) \tag{D.17}$$

respectively, in place of Equations D.13 and D.14.

—| Section 2 |—

Posterior distributions and stochastic assignment rules

We now adapt Equations D.13 and D.14 to handle stochastic assignment rules. We let Π denote a probability distribution over $\mathfrak{G}$ that we interpret as a posterior distribution, assuming that $\mathfrak{G}$ can be embedded in a measurable space.[3] We let $\mathfrak{P}$ denote the collection of all posterior distributions.

Definition D.1 (Posterior assignment rule)**.** *Let Π be a probability distribution over $\mathfrak{G}$ that is constructed upon observing the sample. The posterior assignment rule under Π is a stochastic assignment rule that assigns treatment to individuals with probability $Q^{\Pi}(\boldsymbol{X}) \doteq \int_{\mathfrak{G}} g(\boldsymbol{X}) \cdot \mathrm{d}\Pi$.*

To implement posterior assignment rules in practice, we randomly draw a g from $\mathfrak{G}$ according to Π for each individual in the target population. In this way, similar individuals, who can have similar assignment probabilities, can be assigned to different treatment arms. Moreover, this approach does not require computation of the probability of treatment.

Definition D.2 (Expected welfare risk under Π)**.** *We define the expected welfare risk under Π as*

$$R^{\Pi} \doteq \int_{\mathfrak{G}} R(g) \cdot \mathrm{d}\Pi(g) = \mathrm{E}_P\left(H \cdot \left[T \cdot \left[1 - Q^{\Pi}(\boldsymbol{X})\right] + \left[1 - T\right] \cdot Q^{\Pi}(\boldsymbol{X})\right]\right) \tag{D.18}$$

with its empirical analogue taking the form

$$R^{\Pi}_{S} \doteq \int_{\mathfrak{G}} R_S(g) \cdot \mathrm{d}\Pi(g) = \frac{1}{n}\sum_{i=1}^{n} h_i \cdot \left[t_i \cdot \left[1 - Q^{\Pi}(\boldsymbol{x}_i)\right] + \left[1 - t_i\right] \cdot Q^{\Pi}(\boldsymbol{x}_i)\right] \tag{D.19}$$

The interpretation of R^{Π} is the average welfare loss that the social planner expects from stochastic implementation of g in $\mathfrak{G}$ in the target population when g is distributed according to Π.

We reiterate that stochastic assignment is achieved by randomly drawing g according to Π. This way of selecting assignment rules is reminiscent of the Gibbs classifier in statistical learning theory (Germain et al., 2009) and might offer a computational advantage over other methods if drawing g according to Π is easier than maximising empirical welfare or finding the mode of Π, say. An advantage of stochastic assignment is the possibility for sequential treatment evaluation: the induced assignment of treatment and non-treatment to individuals in the target population by Π is random conditional on $\boldsymbol{X}$, which allows for estimation of the causal effect of treatment in future studies. In this sense, stochastic assignment is well suited to balancing existing evidence about what constitutes the optimal assignment for each

3 For $\mathfrak{G}$ to be embedded in a measurable space, $\mathfrak{G}$ cannot be too rich. We defer to Molchanov (2005) and Gunsilius (2019) for further discussion of this point.

individual against the benefit of further exploration of the treatment effect (Manski, 2000b). We do not, however, study this channel and focus on a purely static problem in this chapter.

Our framework allows the social planner to hold some prior as to what constitutes the best policy. We differentiate the subjective beliefs that the social planner holds, which we encode using the prior distribution Π_0, and their updated beliefs following their observation of sample data, which we encode using Π. In the analysis that follows, we provide finite sample regret guarantees in the form of PAC-Bayes bounds for stochastic assignment. The approach discussed here differs significantly from other Bayesian treatment choice settings, such as those discussed in Chamberlain (2011), because we do not impose any kind of prior belief on P. We instead choose to model the beliefs that the social planner has regarding the optimal policy. Using a decision procedure that is free from specification of the likelihood comes with a desirable robustness property, as we discuss in due course.

Aside from its more conventional role as a means of expressing existing information about what constitutes the best policy, Π_0 can also play several other roles within our framework. For instance, Π_0 can also embed any constraints that are imposed upon the set of policies through truncation of its support. Such a zero density condition can be imposed in lieu of an explicit restriction on $\mathfrak{G}$ and is easy to implement in practice via rejection sampling, with only those policies that satisfy any budgetary, ethical or legal constraints being retained under the sampling procedure. Moreover, Π_0 can also be used to describe the status quo, with restrictions on the shape of Π_0 governing how much policy can deviate. These interpretations of Π_0 naturally extend to Π.

—| Section 3 |—

Optimal stochastic assignment and convergence of welfare

Having defined a criterion by which to assess stochastic assignment rules, we ask what is the optimal posterior assignment rule? Of course, answering this question requires that we make some link between the performance of an assignment rule in the sample versus in the target population, and that we make some judgement about the feasibility of estimating this rule.

—| Subsection 3.a |—

Bounding expected welfare risk

Seeing as experimental data provides only an insight into the welfare performance of any policy in the target population, a very natural question to ask is how much we can expect R_S^{Π} to differ from R^{Π} for any given Π. We provide an answer to this question here.

Theorem D.1. *Suppose that Assumptions D.1 to D.4 are satisfied, and that $n \geq 8$. Then, with at least probability $1 - \varepsilon$ under P^n, for all ε satisfying $0 < \varepsilon < 1$,*

$$R^{\Pi} \leq R_S^{\Pi} + \sqrt{\frac{1}{2n} \cdot \left[d_\ell \left(\Pi, \Pi_0 \right) + \ln \left(\frac{2\sqrt{n}}{\varepsilon} \right) \right]} \tag{D.20}$$

provided that $\Pi \in \mathfrak{P}$ is absolutely continuous with respect to Π_0.

We present Theorem D.1 without proof (see KITAGAWA et al., 2022b for a proof of Theorem D.1).[4]

We note that Theorem D.1 holds for any rules that update Π_0 and deliver Π, and that we have not committed to any particular updating rule to obtain Equation D.20. A feature of Equation D.20 is that the regularisation term—the square root term containing the Kullback-Leibler divergence of Π from Π_0—enters additively, which we find is a convenient feature for establishing convergence of our variational approximation (regularisation can otherwise be effected multiplicatively).[5] The regularisation term, by design, prevents overfitting. To illustrate this point, suppose that Π_0 is uniform over $\mathfrak{G}$: the best response of the social planner absent regularisation is to concentrate probability mass on the optimal (in-sample) g as suggested by data, such that Π is degenerate. In the presence of regularisation, however, this is no longer a best response, since the Kullback-Leibler divergence infinitely penalises degeneracy vis-à-vis uniformity. Rather, the best response of the social planner is to allocate probability mass on all g in $\mathfrak{G}$, albeit concentrating more mass on those g that are associated with low empirical welfare risk. Put differently, the regularisation term controls how far away from Π_0 a stochastic assignment rule can be, with this difference governed by the number of observations in the sample.

The Vapnik-Chervonenkis (VC) dimension is the standard measure of complexity in the statistical learning literature. We instead associate complexity with the Kullback-Leibler divergence. If one is willing to impose distributional constraints on a posterior assignment rule, an advantage of the PAC-Bayes approach and of using the Kullback-Leibler divergence is that complexity is then purely in terms of the selected (stochastic) assignment rule Π rather than in terms of the class of possible stochastic assignment rules $\mathfrak{P}$ or the class of underlying deterministic rules $\mathfrak{G}$. As such, Theorem D.1 does not explictly require any assumption about the VC dimension of $\mathfrak{G}$. The influence of VC dimension for $\mathfrak{G}$ is implicit in our setting: the upper bound on the difference between R_S^{Π} and R^{Π} implied by Equation D.20 is governed

4 The proof builds upon BÉGIN et al. (2016, §Lemma 3), which offers a flexible approach that allows for the recovery of many different PAC-Bayes bounds. The approach centres around a convex function of R^{Π} and R_S^{Π}, which we specify so as to recover the form that is presented in MCALLESTER (2003, see §(8) in BÉGIN et al., 2016). We leverage results in MAURER (2004, §Lemma 3 and §Theorem 1), exploiting the properties of Bernoulli random variables and convex functions, to adapt BÉGIN et al. (2016) and the bound that is presented therein to allow for heterogeneous cost (in lieu of the standard binary loss function that is prevalent in the classification literature and that is studied in BÉGIN et al., 2016).

5 See BÉGIN et al. (2016) for the implications of different convex functions.

by the Kullback-Leibler divergence, which is increasing in the dimension of the support of Π and Π_0, and so is non-decreasing in the complexity of $\mathfrak{G}$.

—| Subsection 3.B |—

Optimal updating rule

In a standard Bayesian setting, unknown parameters index the distribution of data and inference on parameters is conducted with respect to the posterior distribution. Typically, the posterior distribution is constructed from a well-defined likelihood function via Bayes' theorem. We leverage Theorem D.1 to construct Π from Π_0 and R_S^{Π}. This approach is valid since Theorem D.1 holds for all $\Pi \in \mathfrak{P}$. We emphasise that the posterior distribution that we construct is over assignment rules rather than over the data generating processes, which is distinct from Bissiri et al. (2016) and Csaba and Szoke (2020).

Following McAllester (2003) and Germain et al. (2009), we define an optimal posterior distribution, which we denote by Π^*, as a distribution over $\mathfrak{G}$ that minimises the right hand side of Equation D.20. That is, Π^* minimises

$$R_S^{\Pi} + \sqrt{\frac{1}{2n} \cdot \left[d_\ell \left(\Pi, \Pi_0 \right) + \ln \left(\frac{2\sqrt{n}}{\varepsilon} \right) \right]} \tag{D.21}$$

Theorem D.2. *The optimal posterior* Π^* *over* $\mathfrak{G}$ *satisfies*

$$\mathrm{d}\Pi^* \left(g \right) \doteq \frac{\exp \left(-\chi \cdot R_S \left(g \right) \right)}{\int_{\mathfrak{G}} \exp \left(-\chi \cdot R_S \left(g \right) \right) \cdot \mathrm{d}\Pi_0 \left(g \right)} \cdot \mathrm{d}\Pi_0 \left(g \right) \tag{D.22}$$

where

$$\chi \doteq 4n \cdot \sqrt{\frac{1}{2n} \cdot \left[d_\ell \left(\Pi, \Pi_0 \right) + \ln \left(\frac{2\sqrt{n}}{\varepsilon} \right) \right]} \tag{D.23}$$

We present a proof of Theorem D.2 in the appendices.

The posterior distribution that we derive is analogous to the optimal posterior in McAllester (2003) with the difference that our observations are mapped to the unit interval rather than to positive-negative, which is the standard support in classification. This particular distribution is common in the statistical mechanics literature and is a Boltzmann (or Gibbs) distribution, and has the form of exponential tilting of the prior, where the exponential tilting term involves the negative empirical welfare risk. The degree of tilting depends upon the magnitude of $\chi > 0$, which is the inverse of the Lagrange multiplier of the

associated minimisation problem and that corresponds to the root of Equation D.23. The Lagrange multiplier controls the extent to which Π_0 is updated by empirical welfare risk in minimising the upper bound for R^{Π}.

Although Theorem D.2 offers an analytical characterisation of the optimal posterior, we are unable to obtain a closed-form expression for the optimal posterior density. Whilst Equation D.23 does suggest a means to compute this density, it is likely that, in practice, this computation is difficult to perform with any precision. Given that we have, however, established that this density exists, we can consider approximating it using the variational approximation method, as is considered in Alquier et al. (2016).

—| Subsection 3.c |—

Variational approximation of the optimal stochastic assignment rule

We develop a variational approximation of the optimal posterior density. Variational approximation is useful in situations where Gibbs distributions are difficult to sample from directly, such as is the case for graphical models where Markov Chain Monte Carlo (MCMC) sampling is costly (Wainwright and Jordan, 2008). In variational approximation, we choose to approximate the optimal posterior distribution via a family of distributions of our choice, $\mathfrak{V} \subset \mathfrak{P}$. This allows us to develop an analytically tractable upper bound for the welfare regret attained by the resulting stochastic assignment rule.

Aside from guaranteeing tractability in estimation, variational approximation can also be motivated as a convenient way to impose constraints on the set of policies. For instance, a fairness criterion requiring that individuals with similar characteristics have similar probabilities of treatment (Dwork et al., 2012) can be enforced by specifying that $\mathfrak{V}$ is a continuous family, and by limiting the concentration of the density. A similar approach can be used if $\mathfrak{V}$ is a parametric family to limit how much policy can deviate from the status quo, by fixing the parameters of the posterior distribution or restricting them to some set, say.

We approximate Π^* (implicitly defined in Theorem D.2) by minimising the right-hand side of Equation D.20 with respect to posterior distributions in $\mathfrak{V}$, defining

$$\tilde{\Pi} \doteq \underset{\Pi \in \mathfrak{V}}{\operatorname{argmin}} \left(R_S^{\Pi} + \sqrt{\frac{1}{2n} \cdot \left[d_\ell \left(\Pi, \Pi_0 \right) + \ln \left(\frac{2\sqrt{n}}{\varepsilon} \right) \right]} \right) \tag{D.24}$$

We then use this optimal variational posterior distribution to define a new bound for welfare regret from which we can characterise its convergence rate.

Lemma D.1. *With probability* $1-\varepsilon$ *under* P^n*, for all* ε *such that* $0<\varepsilon<1$*, the expected welfare risk under the optimal variational posterior satisfies*

$$R^{\tilde{\Pi}} \leq \inf_{\Pi \in \mathfrak{V}} \left(R^{\Pi} + \frac{a(\lambda, n)}{\lambda} + \frac{d_\ell(\Pi, \Pi_0)}{\lambda} + \frac{\ln(2/\varepsilon)}{\lambda} + \sqrt{\frac{1}{2n} \cdot \left[d_\ell(\Pi, \Pi_0) + \ln\left(\frac{4\sqrt{n}}{\varepsilon} \right) \right]} \right) \tag{D.25}$$

where $\lambda > 0$ *is an arbitrary constant.*

In what follows, we let $\mathfrak{V}$ be a variational family of distributions that assigns positive density to assignment rules in $\mathfrak{L}$ only. Equivalently, we let $\mathfrak{V}$ be a directional family that assigns positive density to unit vectors on the hypersphere, recalling that every policy in the LES class can be uniquely associated with a unit vector (see Equation D.15 and surrounding discussion). A particularly tractable directional family, and one that we use, is the von Mises-Fisher family of distributions, which is characterised by a probability density function satisfying, for all $\kappa > 0$ and m-dimensional unit vectors μ,

$$\mathrm{d}\Pi(\beta; \kappa, \mu) \doteq \frac{\kappa^{m/2-1} \cdot \exp(\kappa \cdot \mu^\top \beta)}{[2\pi]^{m/2} \cdot \mathrm{I}_{m/2-1}(\kappa)} \cdot \mathrm{d}\beta \tag{D.26}$$

We refer the reader to Chapter C for further details about the von Mises-Fisher family of distributions.

We now introduce some further notation that facilitates our analysis. First, we let $\overline{R} \doteq \inf_{\Pi \in \mathfrak{P}} R^{\Pi}$ denote the minimum expected welfare risk amongst assignment rules in $\mathfrak{L}$ and $\overline{\beta}$ denote the vector that parametrises the assignment rule that induces $\overline{R}$. We also add the following assumption that restricts the marginal distribution of $\boldsymbol{X}$.

Assumption D.5 (Margin assumption)**.** *There exists a constant* $\eta > 0$ *such that, for all* m*-dimensional unit vectors* $\beta^\dagger$ *and* $\beta^\ddagger$*,*

$$P\left(\operatorname{sign}(\boldsymbol{X}^\top \beta^\dagger) \neq \operatorname{sign}(\boldsymbol{X}^\top \beta^\ddagger)\right) \leq \eta \cdot \left\| \beta^\dagger - \beta^\ddagger \right\|_2 \tag{D.27}$$

Assumption D.5 is satisfied whenever $\boldsymbol{X}$ has bounded density on the unit hypersphere and is also present in the analysis of Alquier et al. (2016). An interpretation is that the proportion of individuals whose treatment status switches is continuous with respect to the LES, with

$$R(\beta) - \overline{R} \leq 2\eta \cdot \|\beta - \overline{\beta}\|_2 \tag{D.28}$$

being an implication of Assumption D.5.

We now present a high-probability uniform upper bound for the welfare regret of the stochastic assignment rule obtained by variational approximation via the von Mises-Fisher family of distributions.

Theorem D.3. *Suppose that Assumptions D.1 to D.5 are satisfied, that* Π_0 *is a uniform distribution over the unit hypersphere, that* Π *is a von Mises-Fisher distribution, and that* $n \geq 8$*. Then, with probability at least* $1-\varepsilon$ *under* P^n*,*

$$R^{\tilde{\Pi}} - \overline{R} \leq \text{universal constant} \cdot \frac{\ln(n)}{\sqrt{n}} \tag{D.29}$$

where the universal constant is a function of ε*.*

We present proof of Theorem D.3 and provide an analytical expression for the universal constant in the appendices.

The uniform upper bound on welfare regret that is defined by Equation D.29 decays at a rate of $\ln(n)/\sqrt{n}$. This rate is slightly slower than the welfare regret convergence rate of the EWM (deterministic) assignment rule studied in Kitagawa and Tetenov (2018b). A simple comparison of these rates, however, is not quite meaningful for the following reason: we do not know if the convergence rate of $\ln(n)/\sqrt{n}$ that is obtained in Theorem D.3 is sharp or not. Theorem D.3 requires Assumption D.5, whilst Kitagawa and Tetenov (2018b) does not impose this assumption in showing that $1/\sqrt{n}$ is the minimax optimal rate of welfare regret convergence.[6] We do not know what the minimax optimal rate of welfare regret convergence is when Assumption D.5 is additionally imposed and, hence, cannot rule out the possibility that the convergence rate of Theorem D.3 can be improved upon and made faster than $1/\sqrt{n}$. We leave further investigation of this matter for future research.

It is worth emphasising that the regret convergence result of Theorem D.3 imposes weak restrictions on the distribution of data (Assumptions D.3 - D.5) and does not require the specification of a likelihood function or of regression equations. This contrasts with other approaches such as a Bayesian approach, where misspecification of likelihood can lead to non-convergence of the welfare regret even when Bayes optimal policies are constrained to deterministic ones in $\mathfrak{L}$.

Our approach is similar to Alquier et al. (2016) in which the families of distributions for variational approximation are multivariate Gaussian distributions on Euclidean space with flexible covariance ma-

6 Kitagawa and Tetenov (2018b, §Theorem 2.3 and §Theorem 2.4) establishes that the minimax-optimal rate under a stronger condition than our Assumption D.5 is $1/n^{2/3}$. This stronger condition embeds a margin assumption that implies our Assumption D.5 but also embeds the requirement that the first-best treatment rule is contained in the set of admissible decision rules—that $\mathfrak{L}$ contains the deterministic assignment rule that minimises expected welfare risk amongst all assignment rules. We do not make any assumption about whether $\mathfrak{L}$ (or indeed $\mathfrak{G}$ in earlier parts of our analysis) contains the first-best assignment rule, or whether this rule is deterministic or stochastic.

trices. In our approach, we stipulate a class of von Mises-Fisher distributions. Since the empirical welfare criterion for LES rules is invariant to the scale of β, it is natural to consider von Mises-Fisher distributions rather than Gaussian ones. The scale invariance of von Mises-Fisher distributions can simplify optimisation of the variational approximation by reducing the set of optima to a singleton.

—| Section 4 |—

Implementation

To implement our procedure, we restrict attention to $\mathfrak{L}$ and let $\mathfrak{V}$ be the von Mises-Fisher family of distributions. Our goal is then to minimise the objective function,

$$R_S^{\Pi} + \sqrt{\frac{1}{2n} \cdot \left[d_\ell \left(\Pi, \Pi_0 \right) + \ln \left(\frac{2\sqrt{n}}{\varepsilon} \right) \right]} \tag{D.30}$$

with respect to κ and μ, which are the parameters of our chosen variational family (the concentration parameter and the mean direction, respectively). Here, we reiterate that ε relates to the probability with which our high probability bounds hold. We assume that Π_0 is the uniform distribution over the sphere and set ε equal to the 5% level throughout.

We propose numerically minimising the objective function, approximating R_S^{Π} using Monte Carlo draws of β from the von Mises-Fisher distribution for a given realisation of data and for fixed values of the parameters of the von Mises-Fisher distribution. We let β^j be one such pseudo-random draw that we obtain, and compute

$$\hat{R}_S^{\Pi} \doteq \frac{1}{n} \sum_{i=1}^{n} h_i \cdot \left[t_i \cdot \left[1 - \hat{\Pi}_i \right] + \left[1 - t_i \right] \cdot \hat{\Pi}_i \right] \tag{D.31}$$

where

$$\hat{\Pi}_i \doteq \frac{1}{J} \sum_{j=1}^{J} 1 \left(\boldsymbol{x}_i^\top \beta^j \geq 0 \right) \tag{D.32}$$

Fast pseudo-random sampling of von Mises-Fisher random vectors is possible using the rejection sampling method of Wood (1994) or the inversion method of Kurz and Hanebeck (2015). The analogue of Equation D.30 that we then minimise is

$$\hat{R}_S^{\Pi} + \sqrt{\frac{1}{2n} \cdot \left[\left[\frac{m}{2} - 1 \right] \cdot \ln \left(\frac{\kappa}{2} \right) - \ln \left(\mathrm{I}_{m/2-1} \left(\kappa \right) \right) - \ln \left(\Gamma \left(\frac{m}{2} \right) \right) + \frac{\mathrm{I}_{m/2} \left(\kappa \right)}{\mathrm{I}_{m/2-1} \left(\kappa \right)} \cdot \kappa + \ln \left(\frac{2\sqrt{n}}{\varepsilon} \right) \right]} \tag{D.33}$$

which follows from our assumption of a uniform prior and the form that the Kullback-Leibler divergence takes under that assumption.

As is shown in Chapter C, the Kullback-Leibler divergence of the von Mises-Fisher distribution from the uniform distribution over the hypersphere does not depend upon μ and increases at the logarithmic rate in the concentration. In contrast, although R_S^{Π} is a function of the parameters of the von Mises-Fisher distribution, the nature of this relationship is not clear. That several combinations of these parameters induce local minima of R_S^{Π} cannot be ruled out for instance. Moreover, since we construct $\hat{R}_S^{\Pi}$ based upon random draws of β from Π, our objective function is not a smooth function of the concentration nor the mean direction. Given this, we suggest that a grid-based search for the minimum of the objective function is appropriate, with this search limited to values of the concentration parameter between zero and some specified upper limit. Further insight about the behaviour of the objective function is provided in an online appendix.

—| Section 5 |—

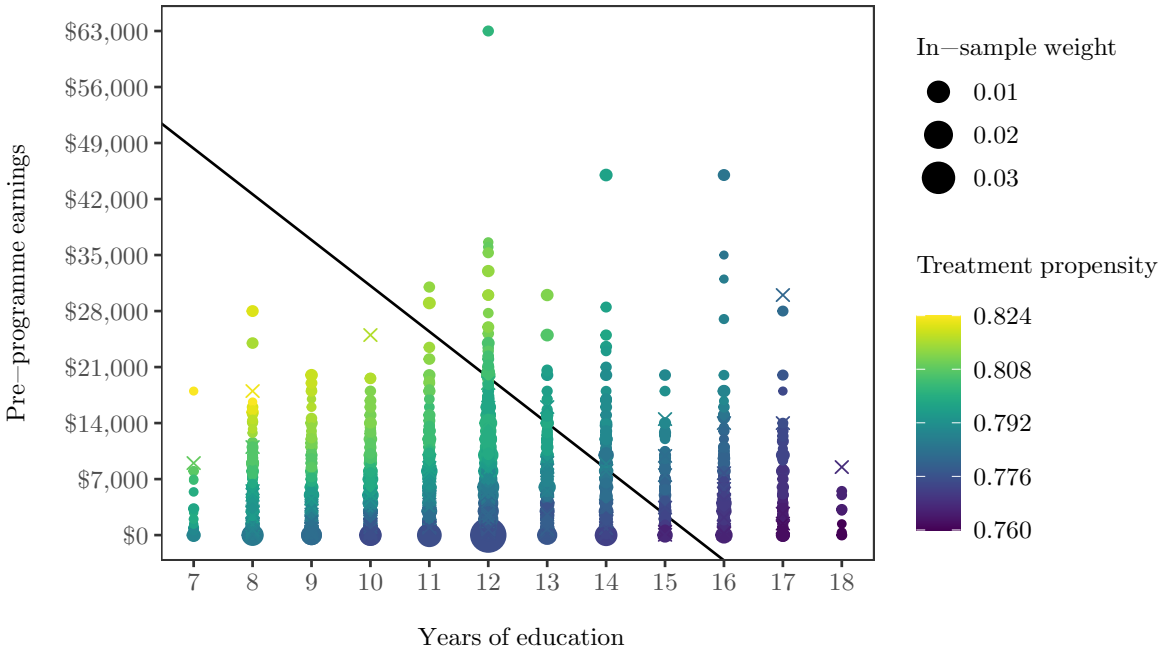
Empirical illustration

We illustrate our procedure using data from the National Job Training Partnership Act (JTPA) Study. Applicants to the Study were randomly allocated to one of two groups. Applicants allocated to the treatment group were extended training, job search assistance and other services provided by the JTPA over a period of 18 months. Applicants allocated to the control group were excluded from JTPA assistance. Along with information collected prior to the commencement of the intervention, the Study also collected administrative and survey data relating to applicants' earnings in the 30 months following its start. Further details about the data and the Study can be found elsewhere (see, for instance, Bloom et al., 1997). We restrict attention to a sample of 9,223 observations for which data on years of education and pre-programme earnings amongst the sample of adults (aged 22 years and older) used in the original evaluation of the programme and in subsequent studies (Abadie et al., 2002; Bloom et al., 1997; Heckman et al., 1997) is available. Applicants in this sample were assigned to the treatment group with a probability of two thirds. Like Kitagawa and Tetenov (2018b), we define T to be the initial assignment of treatment, rather than the actual take-up due to the presence of non-compliance in the experiment. We study stochastic assignment rules.

We follow Kitagawa and Tetenov (2018b) in considering total individual earnings in the 30 months after programme assignment as our principal welfare measure. Moreover, we focus exclusively on the class of linear rules,

$$\begin{aligned}\mathfrak{L} &= \left\{ g : g\left(\boldsymbol{X}\right) = 1\left(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 \geq 0\right) \,\middle|\, \left\|\{\beta_0, \beta_1, \beta_2\}\right\|_2 = 1 \right\} \\ &= \left\{ g : g\left(\boldsymbol{X}\right) = 1\left(\beta_0 + \beta_1 \cdot \text{prior earnings} + \beta_2 \cdot \text{education} \geq 0\right) \,\middle|\, \left\|\{\beta_0, \beta_1, \beta_2\}\right\|_2 = 1 \right\}\end{aligned} \tag{D.34}$$

FIGURE D.1
Variation in treatment propensity across individuals in the JTPA Study sample



JTPA Study sample. This figure illustrates the treatment propensity of individuals under the posterior assignment rule that is induced by $\{\kappa^*, \mu^*\}$. Each point represents the individual characteristics of an individual or several individuals in the (crosses denote individuals with zero in-sample weight). For comparison, individuals to the left of the solid diagonal line are assigned treatment under the optimal deterministic assignment rule of KITAGAWA *and* TETENOV *(2018b).*

that are studied in that paper.

To implement our procedure, we map prior earnings and education to the unit interval,[7] and calculate H as outlined in Equation D.2. We perform this calculation without adjusting post-programme earnings by the average cost of JTPA assistance ($774 per individual) for treated individuals.[8] We then utilise a grid search approach over the parameters of the von Mises-Fisher distribution, specifying a reasonably fine grid over the unit sphere and over a finite subset of the reals.[9]

For each point on our grid, we draw 1,000 values of β from the corresponding von Mises-Fisher distribution and approximate empirical welfare risk as per Equation D.31. We then substitute these values into Equation D.33 to provide an estimate of the objective function.

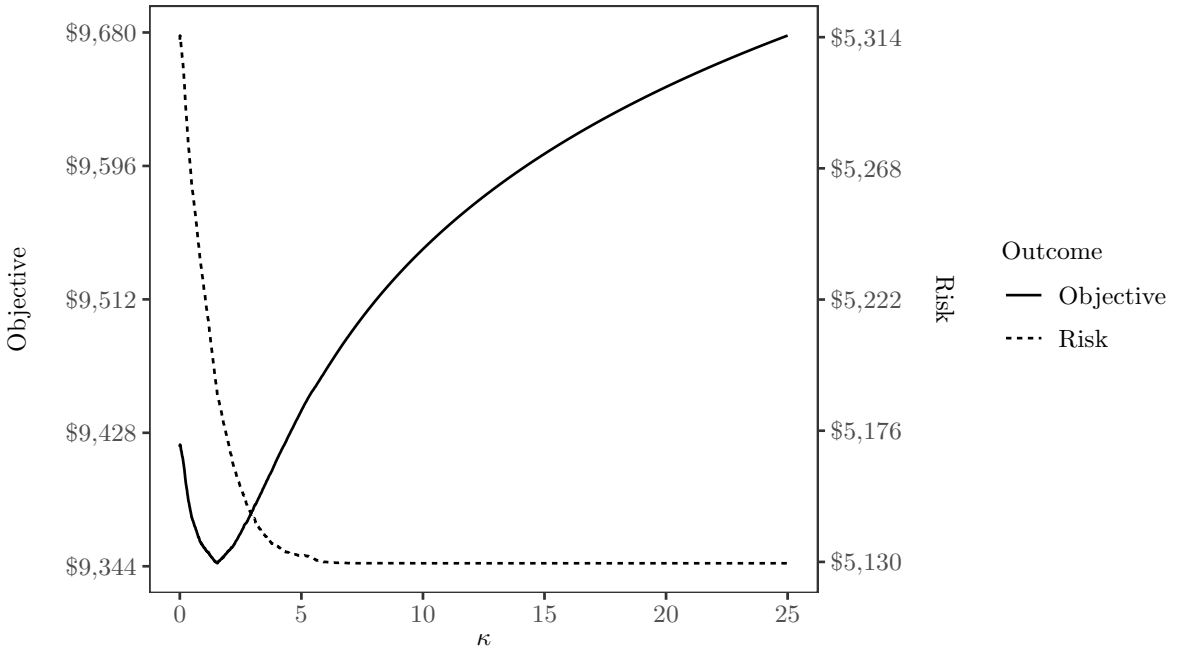
---

7 We map each variable to the unit interval by dividing through by its maximum in the sample. KITAGAWA and TETENOV (2018b) also does this. Such a change of units is useful when the domain of one variable is much larger than the domain of another and the respective coefficients on the two variables reflect this. For instance, in our sample, every individual has between seven and 18 years of education, and no individual earned more than $63,000 prior to the start of the intervention.

8 We adjust post-programme earnings by the average cost of JTPA assistance in an online appendix.

9 We design our grid so as to place an upper limit on the great-circle distance between any point on the sphere and its closest point on the grid. Our grid comprises a total of 10,116 directional vectors combined with a sequence of evenly-spaced concentrations on the zero to five interval. For reference, the surface area of the sphere is 4π, which means that our grid has a density of approximately 0.001.

FIGURE D.2
Behaviour of the objective function at μ^* given variation in κ



JTPA Study sample. This figure illustrates the shape of the objective function and its risk component at μ^* as κ is varied; high concentration is associated with low risk but incurs a large penalty for divergence from the uniform prior.

We find that the objective function is minimised (amongst the class of von Mises-Fisher distributed linear assignment rules) by the stochastic assignment rule with $\kappa = 1.550$ and $\mu = \{0.883, 0.442, 0.158\}$,[10] which we label as κ^* and μ^*, respectively.

The optimal stochastic assignment rule, on average, assigns treatment to individuals in the JTPA Study sample around 83% of the time. This probability is not, however, uniform, and there is some variation in the probability with which distinct individuals are assigned treatment. This variation in assignment propensity can be seen in Figure D.1, which plots the individual characteristics of all individuals in the sample. The propensity with which individuals with distinct characteristics are assigned treatment is represented by the color of each point, and the weight given to individuals in the sample with particular characteristics is represented by the size of each point. The weight attached to a given point is proportional to the sum of post-programme earnings over all individuals with those characteristics.[11] Figure D.1 shows that the optimal stochastic assignment rule is more likely to assign treatment to an individual with few years of education and high prior earnings than an individual with more years of education and lower prior earnings, with the assignment probability ranging from 78% to 84%. The deterministic assignment rule of KITAGAWA and TETENOV (2018b), in contrast, assigns only individuals

10 This directional vector can be represented by an azimuth of 27° and an inclination of 81° using spherical coordinates.

11 To simplify Figure D.1, we scale the weights such that they sum to one.

with few years of education and low prior earnings to treatment, with around 93% of individuals assigned treatment. We plot this deterministic rule as a useful benchmark for comparison in Figure D.1.

It is important to emphasise that it is the regularisation term and, in particular, the Kullback-Leibler divergence that limits the value of the concentration parameter at the optimum, and leads to an interior probability of assignment for all individuals. This can be seen in Figure D.2, which plots the USD equivalent of the objective function (left-hand axis, solid line) and of empirical welfare risk (right-hand axis, dashed line) for a range of values of the concentration parameter, holding fixed $\mu = \mu^*$. We observe that empirical welfare risk decreases as the value of the concentration parameter increases, remaining low and constant once its value is sufficiently large.[12] The intuition here is that large values of the concentration parameter lead to stochastic assignment rules that mimic deterministic ones; we expect the mean direction to eventually coincide with the deterministic assignment rule of Kitagawa and Tetenov (2018b) as the value of the concentration parameter approaches infinity,since there does not exist a (linear) deterministic rule that can improve upon this. Tempering this preference towards large values of the concentration parameter is the Kullback-Leibler divergence of the von Mises-Fisher distribution from the uniform distribution, which is increasing at the logarithmic rate in the concentration and generates the difference between the objective function and empirical welfare risk in Figure D.2. As the value of the concentration parameter increases, the regularisation term begins to dominate.

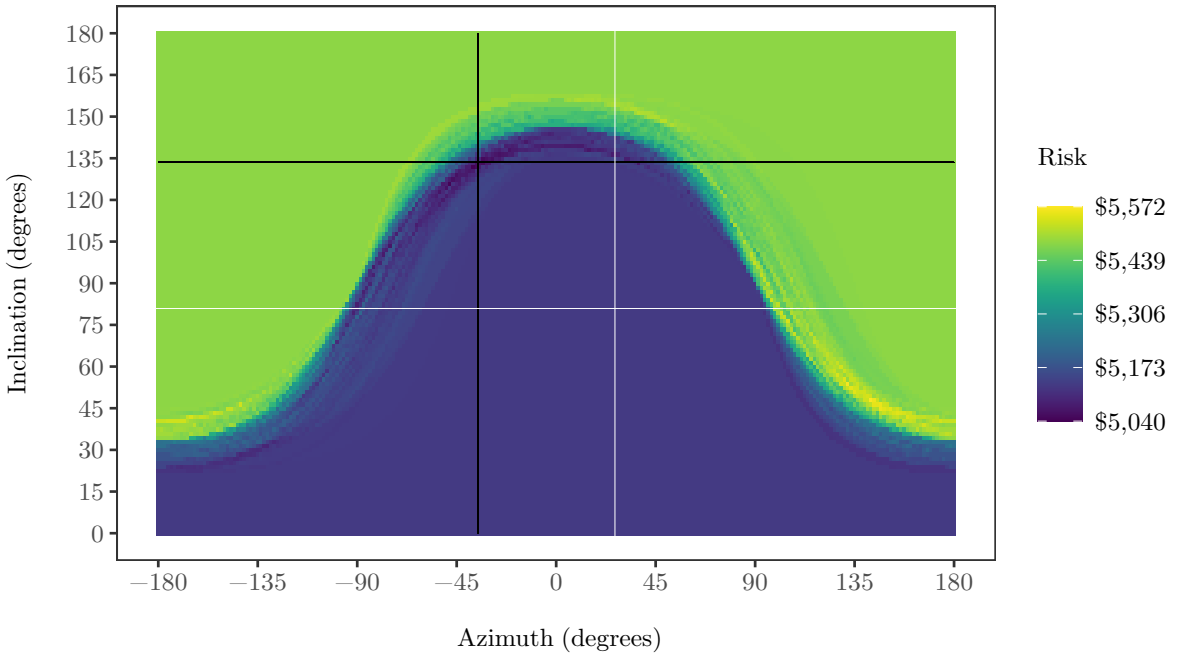
To better understand why μ^* does not coincide with the deterministic assignment rule of Kitagawa and Tetenov (2018b) holding fixed $\kappa = \kappa^*$, we refer to Figure D.3, which plots empirical welfare risk for all vectors on the unit sphere—i.e., the empirical welfare risk associated for each deterministic assignment rule in $\mathfrak{L}$. Figure D.3 utilises the spherical coordinate system

$$\{\beta_0, \beta_1, \beta_2\} = \{\cos(\theta) \cdot \sin(\phi), \sin(\theta) \cdot \sin(\phi), \cos(\phi)\} \tag{D.35}$$

where $-180^\circ \leq \theta < 180^\circ$ is the azimuth and $0^\circ \leq \phi \leq 180^\circ$ is the inclination. It is perhaps convenient to think of the azimuth as related to longitude and the inclination as related to latitude. For non-trivial values of the concentration parameter, the von Mises-Fisher distribution allocates probability mass to the sphere in such a way that its density contours are concentric about the mean direction, with points closer to this direction more likely to occur. The deterministic assignment rule of Kitagawa and Tetenov (2018b) can be seen from Figure D.3 to be located on the boundary between a high risk region (no-one treated) and a moderate risk region (everyone treated). As such, a stochastic assignment rule with a

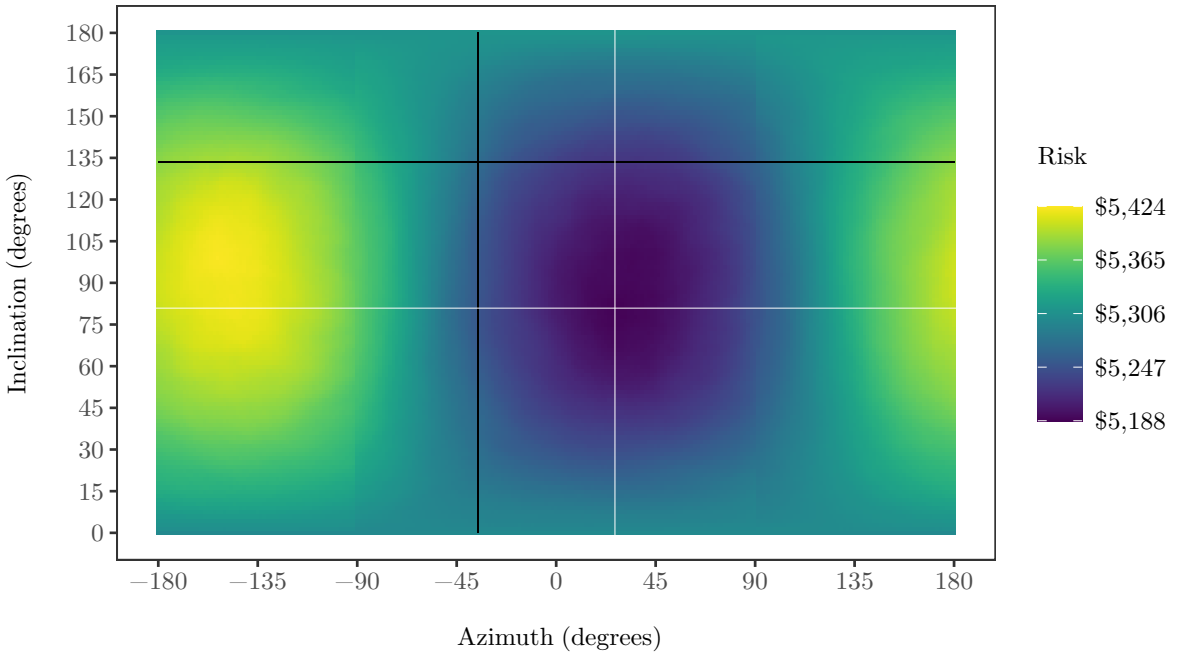
12 That empirical welfare risk decreases for small to moderate values of the concentration parameter is specific to the data and chosen mean direction, and is arguably also attributable to the lack of consideration given to the cost of treatment.

FIGURE D.3
Deterministic assignment rules and empirical welfare risk



JTPA Study sample. This figure illustrates the risk that is associated with (deterministic) assignment rules in $\mathfrak{L}$*. A spherical coordinate mapping is implemented. The intersection of the two white lines is located at* μ^**. The intersection of the two black lines is located at the optimal deterministic assignment rule of* KITAGAWA *and* TETENOV *(2018b), which attains the minimal regret amongst all deterministic linear rules.*

FIGURE D.4
Behaviour of the objective function at κ^* *given variation in the mean direction* μ



JTPA Study sample. This figure illustrates the risk that is associated with (stochastic) assignment rules in $\mathfrak{V}$*; the concentration parameter is fixed at* κ^* *whilst* μ *is varied. The intersection of the two white lines is located at* μ^**. The intersection of the two black lines is located at the optimal deterministic assignment rule of* KITAGAWA *and* TETENOV *(2018b), which attains the minimal regret amongst all deterministic linear rules.*

non-trivial value of the concentration parameter with its mean direction located at this point would approximately allocate probability mass to each of these regions in equal amounts. By shifting the mean direction towards the centre of the moderate risk region, we allocate relatively more mass to rules that induce moderate risk and less mass to rules that induce high risk, which reduces empirical welfare risk overall.

This pattern underlies what we observe in Figure D.4, which plots empirical welfare risk for all directions on the sphere holding fixed $\kappa = \kappa^*$. Despite the apparent discontinuity of risk over deterministic assignment rules, empirical welfare risk (and the objective function) appear to vary smoothly.

—| Appendix D.1 |—

Proofs

↪ *Proof of Theorem D.2.* An optimal posterior minimises

$$\mathrm{E}_{\Pi}\left(R_S\left(g\right)\right) + \sqrt{\frac{1}{2n} \cdot \left[\mathrm{E}_{\Pi}\left(\ln\left(\frac{\mathrm{d}\Pi\left(g\right)}{\mathrm{d}\Pi_0\left(g\right)}\right)\right) + \ln\left(\frac{2\sqrt{n}}{\varepsilon}\right)\right]} \text{ subject to } \int_{\mathfrak{G}} \mathrm{d}\Pi\left(g\right) = 1, \quad \inf_{g \in \mathfrak{G}} \mathrm{d}\Pi\left(g\right) \geq 0 \tag{D.36}$$

Provided that the solution satisfies the non-negativity constraints (the second constraint of Equation D.36), this is equivalent to minimising

$$\int_{\mathfrak{G}} R_S\left(g\right) \cdot \mathrm{d}\Pi\left(g\right) + \sqrt{\frac{1}{2n} \cdot \left[\int_{\mathfrak{G}} \ln\left(\frac{\mathrm{d}\Pi\left(g\right)}{\mathrm{d}\Pi_0\left(g\right)}\right) \cdot \mathrm{d}\Pi\left(g\right) + \ln\left(\frac{2\sqrt{n}}{\varepsilon}\right)\right]} + \xi \cdot \left[\Pi\left(\mathfrak{G}\right) - 1\right] \tag{D.37}$$

where ξ is the Lagrange multiplier. We separate this minimisation into two parts, by minimising Equation D.37 over Π subject to its Kullbuck-Leibler divergence from Π_0 being equal to $c \geq 0$ and, subsequently, by searching for the value of c that minimises the objective function.

We can write the constrained minimisation that forms the first part of the problem as

$$\int_{\mathfrak{G}} R_S\left(g\right) \cdot \mathrm{d}\Pi\left(g\right) + \sqrt{\frac{1}{2n} \cdot \left[c + \ln\left(\frac{2\sqrt{n}}{\varepsilon}\right)\right]} + \xi \cdot \left[\Pi\left(\mathfrak{G}\right) - 1\right] + \frac{1}{\chi} \cdot \left[\int_{\mathfrak{G}} \ln\left(\frac{\mathrm{d}\Pi\left(g\right)}{\mathrm{d}\Pi_0\left(g\right)}\right) \cdot \mathrm{d}\Pi\left(g\right) - c\right] \tag{D.38}$$

where ξ and $1/\chi$ are Lagrange multipliers, and where $c \geq 0$ is a constant, provided that the omitted non-negativity constraints are satisfied at the solution. The associated first order condition of the minimand with respect to $\mathrm{d}\Pi\left(g\right)$ is

$$R_S\left(g\right) + \xi + \frac{1}{\chi} \cdot \left[\ln\left(\frac{\mathrm{d}\Pi\left(g\right)}{\mathrm{d}\Pi_0\left(g\right)}\right) + 1\right] = 0 \tag{D.39}$$

Rearranging, we obtain

$$\mathrm{d}\Pi^{*}\left(g\right)=\frac{\exp\left(-\chi\cdot R_{S}\left(g\right)\right)}{\exp\left(1+\xi\cdot\chi\right)}\cdot\mathrm{d}\Pi_{0}\left(g\right) \tag{D.40}$$

where we emphasise that Π^{*} is a function of χ. In view of the first constraint of Equation D.37,

$$\exp\left(1+\xi\cdot\chi\right)=\int_{\mathfrak{G}}\exp\left(-\chi\cdot R_{S}\left(g\right)\right)\cdot\mathrm{d}\Pi_{0}\left(g\right) \tag{D.41}$$

and so, for all $g\in\mathfrak{G}$,

$$\mathrm{d}\Pi^{*}\left(g\right)=\frac{\exp\left(-\chi\cdot R_{S}\left(g\right)\right)}{\int_{\mathfrak{G}}\exp\left(-\chi\cdot R_{S}\left(g\right)\right)\cdot\mathrm{d}\Pi_{0}\left(g\right)}\cdot\mathrm{d}\Pi_{0}\left(g\right) \tag{D.42}$$

which integrates to one as is required. We reiterate that Equation D.42 is derived for an arbitrary (but non-negative) c, and so holds for any feasible values of χ.

For Π^{*} to satisfy $d_{\ell}\left(\Pi^{*},\Pi_{0}\right)=c$, we require that

$$\begin{aligned}c&=\int_{\mathfrak{G}}\ln\left(\frac{\mathrm{d}\Pi^{*}\left(g\right)}{\mathrm{d}\Pi_{0}\left(g\right)}\right)\cdot\mathrm{d}\Pi^{*}\left(g\right)\\&=-\int_{\mathfrak{G}}\chi\cdot R_{S}\left(g\right)\cdot\mathrm{d}\Pi^{*}\left(g\right)-\ln\left(\int_{\mathfrak{G}}\exp\left(-\chi\cdot R_{S}\left(g\right)\right)\cdot\mathrm{d}\Pi_{0}\left(g\right)\right)\end{aligned} \tag{D.43}$$

This relationship between the radius of the Kullback-Leibler ball and the inverse of the Lagrange multiplier shows that c is strictly monotonically increasing in χ provided that $R_{S}\left(g\right)$ is not constant over those g supported by Π_{0}, since

$$\frac{\mathrm{d}c}{\mathrm{d}\chi}=-\int_{\mathfrak{G}}R_{S}\left(g\right)\cdot\mathrm{d}\Pi^{*}\left(g\right)-\int_{\mathfrak{G}}\chi\cdot R_{S}\left(g\right)\cdot\left[\frac{\mathrm{d}}{\mathrm{d}\chi}\frac{\mathrm{d}\Pi^{*}\left(g\right)}{\mathrm{d}\Pi_{0}\left(g\right)}\right]\cdot\mathrm{d}\Pi_{0}\left(g\right)+\int_{\mathfrak{G}}R_{S}\left(g\right)\cdot\mathrm{d}\Pi^{*}\left(g\right) \tag{D.44}$$

$$=\chi\cdot\int_{\mathfrak{G}}\left[R_{S}\left(g\right)-R_{S}^{\Pi^{*}}\right]^{2}\cdot\mathrm{d}\Pi^{*}\left(g\right) \tag{D.45}$$

$$\geq 0 \tag{D.46}$$

where Equation D.46 is strict if $\chi>0$ and $R_{S}\left(g\right)\neq R_{S}^{\Pi^{*}}$ for at least some g supported by Π_{0}. That χ is non-negative is evident upon further consideration of Equation D.42: posterior distributions associated with negative values of χ assign more mass to policies that are associated with high levels of risk and so cannot be optimal. Accordingly, in the second part of the optimisation, we substitute Π^{*} into Equation D.38 in place of Π and solve the unconstrained minimisation problem with respect to χ. That

is, we solve

$$\min_{\chi}\left(R_S^{\Pi^*}+\underbrace{\sqrt{\frac{1}{2n}\cdot\left[-\chi\cdot R_S^{\Pi^*}-\ln\left(\int_{\mathfrak{G}}\exp\left(-\chi\cdot R_S\left(g\right)\right)\cdot\mathrm{d}\Pi_0\left(g\right)\right)+\ln\left(\frac{2\sqrt{n}}{\varepsilon}\right)\right]}}_{\text{Penalty}}\right) \tag{D.47}$$

We note that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\chi}R_S^{\Pi^*}&=\frac{\mathrm{d}}{\mathrm{d}\chi}\int_{\mathfrak{G}}R_S\left(g\right)\cdot\frac{\exp\left(-\chi\cdot R_S\left(g\right)\right)}{\int_{\mathfrak{G}}\exp\left(-\chi\cdot R_S\left(g\right)\right)\cdot\mathrm{d}\Pi_0\left(g\right)}\cdot\mathrm{d}\Pi_0\left(g\right)\\&=\int_{\mathfrak{G}}R_S\left(g\right)\cdot\frac{\mathrm{d}}{\mathrm{d}\chi}\frac{\exp\left(-\chi\cdot R_S\left(g\right)\right)}{\int_{\mathfrak{G}}\exp\left(-\chi\cdot R_S\left(g\right)\right)\cdot\mathrm{d}\Pi_0\left(g\right)}\cdot\mathrm{d}\Pi_0\left(g\right)\\&=-\int_{\mathfrak{G}}\left[R_S\left(g\right)-R_S^{\Pi^*}\right]^2\cdot\mathrm{d}\Pi^*\left(g\right)\end{aligned} \tag{D.48}$$

which we interpret as the variance of empirical welfare risk under Π^*. Using this result, we further note that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\chi}\text{Penalty}&=\frac{1}{4n\cdot\text{Penalty}}\cdot\left(-R_S^{\Pi^*}-\chi\cdot\frac{\mathrm{d}}{\mathrm{d}\chi}R_S^{\Pi^*}+\frac{\int_{\mathfrak{G}}R_S\left(g\right)\cdot\exp\left(-\chi\cdot R_S\left(g\right)\right)\cdot\mathrm{d}\Pi_0\left(g\right)}{\int_{\mathfrak{G}}\exp\left(-\chi\cdot R_S\left(g\right)\right)\cdot\mathrm{d}\Pi_0\left(g\right)}\right)\\&=\frac{1}{4n\cdot\text{Penalty}}\cdot-\chi\cdot\frac{\mathrm{d}}{\mathrm{d}\chi}R_S^{\Pi^*}\\&=\frac{1}{4n\cdot\text{Penalty}}\cdot\chi\cdot\int_{\mathfrak{G}}\left[R_S\left(g\right)-R_S^{\Pi^*}\right]^2\cdot\mathrm{d}\Pi^*\left(g\right)\end{aligned} \tag{D.49}$$

Together Equations D.48 and D.49 imply that the associated first order condition of Equation D.47 with respect to χ is

$$\frac{1}{4n\cdot\text{Penalty}}\cdot\chi\cdot\int_{\mathfrak{G}}\left[R_S\left(g\right)-R_S^{\Pi^*}\right]^2\cdot\mathrm{d}\Pi^*\left(g\right)=\int_{\mathfrak{G}}\left[R_S\left(g\right)-R_S^{\Pi^*}\right]^2\cdot\mathrm{d}\Pi^*\left(g\right) \tag{D.50}$$

Provided—trivially, we might add—that Π^* is not degenerate and there is variation in empirical welfare risk (the conditions under which Equation D.46 is strict), then this condition reduces to

$$\chi=4n\cdot\sqrt{\frac{1}{2n}\cdot\left[d_\ell\left(\Pi^*,\Pi_0\right)+\ln\left(\frac{2\sqrt{n}}{\varepsilon}\right)\right]} \tag{D.51}$$

which is exactly the condition that appears in Theorem D.2.

Although Equations D.42 and D.51 characterise a possible interior solution of the optimisation, we have yet to guarantee that this proposed solution is a global optimum. To address this, we show that the two possible corner solutions are sub-optimal, such that the first order conditions from which Equation D.51 are derived are applicable. Continuity and differentiability of Equation D.38 are then sufficient[13] to

13 It is possible to show that the difference between the left- and right-hand sides of Equation D.51 is negative at $\chi=0$

guarantee that a fixed point satisfying Equation D.51 exists.

Neither Π_0 nor a degenerate distribution are optimal. To establish that Π_0 is not optimal, we note that Π^* coincides with Π_0 when $\chi = 0$. By marginally increasing χ, such that we move in the direction of Π^*, we obtain a probability distribution that is in the interior. It suffices to show that such a probability distribution reduces the value of the objective function relative to Π_0. Evaluating Equations D.48 and D.49 at $\chi = 0$, we obtain

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\chi} \left[R_S^{\Pi^*} + \text{Penalty}\right]\Big|_{\chi=0} &= \frac{\mathrm{d}}{\mathrm{d}\chi} R_S^{\Pi_0} \\ &< 0 \end{aligned} \tag{D.52}$$

as we require, with any χ satisfying

$$0 < \chi < 2n \cdot R^{\Pi_0} \tag{D.53}$$

yielding a strictly lower value than $\chi = 0$. To establish that a degenerate distribution is not optimal, we note that such a distribution implies infinite divergence from Π_0 (if Π_0 were itself degenerate then our analysis would be meaningless since the prior and posterior distributions would always coincide; and if Π_0 is atomic then the discrete case would apply). Given that Π_0 attains a finite value of the objective function, however, Π_0 is always preferred to a degenerate distribution, and so a degenerate distribution cannot be optimal ■

To prove Theorem D.3, we rely on several intermediate results.

Lemma D.2. *The circular variance of a m-variate von Mises-Fisher random vector with concentration $\kappa \geq 0$ is bounded from above by*

$$\frac{m - 1 + \sqrt{4\kappa^2 + [m+1]^2} - 2\kappa}{m - 1 + \sqrt{4\kappa^2 + [m+1]^2}} \leq \frac{2m}{m - 1 + \sqrt{4\kappa^2 + [m+1]^2}} \tag{D.54}$$

and is an $\mathrm{O}(1/\kappa)$ function

Lemma D.2 is trivially established by subtracting the lower bound of Equation B.11 from one, and so is stated without proof. That the circular variance is an $\mathrm{O}(1/\kappa)$ function is otherwise established in Chapter C using a Hankel expansion.

and positive if χ is sufficiently large, with existence then established using extensions of the intermediate value theorem and its corollary, Bolzano's theorem.

Lemma D.3. *The Kullback-Leibler divergence of a m-variate von Mises-Fisher random vector with concentration $\kappa \geq 0$ is bounded from above by*

$$\frac{m-1}{2} \cdot \left[\ln\left(\frac{m-1}{2} + \sqrt{\kappa^2 + \left[\frac{m+1}{2}\right]^2} \right) - \ln(m) \right] + \sqrt{\kappa^2 + \left[\frac{m-1}{2}\right]^2} - \sqrt{\kappa^2 + \left[\frac{m+1}{2}\right]^2} + 1 \quad \text{D.55}$$

and is an $\mathrm{O}(\ln(\kappa))$ *function*

↪ *Proof of Lemma D.3.* Chapter C shows that, when Π is a von Mises-Fisher distribution and Π_0 is the uniform distribution over the hypersphere,

$$d_\ell(\Pi, \Pi_0) = \left[\frac{m}{2} - 1\right] \cdot \ln\left(\frac{\kappa}{2}\right) - \underbrace{\ln\left(\mathrm{I}_{m/2-1}(\kappa)\right)}_{\text{Bessel fn.}} - \ln\left(\Gamma\left(\frac{m}{2}\right)\right) + \underbrace{\frac{\mathrm{I}_{m/2}(\kappa)}{\mathrm{I}_{m/2-1}(\kappa)} \cdot \kappa}_{\text{Ratio fn.}} \quad \text{D.56}$$

To derive an upper bound on the Kullback-Leibler divergence that does not involve modified Bessel functions or their ratios, we replace the terms labelled Bessel fn. and Ratio fn. in Equation D.56 with appropriate lower and upper bounds, respectively. To do so, we rely on results (and notation) in Chapter B.

First, the term labelled Bessel fn. is bounded from below (recall that the term enters negatively) by

$$\frac{1}{2} \cdot \ln\left(\frac{2}{\kappa}\right) - \ln\left(\Gamma\left(\frac{m}{2}\right)\right) + \frac{m-1}{2} \cdot \ln\left(\frac{\kappa \cdot m}{m - 1 + \sqrt{4\kappa^2 + [m+1]^2}} \right) + \frac{2\kappa^2}{m - 1 + \sqrt{4\kappa^2 + [m+1]^2}} \quad \text{D.57}$$

Second, the term labelled Ratio fn. is bounded from above by

$$\frac{2\kappa^2}{m - 1 + \sqrt{4\kappa^2 + [m-1]^2}} \quad \text{D.58}$$

Substituting Equations D.57 and D.58 into Equation D.56, cancelling terms and noting that

$$\left[\frac{m}{2} - 1\right] \cdot \ln\left(\frac{\kappa}{2}\right) - \frac{1}{2} \cdot \ln\left(\frac{2}{\kappa}\right) = \frac{m-1}{2} \cdot \ln\left(\frac{\kappa}{2}\right) \quad \text{D.59}$$

we obtain

$$\frac{m-1}{2} \cdot \left[\ln\left(\frac{\kappa}{2}\right) - \ln\left(\frac{\kappa \cdot m}{\underline{a} + \sqrt{4\kappa^2 + \overline{a}^2}} \right) \right] - \frac{2\kappa^2}{\overline{a} + \sqrt{4\kappa^2 + \overline{a}^2}} + \frac{2\kappa^2}{\underline{a} + \sqrt{4\kappa^2 + \underline{a}^2}} \quad \text{D.60}$$

where $\underline{a} \doteq m - 1$ and $\overline{a} \doteq m + 1$. To obtain the required result, we simplify the first two terms of Equation D.60 and use the fact that, for any real a (in this case, equal to either $\underline{a}$ or $\overline{a}$, which are

positive reals),

$$\frac{4\kappa^2}{a+\sqrt{4\kappa^2+a^2}} = \sqrt{4\kappa^2+a^2} - a \tag{D.61}$$

which makes use of the formula for the difference of two squares ■

↪ *Proof of Theorem D.3.* With Lemmata D.2 and D.3 to hand, we begin by noting that, together, Assumption D.5 and Lemma D.1 imply that

$$R^{\tilde{\Pi}} \leq \bar{R} + \inf_{\kappa}\left(\inf_{\mu}\left(\text{First term} + \text{Second term} + \sqrt{\text{Third term}}\right)\right) \tag{D.62}$$

where

$$\begin{aligned}
\text{First term} &= \int_{\mathbb{S}^{m-1}} 2\zeta \cdot \left\|\beta - \overline{\beta}\right\|_2 \cdot \mathrm{d}\Pi\left(\beta;\kappa,\mu\right) \\
\text{Second term} &= \frac{1}{\lambda}\cdot\left[a\left(\lambda,n\right) + d_{\ell}\left(\Pi,\Pi_0\right) + \ln\left(\frac{2}{\varepsilon}\right)\right] \\
\text{Third term} &= \frac{1}{2n}\cdot\left[d_{\ell}\left(\Pi,\Pi_0\right) + \ln\left(\frac{4\sqrt{n}}{\varepsilon}\right)\right]
\end{aligned} \tag{D.63}$$

and which we note are functions (either explicit or implicit) in κ and μ.

Focusing on the First term, we use Jensen's inequality to show that

$$2\zeta \cdot \int_{\mathbb{S}^{m-1}} \left\|\beta - \overline{\beta}\right\|_2 \cdot \mathrm{d}\Pi\left(\beta;\kappa,\mu\right) \leq 2\zeta \cdot \sqrt{\int_{\mathbb{S}^{m-1}} \sum_{i=0}^{m-1} \left(\beta_i - \overline{\beta}_i\right)^2 \cdot \mathrm{d}\Pi\left(\beta;\kappa,\mu\right)} \tag{D.64}$$

$$= 2\zeta \cdot \sqrt{\mathrm{tr}\left(\mathrm{E}_{\Pi}\left(\left[\beta-\overline{\beta}\right]\left[\beta-\overline{\beta}\right]^{\top}\right)\right)} \tag{D.65}$$

Rearranging, we obtain

$$\mathrm{E}_{\Pi}\left(\left[\beta-\overline{\beta}\right]\left[\beta-\overline{\beta}\right]^{\top}\right) = \mathrm{E}_{\Pi}\left(\left[\beta-\overline{\beta}+\mathrm{E}_{\Pi}\left(\beta\right)-\mathrm{E}_{\Pi}\left(\beta\right)\right]\left[\beta-\overline{\beta}+\mathrm{E}_{\Pi}\left(\beta\right)-\mathrm{E}_{\Pi}\left(\beta\right)\right]^{\top}\right) \tag{D.66}$$

$$= \mathrm{E}_{\Pi}\left(\left[\beta-\mathrm{E}_{\Pi}\left(\beta\right)\right]\left[\beta-\mathrm{E}_{\Pi}\left(\beta\right)\right]^{\top}\right) + \left[\mathrm{E}_{\Pi}\left(\beta\right)-\overline{\beta}\right]\left[\mathrm{E}_{\Pi}\left(\beta\right)-\overline{\beta}\right]^{\top} \tag{D.67}$$

$$= \text{Variance}\left(\beta\middle|\,\beta\sim\Pi\right) + \left[\frac{\mathrm{I}_{m/2}\left(\kappa\right)}{\mathrm{I}_{m/2-1}\left(\kappa\right)}\cdot\mu-\overline{\beta}\right]\left[\frac{\mathrm{I}_{m/2}\left(\kappa\right)}{\mathrm{I}_{m/2-1}\left(\kappa\right)}\cdot\mu-\overline{\beta}\right]^{\top} \tag{D.68}$$

which relies on results from Chapter C relating to the first two moments of the von Mises-Fisher family of distributions. Restricting the set of von Mises-Fisher distributions to those satisfying $\mu = \overline{\beta}$, the

First term is bounded by

$$2\zeta \cdot \sqrt{\operatorname{tr}\left(\mathrm{E}\left(\left[\beta-\overline{\beta}\right]\left[\beta-\overline{\beta}\right]^{\top}\right)\right)} = 2\zeta \cdot \sqrt{\operatorname{tr}\left(\text{Variance}\left(\beta\middle|\,\beta\sim\Pi\right)+\left[\frac{\mathrm{I}_{m/2}\left(\kappa\right)}{\mathrm{I}_{m/2-1}\left(\kappa\right)}-1\right]^{2}\cdot\mu\mu^{\top}\right)} \\ = \sqrt{8\zeta^{2}}\cdot\sqrt{1-\frac{\mathrm{I}_{m/2}\left(\kappa\right)}{\mathrm{I}_{m/2-1}\left(\kappa\right)}} \tag{D.69}$$

which is proportional to the square root of the circular variance.[14] Lemma D.2 details the behaviour of the circular variance. Using Lemma D.2, Equation D.69 can be bounded from above by

$$\sqrt{8\zeta^{2}\cdot\frac{m-1+\sqrt{4\kappa^{2}+\left[m+1\right]^{2}}-2\kappa}{m-1+\sqrt{4\kappa^{2}+\left[m+1\right]^{2}}}} \leq 4\zeta\cdot\sqrt{\frac{m}{2\kappa}} \tag{D.70}$$

where the right-hand side follows from concavity of the square root, and by decreasing the denominator via elimination of $m-1$ and $m+1$ (both are non-negative constants). We substitute the right-hand side of Equation D.70 into Equation D.62 in place of the First term.

We now focus on the Kullback-Leibler divergence, which appears in both the Second term and the Third term. We observe that the difference between the square roots in Equation D.55 is increasing in κ (i.e, the difference becomes less negative as κ increases and as κ increases in importance relative to m). As such, it suffices to omit the difference between the square roots in Equation D.55 and to simply bound the Kullback-Leibler divergence from above by

$$\frac{m-1}{2}\cdot\ln\left(\frac{m-1+\sqrt{4\kappa^{2}+\left[m+1\right]^{2}}}{2m}\right)+1 \leq \frac{m-1}{2}\cdot\ln\left(\kappa+1\right)+1 \tag{D.71}$$

which is at least one due to the non-negativity of κ.

Substituting Equations D.70 and D.71 as upper bounds on the infimand of Equation D.62, we obtain

$$4\zeta\cdot\sqrt{\frac{m}{2\kappa}}+\frac{1}{\lambda}\cdot\left[a\left(\lambda,n\right)+\frac{m-1}{2}\cdot\ln\left(\kappa+1\right)+\ln\left(\frac{2\mathrm{e}}{\varepsilon}\right)\right]+\sqrt{\frac{1}{2n}\cdot\left[\frac{m-1}{2}\cdot\ln\left(\kappa+1\right)+\ln\left(\frac{4\mathrm{e}\cdot\sqrt{n}}{\varepsilon}\right)\right]} \tag{D.72}$$

which we emphasise is an upper bound on the infimum and is not dependent of β. Our objective is to minimise Equation D.72 by appropriately choosing κ and λ alongside the functional form of a. Accordingly, we let $\kappa = n$ and $\lambda = \sqrt{n}$ alongside $a\left(\lambda,n\right)/\lambda = 1/\lambda$. Given these choices, we can write

14 Recall that the right-hand side of Equation D.62 is preceded by an infimum over μ; replacing the infimum with a specific value delivers an upper bound.

Equation D.72 as

$$4\zeta \cdot \sqrt{\frac{m}{2n}} + \frac{m-1}{2\sqrt{n}} \cdot \ln(n+1) + \frac{1}{\sqrt{n}} \cdot \ln\left(\frac{2e^2}{\varepsilon}\right) + \sqrt{\frac{1}{2n} \cdot \left[\frac{m-1}{2} \cdot \ln(n+1) + \ln\left(\frac{4e \cdot \sqrt{n}}{\varepsilon}\right)\right]} \tag{D.73}$$

which, using the fact that $1/\ln(n)$ and $\ln(n+1)/\ln(n)$ are decreasing in n, we can upper bound by

$$\frac{\ln(n)}{\sqrt{n}} \cdot \frac{1}{\ln(8)} \cdot \left[\underbrace{4\zeta \cdot \sqrt{\frac{m}{2}} + \frac{m-1}{2} \cdot \ln(9) + \ln\left(\frac{2e^2}{\varepsilon}\right) + \sqrt{\frac{1}{2} \cdot \left[\frac{m-1}{2} \cdot \ln(9) + \ln\left(\frac{4e \cdot \sqrt{n}}{\varepsilon}\right)\right]}}_{\text{Universal constant}}\right] \tag{D.74}$$

where we rely on the maintained assumption of Theorem D.3 that $n \geq 8$. The Universal constant is a decreasing function of ε ■

—| Appendix D.2 |—

Accounting for the cost of treatment in the JTPA Study sample

In what follows, we distinguish between the JTPA Study sample, which we sometimes refer to as the raw data, and the cost-adjusted JTPA Study sample, which we sometimes refer to as the costed data. The costed data subtracts the cost of treatment from the outcome of interest (post-programme earnings) of all treated individuals.[15] Following Kitagawa & Tetenov (2018), we assume that the cost of treatment is \$774. We then proceed to search for the optimal stochastic assignment rule, applying the same grid search approach as we outlined in the main text for the raw data, using the costed data.

We find that the objective function is minimised by the stochastic assignment rule with $\kappa = 0.560$ and $\mu = \{+0.872, +0.490, +0.018\}$, which we label κ^a and μ^a, respectively.[16] The value of the objective function and the empirical welfare risk induced by this assignment rule are equivalent to \$9,302 and \$5,155, respectively.

In comparison, Kitagawa and Tetenov (2018b) estimates that the deterministic assignment rule defined by $\beta = \{+0.117, -0.990, -0.086\}$ minimises empirical welfare risk. Relative to the raw analysis, this assignment rule ascribes less weight to individual characteristics, and essentially determines that all individuals with low pre-programme earnings (earnings of around \$5,000 or less) should be treated, irrespective of education.

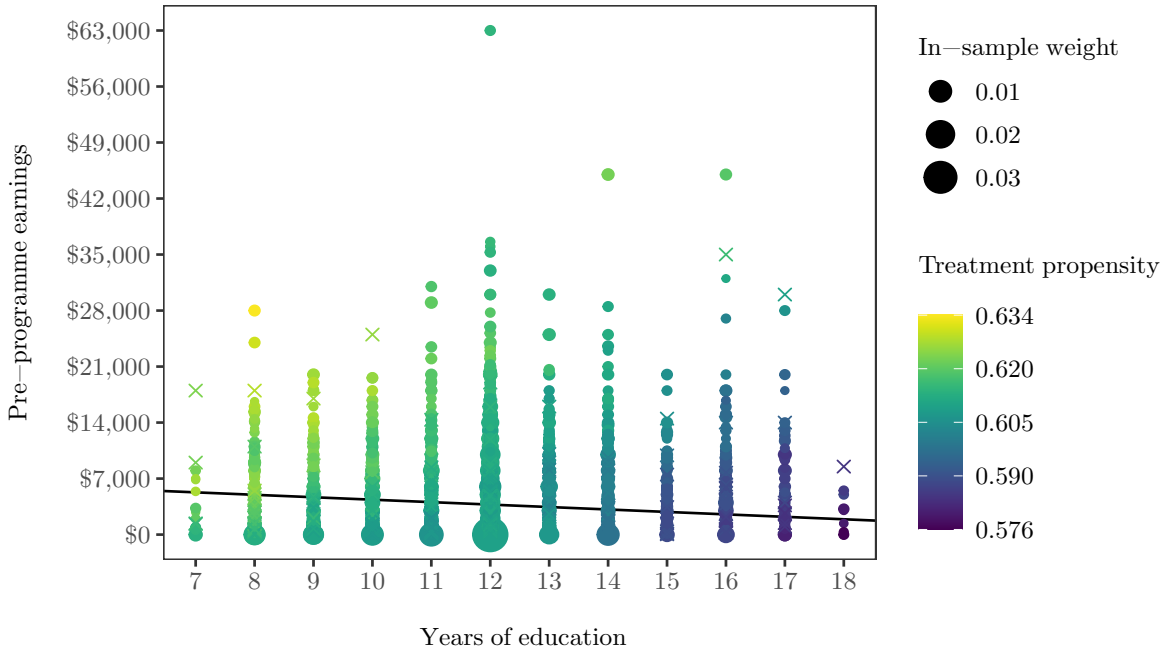
Figures D.5 to D.8 are intended to be comparable to the figures in the main text. There are two differences

15 We observe that this leads individuals with zero post-programme earnings to have a negative outcome. This violates Assumption 1, which requires that all outcomes be bounded and non-negative. We, nonetheless, proceed with this adjustment of post-programme earnings as is. An alternative would be to add \$774 to all post-programme earnings before subtracting the assumed cost of treatment.

16 This directional vector can be represented by an azimuth of 29° and an inclination of 89° using spherical coordinates.

FIGURE D.5

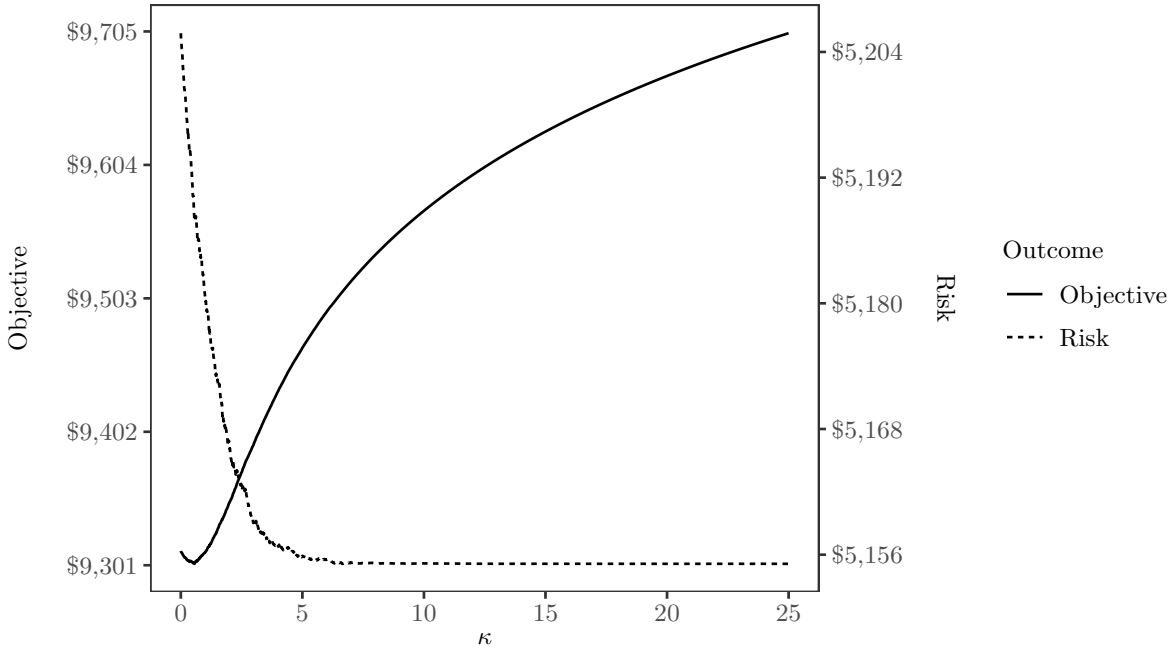
Variation in treatment propensity across individuals in the cost-adjusted JTPA Study sample



Cost-adjusted JTPA Study sample. This figure illustrates the treatment propensity of individuals under the posterior assignment rule that is induced by $\{\kappa^a, \mu^a\}$. Each point represents the individual characteristics of an individual or several individuals (crosses denote individuals with negative in-sample weight). For comparison, individuals to the left of the solid diagonal line are assigned treatment under the optimal deterministic assignment rule of KITAGAWA *and* TETENOV *(2018b).*

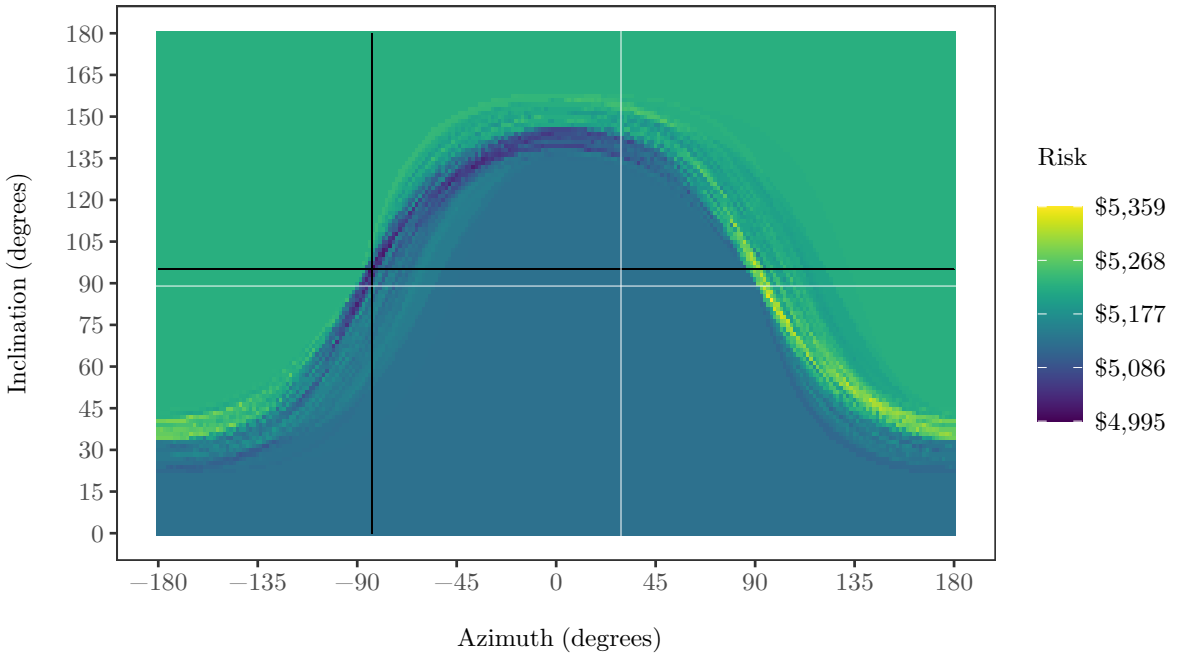
FIGURE D.6

Behaviour of the objective function at μ^* given variation in κ



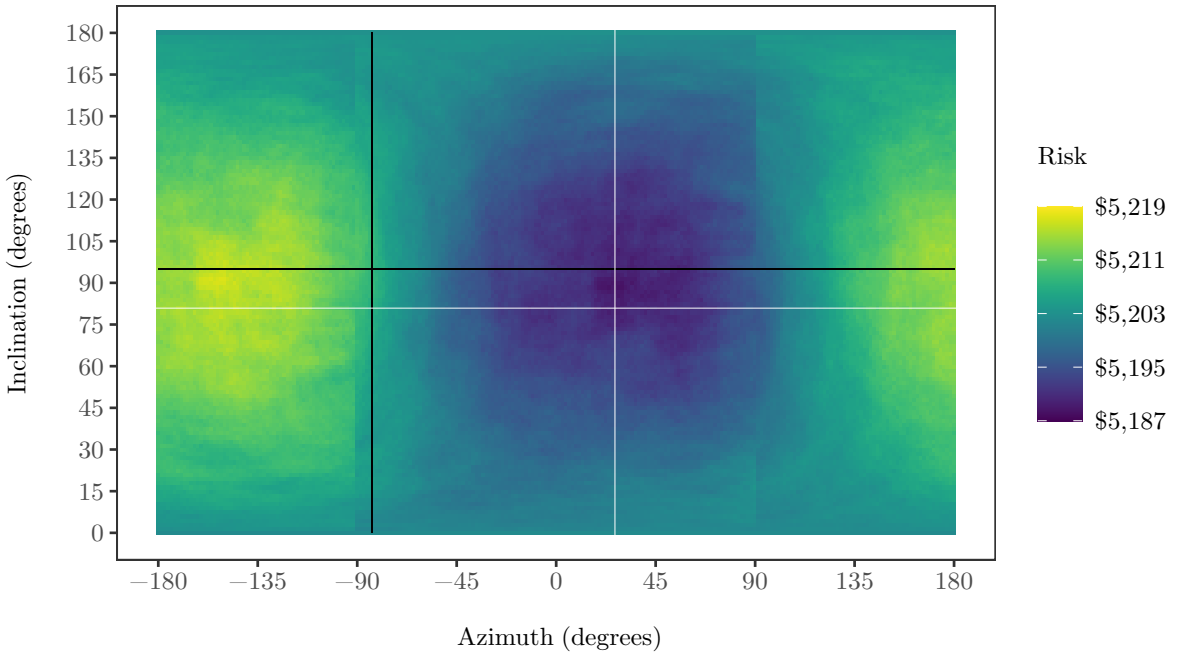
Cost-adjusted JTPA Study sample. This figure illustrates the shape of the objective function and its risk component at μ^a as κ is varied.

FIGURE D.7
Deterministic assignment rules and empirical welfare risk



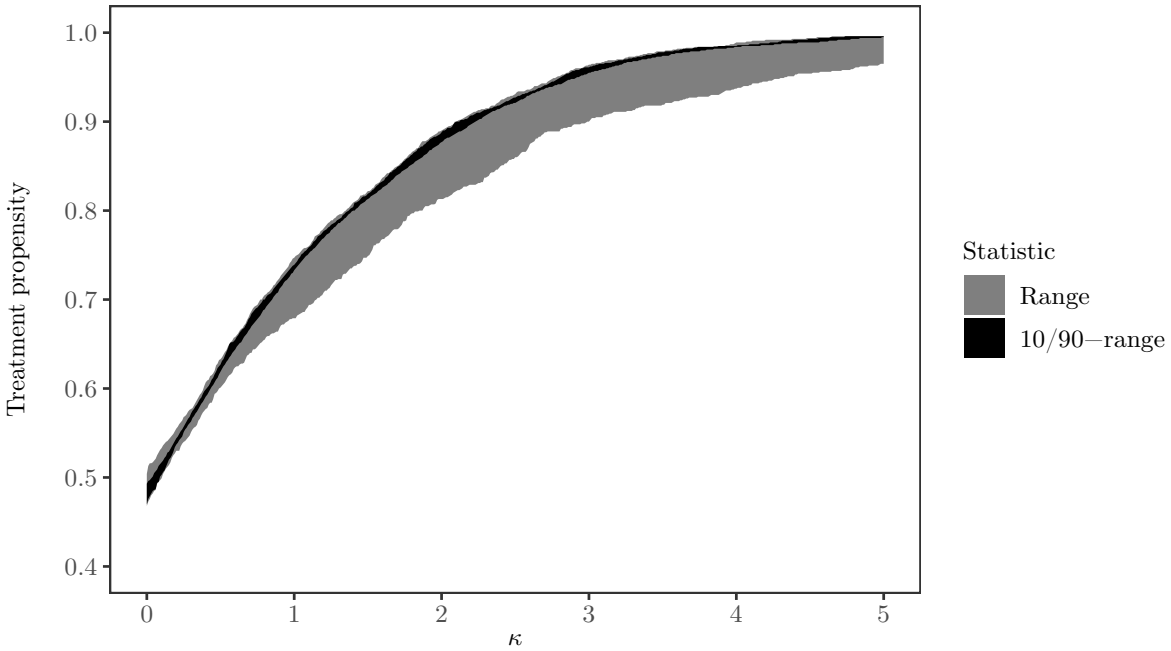
Cost-adjusted JTPA Study sample. This figure illustrates the risk that is associated with (deterministic) assignment rules in $\mathfrak{F}$. A spherical coordinate mapping is implemented. The intersection of the two white lines is located at μ^a. The intersection of the two black lines is located at the optimal deterministic assignment rule of KITAGAWA and TETENOV (2018b), which attains the minimal regret amongst all deterministic linear rules.

FIGURE D.8
Behaviour of the objective function at κ^a given variation in μ



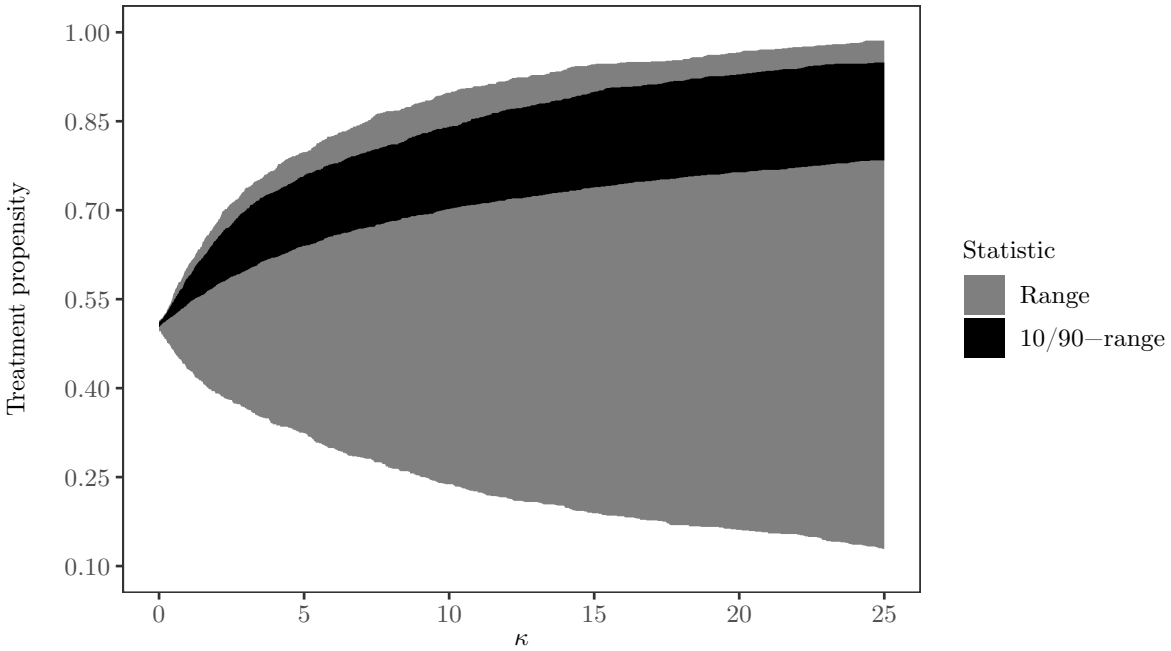
Cost-adjusted JTPA Study sample. This figure illustrates the risk that is associated with (stochastic) assignment rules in $\mathfrak{V}$; the concentration parameter is fixed at κ^a whilst μ is varied. The intersection of the two white lines is located at μ^a. The intersection of the two black lines is located at the optimal deterministic assignment rule of KITAGAWA and TETENOV (2018b), which attains the minimal regret amongst all deterministic linear rules.

FIGURE D.9
Distribution of treatment propensity across individuals (adjusted) at $\{\kappa^a, \mu^a\}$



Cost-adjusted JTPA Study sample. This figure illustrates the variation in treatment propensity across individuals. We draw 1,000 directional vectors and count how many of these vectors, implemented as deterministic assignment rules, assign each individual to treatment. We sort individuals by how often they are assigned to treatment and summarise the implied distribution by the range and the 10^{th} *and* 90^{th} *quantiles.*

FIGURE D.10
Distribution of treatment propensity across individuals (raw) at $\{\kappa^*, \mu^*\}$



JTPA Study sample. This figure illustrates the variation in treatment propensity across individuals. We draw 1,000 directional vectors and count how many of these vectors, implemented as deterministic assignment rules, assign each individual to treatment. We sort individuals by how often they are assigned to treatment and summarise the implied distribution by the range and the 10^{th} *and* 90^{th} *quantiles.*

of note between the results of the raw and costed analyses. First, in Figure D.6, the relationship between the concentration parameter and empirical welfare risk is much more muted, with risk only decreasing by around \$50 as the value of the concentration parameter increases (over the interval that we study). In contrast, empirical welfare risk decreases by around \$200 when the cost of treatment is not accounted for. This reflects the second difference, which is apparent in Figure D.7. Specifically, that the difference in empirical welfare risk between those assignment rules that assign everyone to treatment versus those that assign no-one to treatment is much smaller. The benefit to an increase in the value of the concentration parameter, which is a reduction in the probability mass allocated to those assignment rules that assign no-one to treatment is, accordingly, smaller. The implication is that the penalty term starts to dominate the objective function for smaller values of the concentration parameter.

To supplement our analysis, we include two additional figures in Figures D.9 and D.10. These figures are intended to give some idea about how the concentration of the von Mises-Fisher distribution influences the propensity with which individuals are assigned treatment. For example, given a particular instance of the von Mises-Fisher distribution, one individual might have a high propensity of assignment whereas another might have a low propensity. In other words, the number of directional vectors that are drawn from the von Mises-Fisher distribution for which the first individual is assigned treatment is greater than for the second individual. As the value of the concentration parameter increases, so these directional vectors concentrate around the mean direction, and the propensity of assignment to treatment approaches zero or one for each individual. We see from Figure D.9 that, when $\mu = \mu^*$, the propensity of assignment in the costed data tends towards one for all individuals. In contrast, and to illustrate the possibility of complete dichotomy, for the raw data, we see from Figure D.10 that the propensity of assignment to treatment diverges as the value of the concentration parameter increases. We observe in both datasets that much of the variation in propensity is driven by a few individuals in the tail, who are likely those individuals whose individual characteristics make them outliers. In both cases, the median propensity of treatment (i.e., the probability with which the average individual in the sample is assigned treatment) tends towards one.

—| Appendix D.3 |—

Numerical simulations

We propose several experiments that investigate how various aspects of the sample data that is available to the social planner affect the posterior distribution and its shape when the specified prior distribution is uniform over the sphere. We conduct these experiments to better understand some of the empirical results that we obtain for the main paper.

Aspects of the sample data that we vary include the number of observations, the outcome of interest, and individual characteristics. We conduct these experiments with simulated data and through manipulation of the existing empirical application. Each experiment is comparable to the existing empirical application in any case, in that observed individual characteristics are taken to be education and pre-programme earnings alongside an intercept term, and the outcome of interest is taken to be post-programme earnings. We investigate how these aspects and our variation of them affect the objective function.

The specific questions that we ask, and that inform the design of our experiments, are as follows.

- How does the number of observations influence the shape of the objective function?
- How does the relative influence of education and pre-programme earnings on post-programme earnings influence empirical welfare risk?
- How does the distribution of individual characteristics influence the shape of the objective function?

To address these questions we propose a series of linear specifications that satisfy the bounded outcomes assumption that we require. We concede that these specifications are somewhat contrived, reflecting our need to balance tractability with the requirements of the bounded outcomes assumption. We suppose that

$$\begin{aligned} 0 \leq & X_{\text{earn}} \leq 1 \\ 0 \leq & X_{\text{educ}} \leq 1 \end{aligned} \tag{D.75}$$

which can always be maintained via an appropriate affine map of the individual characteristics (a step that we undertake in any case), and propose that

$$\begin{aligned} Y_1 &= \boldsymbol{X}^\top \alpha_1 + V_1 \text{ with } V_1 \sim \mathrm{N}(0, \sigma_1^2) \text{ such that } 0 \leq V_1 \leq c_v \\ Y_0 &= \boldsymbol{X}^\top \alpha_0 + V_0 \text{ with } V_0 \sim \mathrm{N}(0, \sigma_0^2) \text{ such that } 0 \leq V_0 \leq c_v \end{aligned} \tag{D.76}$$

with $\alpha_1 \in \mathbb{S}^2$ and $\alpha_0 \in \mathbb{S}^2$. The advantage of this specification is that it provides a clear interpretation of the influence of education and pre-programme earnings on post-programme earnings, and facilitates addressing all of the questions that we pose. Moreover, the potential outcomes satisfy

$$\begin{aligned} 0 \leq & Y_1 \leq \alpha_1 + c_v \\ 0 \leq & Y_0 \leq \alpha_1 + c_v \end{aligned} \tag{D.77}$$

We note that the conditional expectations of the potential outcomes are

$$\begin{aligned} \mathrm{E}(Y_1|\boldsymbol{X}) &= \boldsymbol{X}^\top \alpha_1 + \frac{\Phi'(0) - \Phi'(c_v/\sigma_1)}{\Phi(c_v/\sigma_1) - \Phi(0)} \cdot \sigma_1 = \boldsymbol{X}^\top \tilde{\alpha}_1 \\ \mathrm{E}(Y_0|x) &= \boldsymbol{X}^\top \alpha_0 + \frac{\Phi'(0) - \Phi'(c_v/\sigma_0)}{\Phi(c_v/\sigma_0) - \Phi(0)} \cdot \sigma_0 = \boldsymbol{X}^\top \tilde{\alpha}_0 \end{aligned} \tag{D.78}$$

which are both affine functions of individual characteristics.[17] This is an attractive property that we exploit.

Knowledge of the data generating process (Equation D.76 in our framework) is sufficient to determine the optimal assignment rule.[18] Specifically, the optimal assignment rule satisfies

$$g\left(\boldsymbol{X}\right) = 1\left(\mathrm{E}\left(Y_1 \middle| \boldsymbol{X}\right) \geq \mathrm{E}\left(Y_0 \middle| \boldsymbol{X}\right)\right) \tag{D.79}$$

When the conditional expectation of the potential outcomes are affine functions of individual characteristics (and assignment to treatment is at random, which is an assumption that we implicitly maintain throughout) then the optimal assignment rule belongs to the LES class. Hence, Equation D.78 guarantees that the optimal policy has the specific form

$$g\left(\boldsymbol{X}\right) = 1\left(\boldsymbol{X}^\top \left[\tilde{\alpha}_1 - \tilde{\alpha}_0\right] / \|\tilde{\alpha}_1 - \tilde{\alpha}_0\|_2 \geq 0\right) \tag{D.80}$$

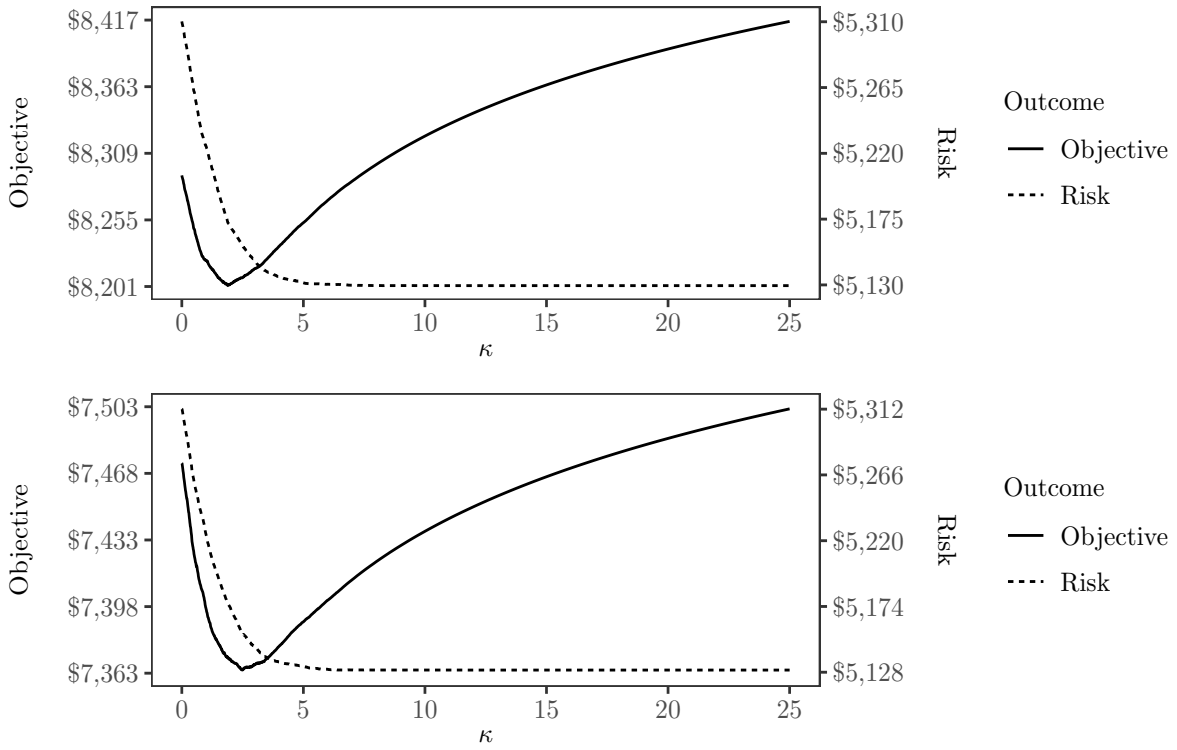
which is not only a member of the LES class but emphasises that the optimal assignment rule can be summarised by a vector on the sphere (provided that the potential outcomes have distinct process). This result provides motivation for our reliance on spherical distributions and, in particular, the von Mises-Fisher distribution. We observe that, abstracting from the issue of sampling variation, empirical welfare risk is minimised when the mean direction of the von Mises-Fisher distribution coincides with the policy defined in Equation D.80. This insight allows us to focus exclusively on the influence of the concentration parameter on empirical welfare risk and the objective function.

Throughout, we are careful to sample using inversion-based pseudo-random sampling methods where possible. Compared to rejection sampling-based pseudo-random sampling methods, inversion-based methods are able to guarantee comparability across experiments despite differences in parameter values. We also note that where we use data from the JTPA Study sample, this data is not adjusted for the

17 Other commonly invoked models fail to meet our test of tractability or the requirements of the bounded outcomes assumption: the standard censored outcome model generates a rectified normal distribution that is non-linear in individual characteristics and is unbounded from above; logarithmic transformation of the outcome implies a non-linear transformation of individual characteristics.

18 We do not make any claims about uniqueness in what follows and, indeed, presented with a finite sample of individual characteristics, it is likely that several assignment rules can attain the same partition of individuals as what we refer to as the optimal assignment rule.

FIGURE D.11
Behaviour of the objective function at μ^* given variation in κ



JTPA Study sample (artificially inflated by copying observations). This figure illustrates how the objective function and its risk component change as the number of observations is doubled (upper panel) and quadrupled (lower panel). We hold μ fixed at μ^* in each case, which differs according to the sample size, and vary κ.

cost of treatment (i.e., we use the raw data). We continue to label those values of the parameters that minimise the objective function by κ^* and μ^*, which we emphasise can vary across the various experiments that we conduct.

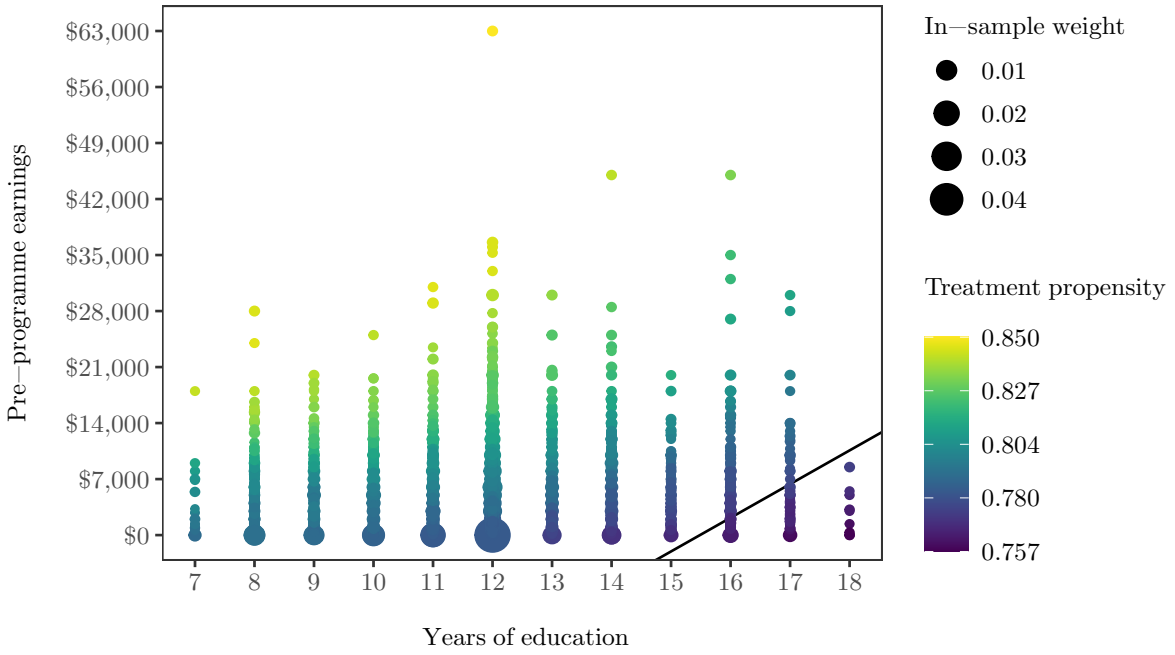
—| SUBAPPENDIX D.3.A |—

VARYING THE NUMBER OF OBSERVATIONS

To investigate the influence that the number of observations has on the objective function and on empirical welfare risk, we copy each observation in the JTPA Study sample either once or three times, thereby doubling and quadrupling the number of individuals in the sample. The mechanical effect of this change is to mute the influence of the penalty term in the objective function, without affecting empirical welfare risk. The immediate implication is that larger values of the concentration parameter can be sustained, since the penalty due to deviating from the uniform distribution is relatively smaller for every such deviation.

We plot the objective function and empirical welfare risk for the doubled and quadrupled JTPA Study samples in Figure D.11. We observe that increasing the number of observations has the direct effect of

FIGURE D.12
(Experiment 1) Variation in treatment propensity across individuals



Simulated data generated using Equation D.82 in conjunction with data from the JTPA Study sample. This figure illustrates the treatment propensity of individuals in a simulated sample under the posterior assignment rule that is induced by $\{\kappa^*, \mu^*\}$. Each point represents the individual characteristics of an individual or several individuals. For comparison, individuals to the left of the solid diagonal line are assigned treatment under the oracle assignment rule of Equation D.83.

decreasing the magnitude of the objective function. Moreover, increasing the number of observations also leads to an increase in κ^*, which increases from 1.550 to 1.890 and then to 2.490. We emphasise that μ^* is also not the same across the two cases. The intuition here is that an increase in the value of the concentration parameter leads to the concentric contour map of the density function becoming more tightly arranged around the mean direction (whatever that may be). Locating the mean direction closer to the boundary between high and moderate regret regions say, such as where the deterministic assignment rule of KITAGAWA and TETENOV (2018b) is located, incurs less of a penalty in this instance since the density function assigns less probability mass to the high regret region than it would for a smaller value of the concentration parameter. For the doubled sample we find that $\mu^* = \{+0.812, +0.577, +0.088\}$, whilst for the quadrupled sample we find that $\mu^* = \{+0.917, +0.394, +0.053\}$.[19]

—| SUBAPPENDIX D.3.B |—

THE VARIANCE OF POST-PROGRAMME EARNINGS

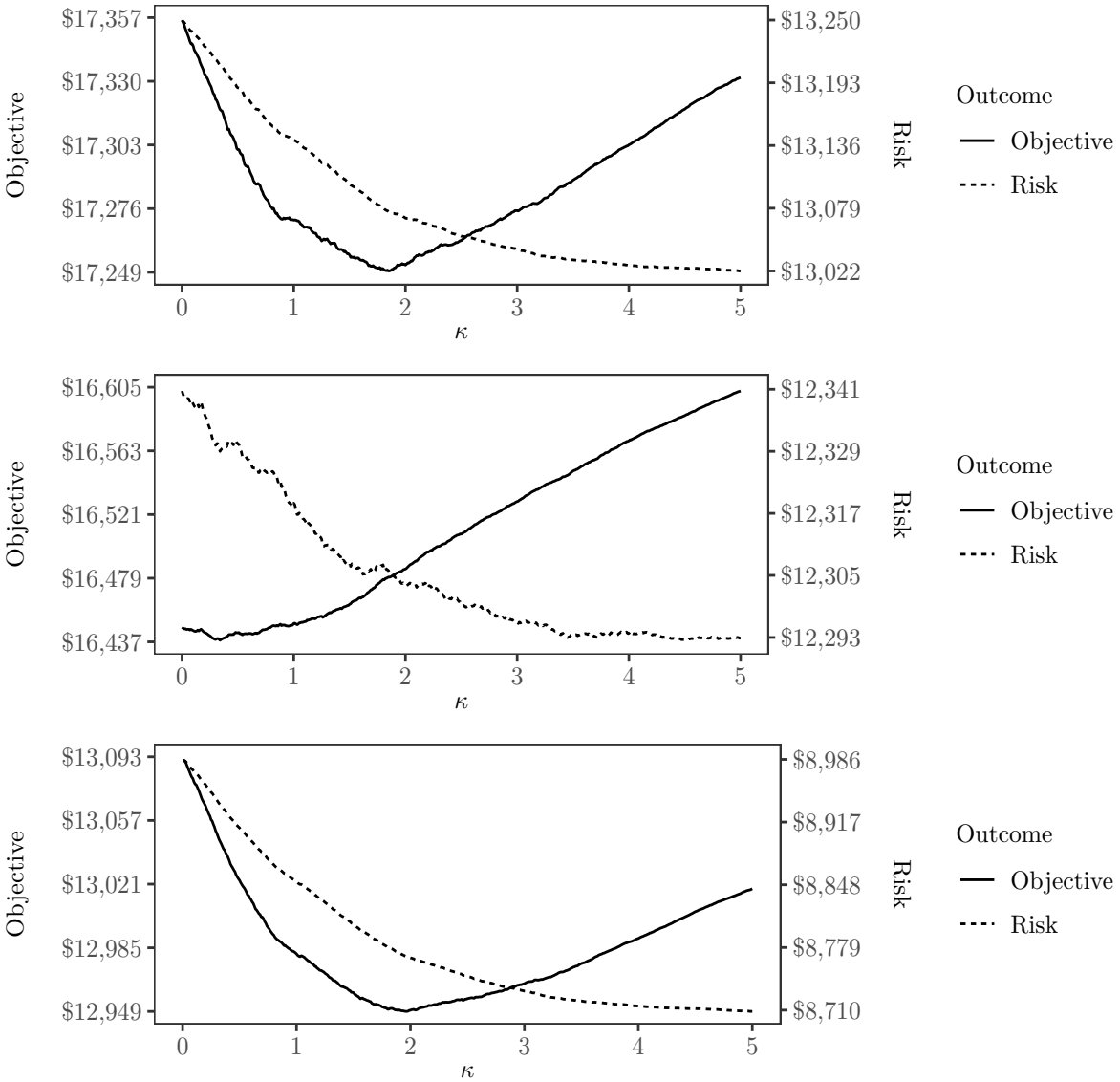
In the following three experiments, we investigate the influence of the variance of post-programme earnings on the results of our method. Specifically, we investigate how altering σ_1 and σ_0 in the specification

19 These directions translate (azimuth:inclination) to $35^\circ : 85^\circ$ and $23^\circ : 87^\circ$, respectively, as compared to $\mu^* = \{+0.883, +0.442, +0.158\}$ or $27^\circ : 81^\circ$ in the original sample.

FIGURE D.13

(Experiments 1–3) Behaviour of the objective function at μ^* given variation in κ

Experiment	μ^*	κ^*
1. Baseline (top panel)	$\{+0.822, +0.565, +0.078\}$	1.850
2. High variance (middle panel)	$\{+0.641, +0.703, -0.309\}$	0.340
3. Low variance (bottom panel)	$\{+0.831, +0.547, +0.105\}$	1.970



Simulated data generated using Equation D.82 in conjunction with data from the JTPA Study sample. This figure illustrates how the shape of the objective function and its risk component changes as the variance of post-programme earnings is increased and decreased. We hold μ fixed at μ^* in each case, which varies according to the sample size, and vary κ. Parameter estimates are presented above.

outlined in Equation D.76 affects our results. In particular, we are interested in whether there is a fundamental change in how empirical welfare risk varies with the parameters of the von Mises-Fisher distribution.

The JTPA Study sample consists of 9,223 observations, and we extract treatment status and individual characteristics directly from this dataset (we refer to the group of individuals in the sample who are treated as the experimental group and to the group of individuals who are not as the control group in line with the terminology surrounding randomised control trials). We then generate potential outcomes according to Equation D.76. We implement the affine map of Equation D.75 via the transformations

$$\begin{aligned} X_{\text{earn}} &\mapsto X_{\text{earn}} / \max\left(X_{\text{earn}}\right) \\ X_{\text{educ}} &\mapsto X_{\text{educ}} / \max\left(X_{\text{educ}}\right) \end{aligned} \tag{D.81}$$

We then regress post-programme earnings in the experimental group and the control group on these characteristics, separately and together, so that the baseline experiment somewhat mimics the JTPA Study sample. Using our simple regressions as a rough guide, we let

$$\begin{aligned} \alpha_1 &= \{+3,040; +86,446; +14,008\} \\ \alpha_0 &= \{-1,086; +82,458; +18,804\} \end{aligned} \quad \text{with } \sigma_1 = \sigma_0 = 15,914 \text{ and } c_v = 5\sigma_1 = 5\sigma_0 \tag{D.82}$$

so that the optimal policy is

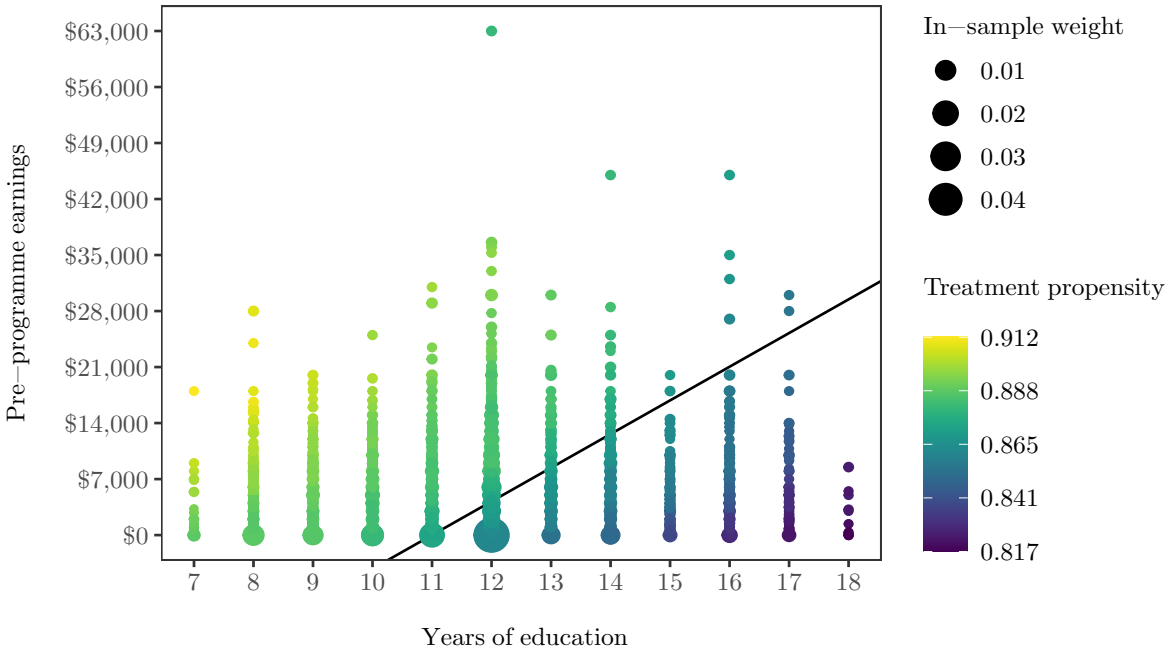
$$g\left(\boldsymbol{X}\right) = 1\left(\boldsymbol{X}^\top \{+0.552, +0.533, -0.641\} \geq 0\right) \tag{D.83}$$

which we plot in Figure D.12. We note that the coefficients of the linear specification are so large because our simple regressions scale individual characteristics but do not scale post-programme earnings, which inflates the effect of pre-programme earnings and years of education. This constitutes our first experiment.

We then propose two further experiments. Our second experiment increases the variance of post-programme earnings by inflating σ_1, σ_0 and c_v in Equation D.82 by a factor of five. Our third experiment similarly reduces the variance of post-programme earnings by deflating σ_1, σ_0 and c_v in Equation D.82 by a factor of five. We note that both experiments leave the oracle policy unchanged from Equation D.83. We present the corresponding estimates of the parameters of the posterior distribution in Figure D.13.

An immediate conclusion that we can draw from these results is that μ^* does not align with the oracle

FIGURE D.14
(Experiment 4) Variation in treatment propensity across individuals



Simulated data generated using Equation D.84 in conjunction with data from the JTPA Study sample. This figure illustrates the treatment propensity of individuals in a simulated sample under the posterior assignment rule that is induced by $\{\kappa^*, \mu^*\}$. Each point represents the individual characteristics of an individual or several individuals. For comparison, individuals to the left of the solid diagonal line are assigned treatment under the oracle assignment rule of Equation D.85.

assignment rule. That being said, it is apparent from Figure D.12 that individuals who are assigned treatment under the oracle assignment rule are more likely to be assigned treatment under the posterior distribution that we obtain.

We plot the behaviour of the objective function and of empirical welfare risk in Figure D.13. We emphasise the non-smoothness of empirical welfare risk for the second experiment, and we suggest that increasing the variance of post-programme earnings makes the problem of finding the optimal assignment rule more difficult.

—| SUBAPPENDIX D.3.C |—

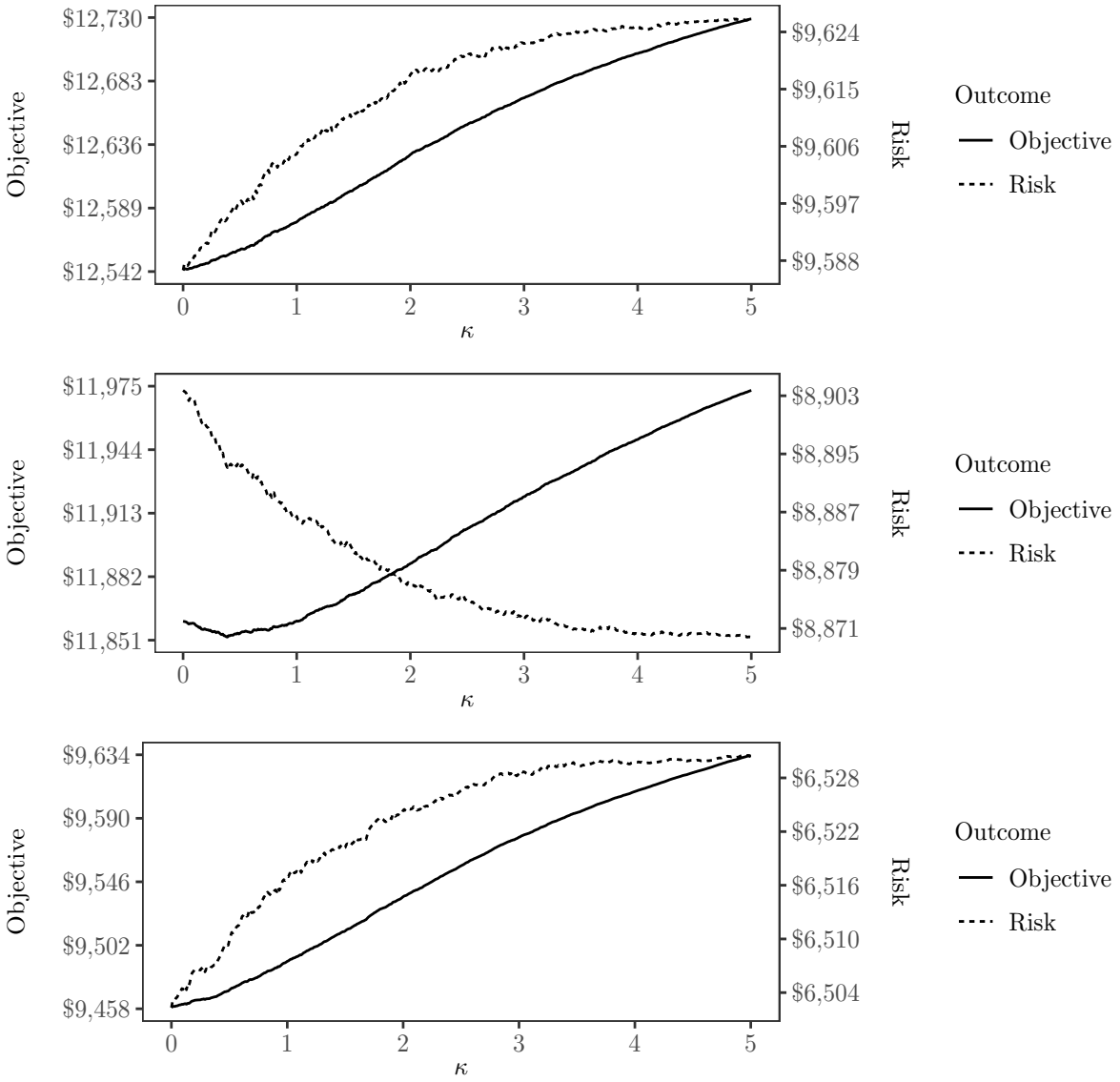
THE LOCATION OF THE ORACLE ASSIGNMENT RULE

A common feature of the data generating process outlined in Equation D.82 for Experiments 1 through 3 is the oracle assignment rule, which is located in the south-east corner of the covariate space (Figure D.12). Two things can be inferred from this. First, that only a small minority of individuals are not assigned treatment under the oracle assignment rule; specifically, those individuals who are educated to

FIGURE D.15

(Experiments 4–6) Behaviour of the objective function at μ^* given variation in κ

Experiment	μ^*	κ^*
4. Baseline (top panel)	$\{+0.932, +0.362, -0.035\}$	0.000
5. High variance (middle panel)	$\{-0.810, -0.526, -0.259\}$	0.390
6. Low variance (bottom panel)	$\{+0.899, +0.425, +0.105\}$	0.000



Simulated data generated using Equation D.84 in conjunction with data from the JTPA Study sample. This figure illustrates how the shape of the objective function and its risk component changes as the variance of post-programme earnings is increased and decreased. We hold μ fixed at μ^* in each case, which differs according to the sample size, and vary κ. Parameter estimates are presented above.

post-graduate level.[20] In other words, only a small subset of the data with extreme characteristics are not assigned treatment. Second, that, all else being equal, average outcomes in the experimental group are higher than average outcomes in the control group, and that this baseline difference is relatively important as compared to pre-programme earnings and years of education. These features, we suggest, are indicative of a relatively easy problem (of determining the optimal assignment rule).

The following three experiments largely replicate Experiments 1 through 3 but shift the oracle assignment rule to the left, narrowing the difference between the baseline average outcomes in the experimental and control groups. In Experiment 4, we assume that

$$\begin{matrix}\alpha_1 = \{+2,442; +86,446; +14,008\} \\ \alpha_0 = \{-489; +82,458; +18,804\}\end{matrix} \text{ with } \sigma_1 = \sigma_0 = 15,914 \text{ and } c_v = 5\sigma_1 = 5\sigma_0 \tag{D.84}$$

so that the optimal policy is

$$g(\boldsymbol{X}) = 1\left(\boldsymbol{X}^\top \{+0.425, +0.579, -0.696\} \geq 0\right) \tag{D.85}$$

which we plot in Figure D.14. Experiments 5 and 6 then mirror Experiments 2 and 3 in that they inflate and deflate, respectively, the variance of post-programme earnings. We suggest that Equation D.84 is a more difficult problem than Equation D.82.

We specifically design these experiments so that the oracle assignment rule is such that as close to 50% of the sample is assigned treatment as is possible without altering the importance of pre-programme earnings relative to years of education in the earnings process. We present the corresponding estimates of the parameters of the posterior distribution immediately in Figure D.15.

A curious feature of these results is that, for Experiments 4 and 6, the posterior distribution is uniform. We note that the mean direction is irrelevant in this case, but we include it anyway because knowledge of it is necessary to interpret Figure D.15. This uniformity of the posterior distribution is apparently driven by how empirical welfare risk increases alongside the concentration in Figure D.15. A possible explanation for this feature is that this data generating process is indeed hard. We suggest that the narrow gap between the baseline average outcomes of the experimental and control groups makes treating everyone versus treating no-one equally appealing (or unappealing). For instance, contrasting Figure D.7 with its analogue in the main text (for the raw data), we see that the difference in empirical welfare risk between the treat everyone rules and treat no-one rules that occupy the southern and northern

20 Or, more precisely, that are educated for an equivalent amount of time as would be required to obtain a post-graduate degree, since we do not observe education level.

regions of the heatmaps, respectively, narrows. A similar effect happens here. Given the fact that the density contours of the von Mises-Fisher distribution are concentric, there is then no natural region of the sphere to locate in order to minimise empirical welfare risk unless the value of the concentration parameter is particularly large and the posterior distribution allocates substantial probability mass to an extremely localised area on the frontier between the collections of treat everyone rules and treat no-one rules. High concentration is heavily penalised by the Kullback-Leibler divergence though, and so uniformity cannot be improved upon. The fact that a different pattern is observed for Experiment 5 is compatible with this argument, and could be achieved if the high variance of post-programme earnings makes a few observations in the sample pivotal (recall that the weight attached to each observation is proportional to post-programme earnings; inflating the variance of post-programme earnings can lead to greater concentration of weight on a few observations, since extreme outliers are more likely).

—| Subappendix D.3.d |—

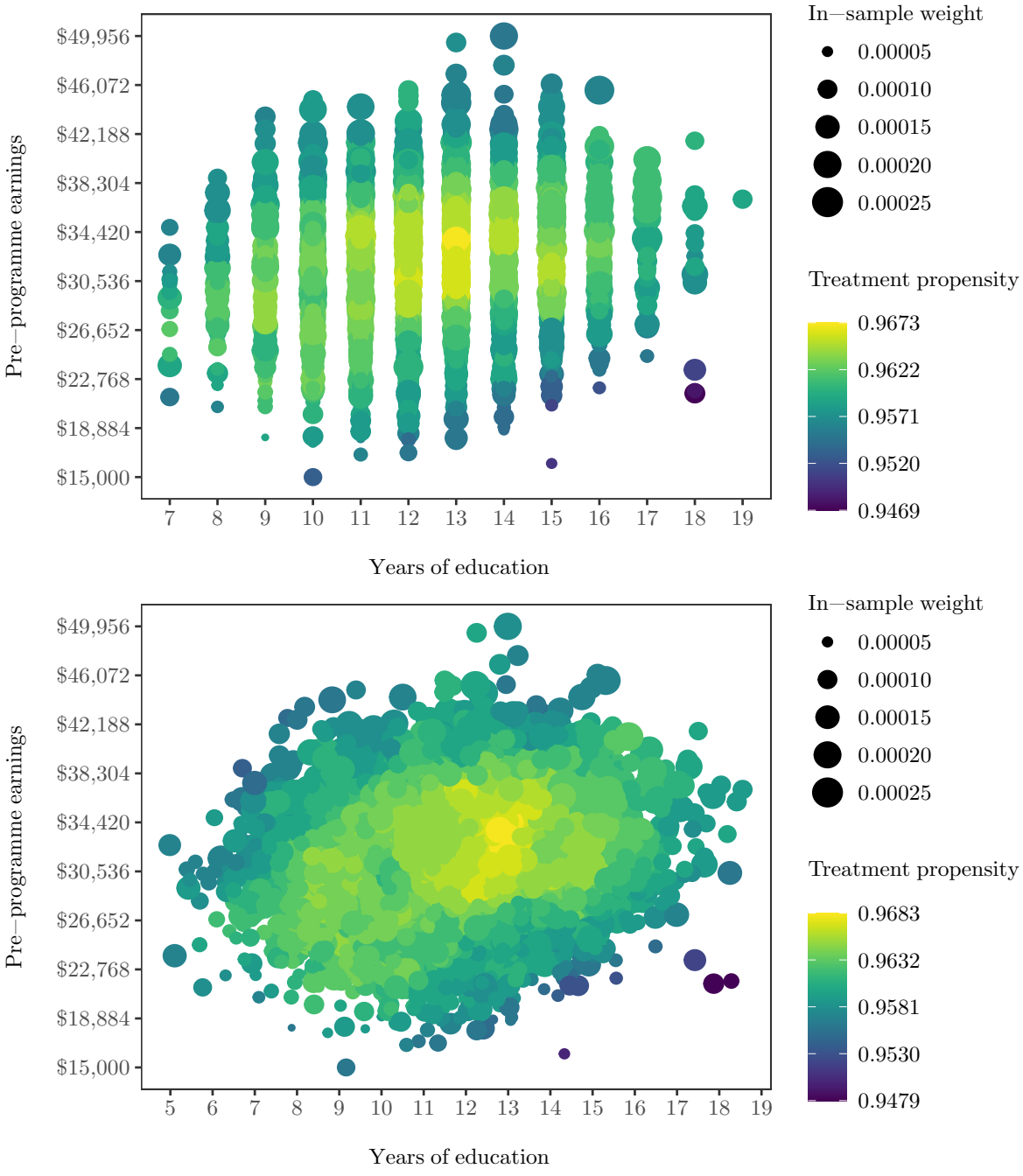
Individual characteristics and their distribution

The final experiments that we undertake vary the distribution of individual characteristics. Our intention is to understand whether the JTPA Study sample is, in some sense, special and whether altering individual characteristics substantially alters our results.

Experiments 7 and 8 mirror Experiment 1 in how the outcome is generated: both experiments rely on Equation D.76 and the parameter values that we outline in Equation D.82 for Experiment 1. The distinction between Experiments 7 and 8 and Experiment 1 is how individual characteristics are generated. Whereas Experiment 1 uses data taken from the JTPA Study sample, Experiments 7 and 8 generate individual characteristics according to a bivariate normal distribution with the mean vector equal to half of the maximum of pre-programme earnings and the average number of years of education, respectively. We estimate the covariance matrix of pre-programme earnings and years of education in the JTPA Study sample, and set the covariance of the bivariate normal distribution equal to this (Pearson correlation coefficient of 0.126). In the case of Experiment 7, we discretise years of education by assigning each observation to one of 12 equal-sized bins. We then map each characteristic to the unit interval by means of the aforementioned linear transformation.

We plot the individual characteristics that we use in Experiments 7 and 8 in Figure D.16. Absent from either plot is the oracle assignment rule. This is not an oversight. Rather, the oracle assignment rule is such that it recommends that all individuals be assigned treatment (i.e., it to the south-east of the plotting area). We plot the objective function and empirical welfare risk for Experiments 7 and 8 in Figure D.17. Due to the similarity of Figure D.17 to the other figures that we have presented, we

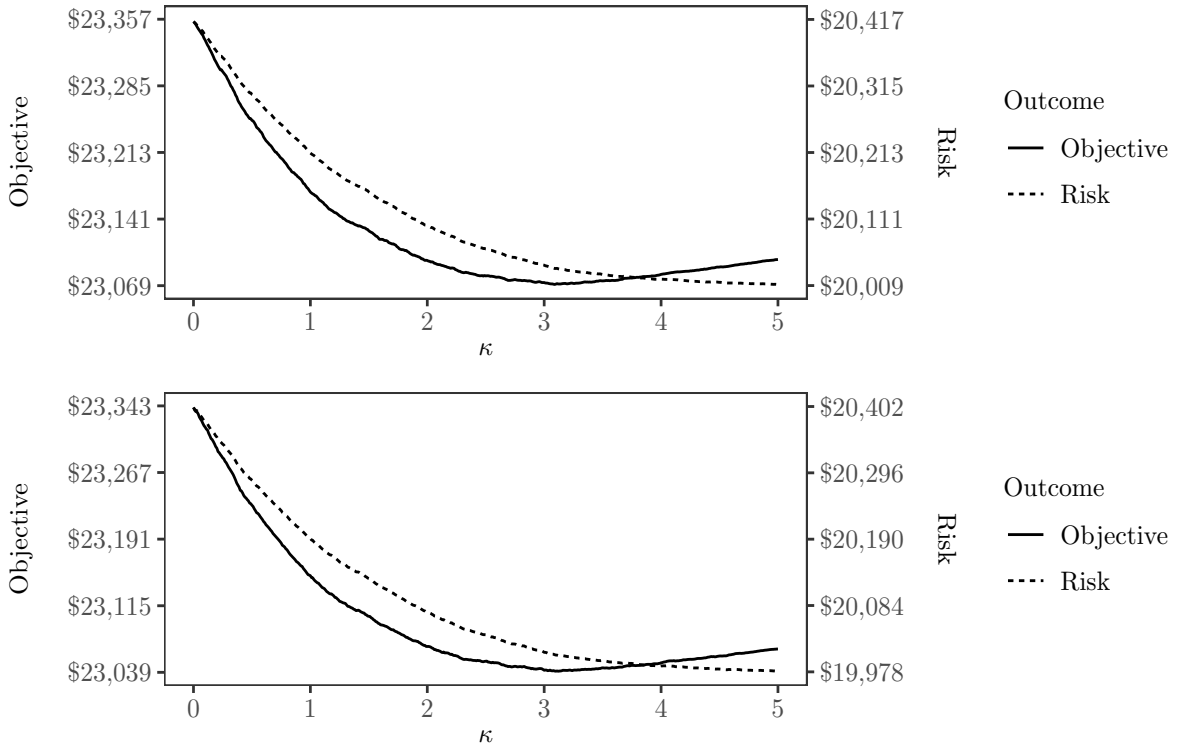
FIGURE D.16
(Experiments 7–8) Variation in treatment propensity across individuals



Simulated data generated using Equation D.82 and a bivariate normal distribution. This figure illustrates the treatment propensity of individuals in simulated samples under the posterior assignment rule that is induced by $\{\kappa^*, \mu^*\}$. Each point represents the individual characteristics of an individual or several individuals. Every individual is assigned to treatment under the oracle assignment rule.

FIGURE D.17
(Experiments 7–8) Behaviour of the objective function at μ^* given variation in κ

Experiment	μ^*	κ^*
7. Discrete (top panel)	$\{+0.794, +0.405, +0.454\}$	3.090
8. Continuous (bottom panel)	$\{+0.801, +0.408, +0.438\}$	3.120



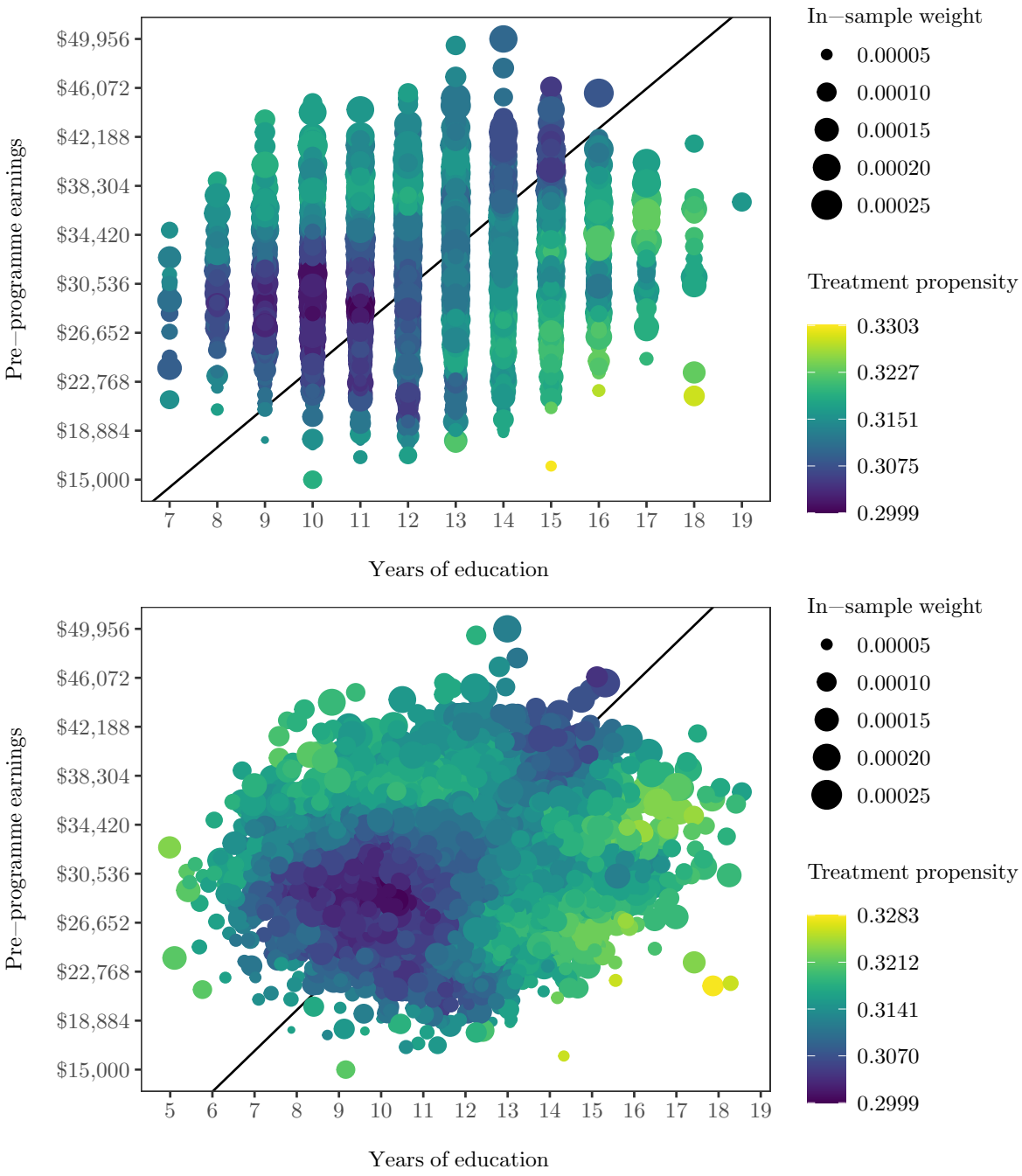
Simulated data generated using Equation D.82 and a bivariate normal distribution. This figure illustrates how the shape of the objective function and its risk component at μ^* changes as κ increases. Parameter estimates are presented above.

do not provide any further discussion of this figure. What is somewhat interesting though is how the propensity of treatment is highest for those individuals with individual characteristics located around their mean values. We present the corresponding estimates of the parameters of the posterior distribution in Figure D.17. We find that μ^* does not align with the oracle assignment rule, but does dictate that all individuals are assigned treatment (as they are under the oracle assignment rule); κ^* is also relatively large, as compared to its value in the previous experiments.

So as to make the assignment problem harder—at least, what we understand to be harder—we propose Experiments 9 and 10. These experiments follow Experiments 7 and 8 in how individual characteristics are generated, but differ slightly in how they generate the outcome. Whilst Experiments 9 and 10 broadly follow Experiments 7 and 8 with respect to how the outcome is generated, they shift the process for the potential outcomes and, thereby, the oracle assignment rule. Specifically, in Experiments 9 and

FIGURE D.18

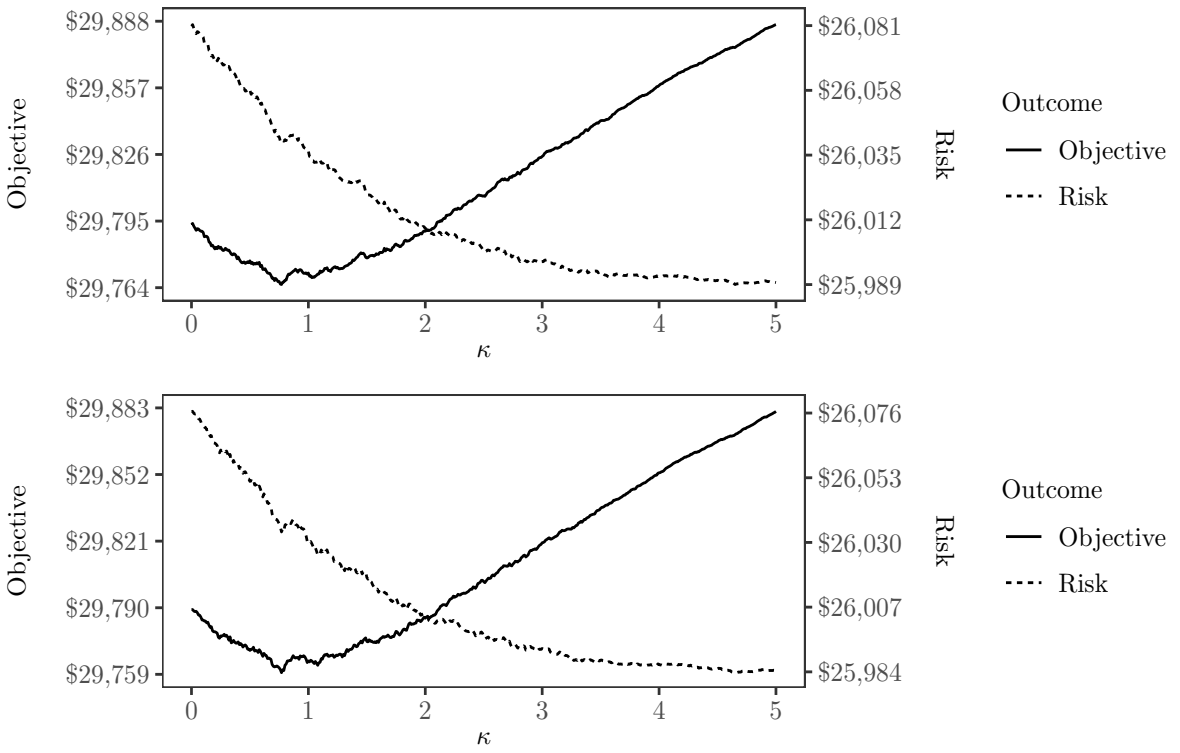
(Experiments 9–10) Variation in treatment propensity across individuals



Simulated data generated using Equation D.86 and a bivariate normal distribution. This figure illustrates the treatment propensity of individuals in simulated samples under the posterior assignment rule that is induced by $\{\kappa^*, \mu^*\}$. Each point represents the individual characteristics of an individual or several individuals. For comparison, individuals to the left of the solid diagonal line are assigned treatment under the oracle assignment rule, which partitions the sample into two approximately equal-sized groups.

FIGURE D.19
(Experiments 9–10) Behaviour of the objective function at μ^* given variation in κ

Experiment	μ^*	κ^*
9. Discrete (top panel)	$\{-0.766, -0.470, -0.438\}$	0.770
10. Continuous (bottom panel)	$\{-0.766, -0.470, -0.438\}$	0.770



Simulated data generated using Equation D.86 and a bivariate normal distribution. This figure illustrates how the shape of the objective function and its risk component at μ^* changes as κ increases. Parameter estimates are presented above.

10, we assume that

$$\begin{aligned}\alpha_1 &= \{+1,286; +86,446; +14,008\} \\ \alpha_0 &= \{-668; +82,458; +18,804\}\end{aligned} \quad \text{with } \sigma_1 = \sigma_0 = 15,914 \text{ and } c_v = 5\sigma_1 = 5\sigma_0 \tag{D.86}$$

such that the oracle policy is

$$g(\boldsymbol{X}) = 1\left(\boldsymbol{X}^\top \{+0.098, +0.636, -0.765\} \geq 0\right) \tag{D.87}$$

The effect of this change is to maintain the slope of the oracle assignment rule (i.e., the contribution of pre-programme earnings relative to years of education) but to alter the intercept. In other words, to narrow the difference in baseline average outcomes between the experimental and control groups. In this regard, Experiments 9 and 10 are similar in intent to Experiment 4.

We plot the individual characteristics that we use in Experiments 9 and 10 in Figure D.18. The oracle assignment rule is designed to partition the sample into approximately two groups of equal size. We plot the objective function (left-hand panel) and empirical welfare risk (right-hand panel) for Experiments 9 and 10 in Figure D.19. We draw attention to several differences between the results that we obtain for Experiments 9 and 10 versus Experiments 7 and 8. First, we highlight the fall in the average propensity of treatment, as is apparent in Figure D.18. Whereas in Experiments 7 and 8 (and the preceding experiments too), the propensity of treatment is close to one (i.e., everyone is assigned treatment), here the propensity is closer to one third. There is more randomisation. Second, individuals who have a relatively high propensity of treatment as compared to their peers in Experiments 7 and 8 have a relatively low propensity of treatment in Experiments 9 and 10 (visually, the colours are inverted), as is apparent in Figure D.18. That is, individuals whose characteristics are close to the mean values have a low propensity of treatment. Third, the value of the concentration parameter decreases from above three to below one (i.e., the posterior distribution is less concentrated and more uniform), as is apparent in Figure D.19. We present the corresponding estimates of the parameters of the posterior distribution in Figure D.19. Interestingly, the mean direction is, in both cases, almost the exact opposite of the corresponding mean direction in Experiments 7 and 8 (i.e., the mean direction is approximately the negative of the mean direction in Experiments 9 and 10).

—| Chapter E |—

The effect of additional children on maternal labour supply

The principal objective of this chapter is to estimate the effect of additional children on maternal employment under weak restrictions on the assumed data generating process. I frame this question in the context of the U.S. labour market, exclusively studying mothers with at least two children. I exploit natural variation in the characteristics (specifically, the birthdate and sex) of existing children to identify the employment response of mothers to additional children. The model that I employ is a minimally restrictive single equation model that incorporates an instrumental variable restriction; the model features a dichotomous outcome and a dichotomous treatment, and is partially identifying.

Understanding the sign and magnitude of the employment response of mothers to additional children is important for the design and evaluation of policy. Although family size is often, in itself, not a stated objective of policy and is not necessarily (directly) manipulable in any case, the effect of additional children on maternal employment is highly relevant to policymakers for its influence on present and future aggregate outcomes. Do labour supply responses amongst working-age mothers exacerbate or offset the long-term effect of falling fertility on public finances? Children are an important feature of the tax code and tax policy is often designed with families with children in mind—recent budget announcements indicate an intention by government to provide or increase childcare provision with the intention of improving the incentive to work amongst parents.[1] Yet, do children increase or reduce maternal employment? Understanding how children affect maternal labour supply is necessary to understand both how the tax and benefit schedule affects the household budget constraint, and whether these policies are the most suitable for achieving their stated aims. Whilst low maternal employment relative to non-maternal employment (see OECD, 2007) is a concern of policymakers, it is unclear whether this effect is driven by

1 The budget for the 2018 fiscal year included the provision of tax relief to help families struggling with child and dependent care expenses, on top of the existing Earned Income Credit that lowers marginal tax rates for working-age mothers with children in low-income groups. Since then, the American Rescue Plan and the President's (proposed) Budget for the 2024 fiscal year include an expansion of child tax credits and of access to affordable childcare. Similarly, in the U.K., the 2017 Spring Budget included the provision of financial support to parents for each child under the age of 12 years and of free childcare allowance of up to 30 hours per week to working parents with a three or four year-old child, with the express intention of providing support for ordinary working families, and for women in the workplace. The 2023 Spring Budget extended the childcare allowance to children aged between nine months and two years of age.

children or by the pre-existing opportunities of mothers for employment relative to non-mothers.

Existing empirical evidence as to the effect of children on maternal employment is mixed, despite apparent similarities in sample construction—restricting attention to married mothers with at least two children so as to exploit variation in family size that is associated with the birthdate and sex of these children—across works.

Angrist and Evans (1998), Iacovou (2001), and Al-Khaja (2016) adopt a reduced form perspective, recovering a Local Average Treatment Effect (Imbens and Angrist, 1994) that is the effect of children on maternal employment amongst an unidentifiable subset of the sample population. Importantly, this subset is not invariant and is sensistive to the institutions and cultural norms that are present, and can lead to estimates that are not necessarily representative of the effect of children across a population. For instance, Angrist and Evans (1998) finds that additional children reduce the probability that a mother is employed in the context of the U.S. labour market, whereas Iacovou (2001) and Al-Khaja (2016) find the opposite in the context of the British and Egyptian labour markets, respectively. It is unclear whether these reported differences represent an actual difference in maternal labour supply across the three countries, or whether they are due to differences in preferences over family composition and who is affected by the plurality or sex of a child.

I, instead, adopt a structural perspective, and focus on identifying and estimating an Average Treatment Effect—and its primitives, the Average Structual Functions (Blundell and Powell, 2004)—that can inform both *ex-ante* policy design and *ex-post* policy evaluation.[2,3] A similar approach is taken by Chesher and Rosen (2020)—that subsumes Chesher and Rosen (2013)—with the stated aim of illustrating how to apply minimally restrictive models to data and to demonstrate what such models deliver.

The model that I study generalises the non-parametric model of treatment non-compliance that is studied by Balke and Pearl (1997; hereafter, the Balke–Pearl model). The Balke–Pearl model permits any form of dependence between the employment decision and the fertility decision in the context of maternal labour supply, and constitutes a useful benchmark. Put differently, the Balke–Pearl model does not preclude the existence of any response types. This property contrasts favourably with control function approaches (see Wooldridge, 2015 for a discussion of control functions) and special regressor methods (see Racine et al., 2014 §I.2 for a discussion of special regressors) that either restrict these unobservable components of the economic environment or rely upon continuous variation that may not be present.

2 The Average Treatment Effect is sometimes called the Average Causal Effect.

3 The Average Structural Functions are the means of the so-called potential outcome distributions introduced in D. B. Rubin (1974)—and arguably preceding that in Splawa-Neyman (1923).

The inherent freedom of the Balke–Pearl framework does, however, mean that multiple behaviours—the measures of each of which combine to constitute the object of interest—are associated with a single outcome, which is the issue at the core of the anatomy of the selection problem (Manski, 1989). The Balke–Pearl model only partially identifies the Average Treatment Effect and its primitives, and this property is shared by the model that I study.

I generalise the Balke–Pearl model in two ways. First, I allow covariates to directly influence both response and treatment. For example, I allow mothers of different ethnicities to hold different preferences over how many children to have and whether to work. Second, I allow the instrument to have discrete—rather than dichotomous—support. For example, I allow for the birthdate and sex of children to be used in combination. The model—or, more precisely, models—that Chesher and Rosen (2020) studies also embeds these features, but maintains a monotonicity restriction that I do not. By relaxing the assumption of weak separability (or monotonic response) that is made in that paper,[4] I allow for a richer set of behaviours. In particular, I allow some mothers to move into employment given an increase in family size, and for others to move out-of employment given a similar increase. Such behaviour is ruled-out by monotonic response, but is an important feature of the model of Gronau (1977); there, differences in the endowments of and prices that are faced by families generate different employment responses at the extensive margin. To accommodate both generalisations, I extend the statistical independence restriction that the Balke–Pearl model embeds.

Aside from relaxing the assumption of weak separability, I also extend Chesher and Rosen (2020) by providing more recent evidence of the effect of additional children on maternal employment. Whereas Chesher and Rosen (2020) exclusively uses data from 1980—the original sample considered by Angrist and Evans (1998)—I additionally use data from 2010 and 2015, and so can provide some insight into whether this effect has changed over time. I do, however, differ from Chesher and Rosen (2020) in which covariates I include, focusing on age and ethnicity or race rather than education level.

An interesting property of the Balke-Pearl model is the lack of information content that the treatment equation provides; the response equation and the statistical independence restriction constitute a sufficient representation of the model. In this sense, both the Balke-Pearl model and the model that is studied in Chesher and Rosen (2020) can be expressed as single-equation models, with the only difference between them being the assumption of weak separability that is present in the analysis of Chesher and Rosen (2013). Monotonicity is a powerful restriction, in that it effectively guarantees identification of a non-zero effect and, in many cases, the sign of a treatment effect. Mourifié (2015) also studies a tri-

4 I reiterate that the stated aim of Chesher and Rosen (2020) is to illustrate how to apply minimally restrictive models to data and to demonstrate what such models deliver. Imposing weak separability—irrespective of its merits—is compatible with this objective, since it leads to a simpler and more intuitive identification analysis.

angular model (i.e., a model featuring both a response equation and a treatment equation, as laid-out in Strotz and Wold, 1960), and imposes weak separability on both of these equations. Mourifié (2015) states sharp bounds on the Average Treatment Effect and its primitives that are robust to the failure of the support condition that is required in Shaikh and Vytlacil (2011). Unlike for the Balke-Pearl model, however, the treatment equation is not redundant and has information content.

Richardson and Robins (2014) extends Balke and Pearl (1997) in allowing instruments to be discrete rather than dichotomous, and is the closest paper to this work. Richardson and Robins (2014) also varies the strength of the statistical independence restriction therein, with the presented bounds coinciding with those that I obtain when covariates are assumed to have a single point of support. Kitagawa (2021) extends Balke and Pearl (1997) in allowing for continuous response and varies the strength of the statistical independence restriction therein. The model that is studied in Kitagawa (2021) is, like the Balke-Pearl model that it nests, partially identifying and is falsifiable via the generalised instrumental inequalities of Kédagni and Mourifié (2020). Chesher and Rosen (2017) provides a general characterisation of sharp bounds on the identified set for a large of class of models and a wide choice of criteria using results from random set theory, with the analysis therein able to replicate the findings of Balke and Pearl (1997) and Kitagawa (2021). Although this characterisation equally applies when treatment is discrete or continuous, Gunsilius (2020) advances a path-sampling approach as a more tractable alternative for triangular models featuring large support.

A common issue for the application of the aforementioned models and their associated methods is that naïve estimation of the Average Treatment Effect and other partially identified objects is typically biased. I adopt a frequentist perspective and a method described in Chernozhukov et al. (2013) that is half-median unbiased. Chernozhukov et al. (2013) is but one paper that discusses inference on moment inequalities—other papers in the literature include Andrews and Soares (2010), and Bugni (2010), and references therein. One paper of particular note is Kaido et al. (2019), which proposes a bootstrap procedure to correct for projection conservatism. Projection conservatism—that confidence regions do not have the correct size, and contain the truth in more than the pre-specified proportion of samples—is of concern in partially identifying models where structural characteristics of interest are not identified directly but as the projection of an identified set of several parameters. For instance, the Average Treatment Effect is identified as a projection of the Average Structural Functions (their difference).

An alternative approach is to undertake Robust Bayesian inference, which is the subject of Giacomini and Kitagawa (2021b) and Kline and Tamer (2016). The principal distinction between robust Bayesian inference and Bayesian inference is that the former specifies multiple priors. Moon and

Schorfheide (2012) shows that Bayesian credible regions differ substantially from frequentist confidence regions, even asymptotically, in partially identifying settings. Robust Bayesian credible regions though do not share this property; Kitagawa (2021) shows that robust Bayesian inference using the lower envelope (the posterior lower probability) has a frequentist interpretation under certain conditions.[5] Attractive features of robust Bayesian inference are that they do not suffer from projection conservatism and are often less expensive than frequentist approaches that rely on grid-based inversion of a test-statistic. I do not adopt a Bayesian perspective (robust or otherwise) in this chapter.

—| Section 1 |—
Framework

I suppose that the social planner (or the econometrician acting on her behalf) observes experimental or observational data that comprises a probability distribution over $\{Y, T, Z, X\}$ that I denote by P^n and that is constructed from n independent and identically distributed draws. Here, $Y \in \{0, 1\}$ denotes response, which is some measured binary outcome of interest; $T \in \{0, 1\}$ denotes treatment, which is an indicator for some measured binary choice; $Z \in \mathbb{R}$ denotes an instrument, which is an indicator for the occurence of some event; and $X \in \mathbb{R}$ denotes covariates, which are some measured individual characteristics. The population from which the experimental or observational data is drawn comprises a probability distribution over $\{U, Z, X\}$ that I denote by P. Here, U is unobserved heterogeneity that relates to Y and T via the relationship

$$\begin{aligned} y &= 1\left(s_y\left(t, z, x, u\right) \geq 0\right) \\ t &= 1\left(s_t\left(y, z, x, u\right) \geq 0\right) \end{aligned} \tag{E.1}$$

I refer to P^n as the empirical distribution and to P (together with Equation E.1) as the data-generating process.

Throughout my analysis, I maintain several assumptions that I collectively refer to as the generalised Balke-Pearl model for their resemblence to the assumptions that the eponymic model embeds.

Assumption E.1 (Exclusion)**.** *The data generating process satisfies*

$$\begin{aligned} s_y &= h_y\left(t, x, u\right) \\ s_t &= h_t\left(z, x, u\right) \end{aligned} \tag{E.2}$$

such that the generalised Balke-Pearl model is triangular in Y, T and Z.

5 Kitagawa (2021) notes that at least one of these conditions is fragile; equality of bounds, such as when an instrumental variable has no influence over treatment, invalidates this frequentist interpretation.

The Balke-Pearl model is a structural equation model, such that each structural equation specifies a causal relationship. Given its recursive nature, the Balke-Pearl model induces a probability distribution that is Markovian; together with the statistical independence restrictions that it embeds, I note that the Balke-Pearl model can instead be presented as a (Pearlian) Directed Acyclic Graph (DAG; see Pearl, 2009 for a comprehensive treatment of the equivalence of structural equation models and DAGs, and Dawid, 2008 for a discussion of the difference between Pearlian DAGs and probabilistic DAGs). The generalised Balke-Pearl model also has this dual representation.

Assumption E.2 (Discrete support)**.** *The data generating process satisfies* $\mathfrak{z} < \infty$ *and* $\mathfrak{x} < \infty$*, with* $\mathfrak{z} \leftmapsto x \geq 2$ *holding over* $\mathfrak{X}$*.*

Assumption E.3 (Statistical independence)**.** *The data generating process satisfies*

$$U|\{Z, X\} \sim \text{Uniform}(0, 1) \tag{E.3}$$

for which h *is appropriately normalised.*

The generalised Balke-Pearl model exhibits a number of features that make it ideally suited to the study of the effect of additional children on female employment.

First, the generalised Balke-Pearl model permits endogenous selection of T. That is, the model allows for dependence between the employment and fertility decisions of mothers. In particular, it is likely that mothers factor their opportunities for employment into their decision about whether to have additional children, and *vice versa*. For example, mothers who face poor employment opportunities (and a low opportunity cost of pregnancy) relative to others may choose to have an additional child. This story hints at mothers with three or more children also being mothers who are in any case less likely to be employed; whilst this may or may not be true, the existence of some indirect mechanism that generates correlation between response and treatment is widely accepted.

Second, the generalised Balke-Pearl model permits rich heterogeneity of response both across and within sub-populations labelled by X. Heterogeneity of response across sub-populations can be generated by parametric models and so does not warrant much discussion.[6,7] Of greater interest, however, and warranting more-detailed discussion, is heterogeneity of response to treatment within sub-populations. The

6 As a simple example, the generalised Balke-Pearl model can allow Hispanic and non-Hispanic mothers (if ethinicity, or race, is an observable individual characteristic) to have different employment rates and to exhibit different effects of additional children on maternal employment (as an extreme case, if the effect of additional children on maternal employment amongst Hispanic mothers has the opposite sign to that for non-Hispanic mothers, say).

7 The non-parametric specification does, however, mean that X can always be written as a scalar without loss of generality, a restatement that is with loss in a parametric specification.

flexible nature of both the response equation and the treatment equation means that the generalised Balke-Pearl model can allow different mothers to experience different effects of additional children. Broadly speaking, a mother may increase her labour supply in response to additional children, may decrease her labour supply, or may leave her labour supply unchanged. The generalised Balke-Pearl model can generate all three forms of response. In particular, I emphasise that response can be non-monotonic, in that some mothers may increase their labour supply whilst others decrease their labour supply. That mothers may exhibit opposite responses at the extensive margin is due to non-separability of the response equation in unobserved heterogeneity.[8] Such behaviour is ruled-out by monotonic treatment response, but is an important feature of the model of Gronau (1977); there, differences in the endowments of and prices that are faced by families generate different employment responses at the extensive margin. For example, it is intuitive to think of children as both affecting the cost of employment (the need to provide childcare) and the value of employment (children increase the marginal utility of income); if the first effect dominates the second then it is feasible that a mother increases her labour supply in response to additional children, but the opposite will occur if the reverse is true.

Third, the generalised Balke-Pearl model features a lack of restrictions on both the dimension and the distribution of unobserved heterogeneity. Underpinning this feature of the generalised Balke-Pearl model is the flexibility of the structural equations, which I discuss in Section 2; specifically, that unobserved heterogeneity may be partitioned into a finite number of equivalence classes. Generally, restrictions on unobserved heterogeneity are not verifiable, and may impose non-trivial constraints on the economic environment—particularly in parametric models. These constraints may include restrictions on the distributions of endowments and of prices. The credibility of the generalised Balke-Pearl model is that it imposes no constraint of the shape of these distributions, and can allow for a large class of features of the economic environment to influence the employment decision and the fertility decision.

—| Subsection 1.a |—

The meaning and credibility of the restrictions

Although the generalised Balke-Pearl model is minimally restrictive, it is not entirely non-restrictive. In particular, the model embeds several explicit restrictions—Assumptions E.1 to E.3—on the data generating process and a further restriction that is implicit—a rank condition. I discuss each of these assumptions in turn, suggesting an interpretation for each and scenarios that would lead to their violation.

The exclusion restriction—Assumption E.1—can be restated as $Y \perp\!\!\!\perp Z | \{T, X, U\}$, so precluding the

8 Vytlacil (2002) shows that monotonicity is equivalent to weak separability, which precludes one of these responses.

instrument from directly influencing response (contigent upon its antecedents, response is deterministic). In the context of maternal employment, the exclusion restriction requires that having twins or higher plurality siblings does not directly influence the employment decision. Violation of the restriction occurs if twins incur additional costs versus non-plural siblings; for example, if having children of the same age requires duplication, versus handing-down or re-using. This reasoning similarly applies if the sex of existing children is used in place of plurality of birth; for example, if female and male children participate in different activities at distinct venues, or if they attend different schools. The argument in both scenarios is that the instrument—whether a plural second birth or the sex of existing children—influences the employment decision via its effect on the budget constraint, and not simply through its influence on the decision of whether to have additional children.

The rank condition can be stated as $T \not\perp\!\!\!\perp Z$, so that the treatment equation is a non-trivial function of the instrument. In the context of maternal employment, a plural second birth means that a mother has three or more children, and so automatically satisfies this requirement by definition. If the sex of existing children is used in place of plurality of birth, violation of the restriction occurs if parents either do not hold a preference over the sexes of their children or if parents are equally balanced over all possible combinations (i.e., as many parents prefer mixtures of female and male children as prefer just female or just male).

The discrete support restriction—Assumption E.2—is a statement about both the amount of variation in the instrument and the sort of covariates that can be incorporated. The discrete support restriction does not, however, otherwise impose any constraint on the marginal distribution of the instrument and the covariates—pairs of realisations can exhibit dependence or not occur at all. As a restriction on the observable exogenous economic variates, the discrete support restriction is verifiable (i.e., it can be tested and either accepted or rejected). In the context of maternal employment, the discrete support restriction requires that there are at least some mothers for every combination of age, age at first birth, and ethnicity and race that experience a plural second birth. Violation of the restriction occurs if a plural second birth is too rare an event; for example, if the aforementioned characteristics are so finely granulated that there are fewer plural second births than there are groups. If the sex of existing children is used in place of plurality of birth, violation of the restriction is unlikely due to the natural sex ratio (i.e., it is nearly—although not equally—as likely that a mother has a female child as a male child).

The statistical independence restriction—Assumption E.3—can be restated as $U \perp\!\!\!\perp \{Z, X\}$, since the distributional assumption (i.e., that unobserved heterogeneity is uniformly distributed) is simply a normalisation accorded by the non-parametric nature of the structural equations. In the context of maternal employment, the statistical independence restriction implies that mothers do not base their age, age at

first birth, and ethnicity and race on their opportunities for employment, nor is the probability of a plural second birth dependent upon these opportunities. Given the immutable and predetermined nature of these characteristics, there is strong reason to suppose that this restriction is satisfied in this regard. Of greater concern is whether any of these economic variates are statistically associated with a further variable that influences Y but that is not included explicitly in the model. For example, the statistical independence restriction is violated if education is statistically associated with age at first birth or ethnicity and race, since education likely influences the probability of employment (through spousal income if more highly educated mothers marry equally educated men, and through the offered wage). This reasoning similarly applies if the sex of existing children is used in place of plurality of birth.

Together, these restrictions define Z as an instrumental variable (for the effect of treatment on response).[9] I note that the statistical independence restriction may be weakened; specifically, the statistical independence restriction may be weakened by partitioning U into a part that enters the response equation and a part that enters the treatment equation. A sufficient restriction is then that Z is statistically independent of that part of U that enters the response equation, although it can be statistically associated with the part that enters the treatment equation. Kitagawa (2021) further weakens this restriction to marginal statistical independence, but it is unclear what the interpretation of this restriction is in a structural equation model.

—| Section 2 |—

Equivalence classes and latent types

The generalised Balke-Pearl model is a structural equation model, such that each structural equation specifies a causal relationship. The response equation and the treatment equation then fully describe behaviour; in particular, the outcome equation describes maternal employment for both the realised treatment and the unrealised treatment. The potential outcome notation that is developed in D. B. Rubin (1974) advances this idea by defining both the realised outcome and the unrealised outcome as variates, and is complementary to the structural approach that I take in this chapter. I adopt the potential outcome notation briefly so as to illustrate an important point that is made in Pearl (2009).

$$\begin{aligned} y_t(x) &\doteq 1\left(h_y(t,x,u) \geq 0\right) \\ t_z(x) &\doteq 1\left(h_t(z,x,u) \geq 0\right) \end{aligned} \qquad \text{E.4}$$

9 This definition of an instrumental variable is consistent with the definition that is stated in Pearl (2009).

From the definition of the potential outcomes, it is evident that the potential outcomes share the support of the response and of treatment—the potential outcomes are dichotomous random variables. Moreover, all variation in the potential outcomes is due to variation in U, since each potential outcome designates t, z and x. It immediately follows that unobserved heterogeneity may be collected into sets that deliver common values of the potential outcomes—equivalence classes.

An equivalence class is a set in u-space such that any two members of a given class induce the same potential outcomes as each other for every possible designation of the observable exogenous random variates. I note that this definition of an equivalence class does not rely on a support condition (for unobserved heterogeneity)—the definition is valid regardless of the dimension and distribution of unobserved heterogeneity. This observation is consistent with the claim that is made in Balke and Pearl (1997) that the Balke-Pearl model can allow for unobserved heterogeneity of arbitrary dimension. Each equivalence class collects values of unobserved heterogeneity that deliver common values of the potential outcomes. Given that each potential outcome may take a finite number of values and there are a finite number of potential outcomes, unobserved heterogeneity may be collected into a finite number of equivalence classes (Pearl, 2009). The number of equivalence classes increases exponentially in $\mathfrak{z}$ and $\mathfrak{x}$.

It is convenient to think of each equivalence class as a latent type (Pearl, 2009), in that each class describes a common set of behaviours. By appropriately defining the structural equations as index functions, the Balke-Pearl model and its generalisation can be restated in terms of response types. Doing so establishes an equivalence between the potential outcome framework and the structural equation framework that is often disregarded by proponents of each approach. Moreover, these latent types can always be projected onto the unit continuum using a probability transformation, thereby confirmation the normalisation that I impose.

The appropriate interpretation of the associated statistical independence restriction that arises in this potential outcome–*cum*–structural equation framework is that latent types are statistically independent of covariates and the instrument. In the context of maternal employment, this restriction is effectively equivalent to the statement that mothers are *ex-ante* identical up to a finite number of latent types and that covariates and the instrument are randomly allocated. Put differently, mothers of the same latent type would make the same employment and fertility decisions but make (possibly) different decisions due to *ex-post* variation in their covariates and the instrument.

—| Section 3 |—

Identification

I extend Balke and Pearl (1997) in allowing covariates to influence the economic system, as a cause of the response and of treatment. A contribution of this chapter is that I state sharp bounds on structural parameters of interest for this case, which represents a simple extension of Balke and Pearl (1997). These bounds coincide with those that are stated in Balke and Pearl (1997) for the special case in which the analysis of Balke and Pearl (1997) is applied to a sub-population x rather than the full population. Such a refinement of the population does not invalidate the analysis of Balke and Pearl (1997) due to the independence restriction that I invoke (that covariates are statistically independent of unobserved heterogeneity). However, a refinement of the population also alters the object that Balke and Pearl (1997) bounds; specifically, I state sharp bounds on the Average Treatment Effect in the sub-population x and its primitives (as opposed to the ATE and its primitives). I follow Heckman et al. (2006) in shortening the Average Treatment Effect in the sub-population x to $\text{ATE}(x)$.[10]

An intuitive interpretation of $\text{ATE}(x)$ is the change in the employment rate between two counterfactuals, which I write as

$$\text{ATE}(x) \doteq \bar{Y}_1(x) - \bar{Y}_0(x) \tag{E.5}$$

The first counterfactual is where all mothers in the sub-population x have three or more children.

$$\bar{Y}_1(x) \doteq \Pr(Y_1(x) = 1|x) \tag{E.6}$$

The second counterfactual is where all mothers in the sub-population x have two children.

$$\bar{Y}_0(x) \doteq \Pr(Y_0(x) = 1|x) \tag{E.7}$$

If the employment rate of the first counterfactual is greater than that of the second counterfactual then $\text{ATE}(x)$ is positive, such that additional children increase maternal employment.

Broadly speaking, a mother may increase her labour supply in response to additional children, may decrease her labour supply, or may leave her labour supply unchanged. This motivates a further intuitive interpretation of $\text{ATE}(x)$: as the difference in the measures of the first two types of behaviour in the sub-population x. Although these behaviours are distinct from the compliers and defiers that are introduced

10 This parameter is otherwise referred to as CATE(x) in Abrevaya et al. (2015) and is closely related to the Average Structural Functions in Blundell and Powell (2004).

in Imbens and D. B. Rubin (1997)—in that they relate to the response equation rather than the treatment equation—they, nonetheless, have a similar flavour to those two concepts. It follows that $\text{ATE}(x)$ is positive if the proportion of mothers who move into employment exceeds the measure of proportion of mothers who move out of employment, such that there is a net increase in employment. These behaviours are exactly defined by the response type. I note that this interpretation of $\text{ATE}(x)$ applies only when the outcome and treatment are dichotomous.

I let $Q_{yt.z}(x) \doteq \Pr(Y = y, T = t|x, z)$, which extends the notation in Balke and Pearl (1997). The generalised Balke-Pearl bounds are, for all $\{a, b\} \in \mathfrak{Z} \mapsto x$,

$$\left\{ \begin{array}{c} Q_{10.a}(x) \\ (Q_{10.a} + Q_{11.a} - Q_{00.b} - Q_{11.b}) \circ (x) \end{array} \right\} \leq \bar{Y}_0(x) \leq \left\{ \begin{array}{c} 1 - Q_{00.a}(x) \\ (Q_{01.a} + Q_{10.a} + Q_{10.b} + Q_{11.b}) \circ (x) \end{array} \right\} \tag{E.8}$$

and

$$\left\{ \begin{array}{c} Q_{11.a}(x) \\ (Q_{00.a} + Q_{11.a} - Q_{00.b} - Q_{01.b}) \circ (x) \end{array} \right\} \leq \bar{Y}_1(x) \leq \left\{ \begin{array}{c} 1 - Q_{01.a}(x) \\ (Q_{00.a} + Q_{11.a} + Q_{10.b} + Q_{11.b}) \circ (x) \end{array} \right\} \tag{E.9}$$

These bounds are sharp.[11]

Although the $\text{ATE}(x)$ and its primitives are informative, they may be of limited policy relevance; policy-makers are often unable to discriminate on the basis of covariates. For instance, equality legislation often prevents policy-makers from discriminating on the basis of ethnicity or race. Of greater importance to policy-makers in such an environment is ATE and its primitives. The policies that were recently announced by the U.K. government, and that are discussed earlier, are examples of policies that are applied uniformly. What is of interest to the policy-maker then is whether the effect of the policy is positive or negative on average across the population. To obtain the aggregate outcome or effect, it is sufficient to take the weighted Minkowski sum over $\text{ATE}(x)$ or its primitives (i.e., take the weighted average over the minimum of each set to obtain the aggregate lower bound, and the weighted average over the maximum of each set to obtain the aggregate upper bound).[12] Here, each weight corresponds to the relative size of the sub-population x in the population, and can be calibrated to the population that is observed or else to some other hypothetical population for *ex-ante* analysis.

11 Proof of this immediately follows from Chesher and Rosen (2017). I, nonetheless, include a discussion of the result in the appendices, and a formal derivation in Section 6.

12 The continuous equivalent of the Minkowski sum is the Aumann integral (Aumann, 1965).

—| Section 4 |—

Data and sample composition

I use data from the Integrated Public Use Microdata Series (IPUMS; Ruggles et al., 2015); the data consist of individual-level observations of households in the U.S., and were collected as part of the decennial census of 1980, and the American Community Survey (ACS) of 2010 and of 2015.

A notable difference between the earlier census and the ACS is the timing of collection—whereas the earlier census presents a snapshot of society on April 1st of that year, the ACS is collected throughout a calendar year. The ACS is less informative, in that the birth year that is imputed by the U.S. Census Bureau is incorrect for all individuals who have a birthday that falls after the date of survey. Although this error is random, and averages across the sample, an unfortunate consequence is that it is not necessarily possible to distinguish two children that are born ten or 11 months apart from twins. Furthermore, the ACS does not contain information about the number of children ever born (a measure of fertility and infant mortality), nor does it contain data on the age at or quarter of marriage.

I refine the census and ACS data according to a number of rules. First, I include only women who are aged between 21 and 35 years that are married and whose spouse is present. I exclude women who belong to a married couple where either the woman or her spouse have been married more than once. Second, I include only women who have at least two children. I exclude women who are identified as the adopted- or step-mother of any child, or whose spouse is identified as the adopted- or step-father of any child. I exclude women whose first child is 18 years old or older, or whose second child is less than one year old. Third, I exclude women whose first child was born before her 15th birthday or that of her spouse. Fourth, I exclude women whose age or sex, or that of her children, has been imputed by the U.S. Census Bureau.

I restrict attention to married women for two reasons. First, to retain comparability with Angrist and Evans (1998), which follows many of the same sample selection rules.[13] Second, by restricting attention to married women, I am able to investigate the effect of children on maternal employment for a stable sub-population. If I do not impose this focus, then it is possible that any observed change in the effect of children on maternal employment over time is not due to a change in the employment response of mothers, but due to changes in household composition.

I report summary statistics for the sample in Table E.1 and in Table E.2. Married women represent a

13 I differ from Angrist and Evans (1998) in that I do not exclude women whose first child was born before marriage, and I do not exclude women whose reported number of children ever born differs from their reported number of children. The ACS does not contain the information that is required for these rules.

particularly high fertility group in the population, and have exhibited a slight upward trend in fertility since 1980 that has partially offset declining fertility amongst unmarried women. The most striking statistic as evidence of this overall decline in fertility is a fall in the number of children that all women aged between 21 and 35 years report having, from 1.153 children in 1980 to 0.844 in 2015. Set against this decline in fertility is an increase in the female employment rate, both amongst married and unmarried women. The principal objective of this chapter is to discover whether this increase in female employment is (in part) due to the documented decline in fertility, or is despite it; and in the context of maternal employment, whether the observed increase in the employment rate of mothers has been dampened by an increase in family size.

I remark upon two further trends that are evident in Figure E.2 . First, that a husband's labour supply remains stable across the study period, both at the extensive margin and at the intensive margin. This trend is well-documented elsewhere, and suggests that male labour supply is inelastic. If children do influence the employment decision then it is likely that this effect accrues on the employment decision of the mother. Second, (deflated) household income increases from (1999) $50,900 in 1980 to (1999) $55,034 in 2015 amongst married couples. This change in household income is of the order of ten percentage points, and is much more subdued than historic growth rates would imply. Indeed, decomposing household income into wife's labour income and husband's labour income, it is apparent that this increase is driven purely by an increase in wife's labour income—real wage growth is otherwise stagnant. It is reasonable to suggest that this increase in wife's labour income is due to the documented increase in maternal employment.

—| Subsection 4.a |—

Variable definition

In presenting the Balke-Pearl model in the context of maternal employment, I have stated particular definitions for each of the variables. I note that I subject some of these definitions to change—I propose various definitions of covariates and of the instrument. I emphasise that I continue to define the outcome as a dichotomous indicator for whether a mother worked for pay in the previous year; and I define treatment as a dichotomous indicator for whether a mother has two children, or three or more children.

I continue to define age, age at first birth, and ethnicity and race as covariates. However, in some instances I restrict the vector of covariates that I include to be a proper subset of these covariates. Including the full set of covariates is not feasible in all cases, and leads to degeneracy of the estimator. To guarantee non-degeneracy of the estimator whilst retaining as much detail of covariates as possible,

FIGURE E.1
Fertility and labour supply measures

	Means (and standard deviations)					
	PUMS 1980		ACS 2010		ACS 2015	
Women aged 21-35						
Children ever born	1.238	(1.295)	-	-	-	-
Reported number of children	1.153	(1.236)	0.929	(1.216)	0.844	(1.186)
Has two or more children	0.374	(0.484)	0.292	(0.455)	0.263	(0.440)
Worked in previous year	0.732	(0.443)	0.773	(0.419)	0.787	(0.409)
Observations	1,380,870		263,009		271,322	
Women aged 36-50						
Children ever born	2.818	(1.921)	-	-	-	-
Reported number of children	1.725	(1.454)	1.344	(1.247)	1.386	(1.258)
Has two or more children	0.531	(0.499)	0.439	(0.496)	0.455	(0.498)
Worked in previous year	0.665	(0.472)	0.768	(0.422)	0.772	(0.419)
Observations	885,405		313,578		285,436	
Women aged 21-35 with two or more children and no adopted children						
Children ever born	2.588	(0.905)	-	-	-	-
Reported number of children	2.500	(0.797)	2.532	(0.815)	2.538	(0.828)
Has three or more children	0.359	(0.480)	0.380	(0.485)	0.380	(0.485)
Worked in previous year	0.560	(0.496)	0.668	(0.471)	0.679	(0.467)
Observations	486,736		67,073		60,216	
Married women aged 21-35 with two or more children and no adopted children						
Children ever born	2.530	(0.842)	-	-	-	-
Reported number of children	2.476	(0.770)	2.494	(0.785)	2.506	(0.811)
Has three or more children	0.349	(0.477)	0.361	(0.480)	0.361	(0.480)
Worked in previous year	0.532	(0.499)	0.631	(0.482)	0.628	(0.483)
Observations	333,334		38,958		35,036	

The table reports means of variables (that are equivalent to probabilities in some cases). Standard deviations are reported in parentheses. Children ever born is the number of biological children that a mother reports ever having, and is available only in the PUMS. Reported number of children is the number of children that a mother reports having in the household; Has two or more children and Has three or more children are defined in terms of Reported number of children. Worked in previous year indicates whether a mother worked for profit, pay, or as an unpaid family worker in the previous year. All samples exclude mothers whose age or gender have been imputed by the Census Bureau or have children whose age or gender have been imputed. The married sample consists of women who are married and have been married only once and are married to spouses that have been married only once.

I discretise both age and age at first birth, and partition ethnicity and race into several broad classes. I partition age and age at first birth into five sets: 15 to 18 years, 18 to 22 years, 22 to 25 years, 25 to 30 years, and 30 to 35 years. I design these partitions—specifically the first three partitions—to reflect the current U.S. education system, noting that individuals who are currently undertaking an education

FIGURE E.2
Fertility measures, child gender, and labour market outcomes

	Means (and standard deviations) for women aged 21-35 with two or more children								
	PUMS 1980			ACS 2010			ACS 2015		
		Married couples			Married couples			Married couples	
	All women	Wives	Husbands	All women	Wives	Husbands	All women	Wives	Husbands
Children ever born	3.561	3.553		-	-		-	-	
	(0.919)	(0.887)		-	-		-	-	
Reported number of children	2.550	2.529		2.661	2.640		2.684	2.661	
	(0.811)	(0.795)		(0.864)	(0.847)		(0.887)	(0.876)	
Has more than two children	0.399	0.388		0.472	0.463		0.478	0.466	
	(0.490)	(0.487)		(0.499)	(0.499)		(0.500)	(0.499)	
Male first-born child	0.514	0.514		0.512	0.513		0.506	0.505	
	(0.500)	(0.500)		(0.500)	(0.500)		(0.500)	(0.500)	
Male second-born child	0.512	0.511		0.509	0.507		0.512	0.511	
	(0.500)	(0.500)		(0.500)	(0.500)		(0.500)	(0.500)	
First two children male	0.265	0.265		0.263	0.263		0.260	0.259	
	(0.442)	(0.441)		(0.440)	(0.440)		(0.439)	(0.438)	
First two children female	0.240	0.239		0.242	0.242		0.242	0.242	
	(0.427)	(0.427)		(0.428)	(0.429)		(0.428)	(0.428)	
First two children share gender	0.505	0.504		0.505	0.506		0.502	0.501	
	(0.500)	(0.500)		(0.500)	(0.500)		(0.500)	(0.500)	
Plural second birth	0.009	0.008		0.013	0.012		0.013	0.012	
	(0.093)	(0.091)		(0.112)	(0.109)		(0.112)	(0.110)	
Age	30.215	30.175	32.625	31.246	31.158	33.931	31.500	31.438	34.173
	(3.497)	(3.521)	(4.540)	(3.143)	(3.187)	(4.669)	(3.006)	(3.039)	(4.606)
Age at first birth	21.226	21.414	23.857	22.101	22.348	25.118	22.415	22.678	25.416
	(3.028)	(2.991)	(3.845)	(3.726)	(3.729)	(4.812)	(3.729)	(3.714)	(4.821)
Observations	376,161	297,124		41,089	30,258		36,196	26,805	

Continued ↪

FIGURE E.2
Fertility measures, child gender, and labour market outcomes

	Means (and standard deviations) for women aged 21-35 with two or more children								
	PUMS 1980			ACS 2010			ACS 2015		
		Married couples			Married couples			Married couples	
	All women	Wives	Husbands	All women	Wives	Husbands	All women	Wives	Husbands
Worked for pay	0.545	0.535	0.974	0.634	0.628	0.945	0.626	0.619	0.958
	(0.498)	(0.499)	(0.158)	(0.482)	(0.483)	(0.228)	(0.484)	(0.486)	(0.201)
Weeks worked	17.8↔21.6	17.3↔21.0	45.5↔48.5	26.7↔29.5	26.3↔29.1	43.3↔46.4	26.7↔29.4	26.1↔28.8	45.5↔48.1
Hours per week	13.3↔15.8	12.9↔15.5	39.2↔43.7	-	-	-	-	-	-
Labour income	7,022	6,677	37,981	12,789	12,508	33,378	13,180	13,095	35,542
	(11,138)	(10,846)	(25,498)	(18,036)	(17,858)	(34,418)	(19,914)	(20,077)	(36,735)
Family income	50,903	50,900		52,723	52,730		54,803	55,034	
	(28,111)	(27,910)		(43,302)	(43,495)		(46,623)	(46,676)	
Non-wife income	43,881	44,222		39,934	40,222		41,623	41,939	
	(26,633)	(26,628)		(38,182)	(38,554)		(40,952)	(40,869)	
Observations	376,161	297,124		41,089	30,258		36,196	26,805	

The sample includes mothers aged between 21 and 35 years with at least two children, whose age or gender (or that of any children) has not been imputed by the Census Bureau. Married couples include only those couples that are married at the date of survey and have only been married to each other. Labour market outcome variables are reported for the previous year. Weeks worked and Hours per week are set-valued and so the reported means are set-valued (an upper bound of 84 hours is used for Hours worked). Incomes are adjusted to 1999 USD values using Census Bureau adjustment factors. Non-wife income is defined as Family income less the value of Labour income for a wife.

programme may postpone having children until after they have completed that programme. I define ethnicity as a dichotomous indicator for whether a mother is of Hispanic origin. I define race as a three-valued indicator for whether a mother is black, white, or other race.

I propose four definitions of the instrument. First, I define the instrument as a dichotomous indicator for a plural second birth (a mother gives birth to twins or higher plurality siblings). Second, I define the instrument as a dichotomous indicator for whether a mother's first two children share the same sex. Third, I propose separating the sex composition of a mother's first two children into its constituent pairs. That is, I define the instrument as a four-valued variable that takes different values depending upon the sex of the first child and of the second child. Fourth, I combine a plural second birth with the sex composition of a mother's first two children. I note that the influence on the fertility decision is different for a plural second birth and sex; in particular, a plural second birth acts mechanically, in that a mother that experiences a plural second birth has three or more children by definition. In contrast, sex influences the fertility decision through parental preferences. If a mother experiences a plural second birth though, parental preferences are irrelevant to the fertility decision. To reflect this asymmetry, I define the instrument as a five-valued variable with a hierarchical structure that assigns priority to a plural second birth; if a mother does not experience a plural second birth then the instrument takes either the second through fifth value, depending upon the sex composition of her first two children.

The generalised Balke-Pearl model imposes an exclusion restriction on the assumed data generating process. Such a restriction is violated if twins incur additional costs versus non-plural siblings; for example, if having children of the same age requires duplication, versus handing-down or reusing. These additional costs also include non-pecuniary costs that are especially high during early childhood, and that may extend beyond a year (recall that I exclude mothers whose second child is less than one year old) in the case of twins. Due to these early childhood considerations, I refine the sample to exclude mothers whose second child is aged under five years whenever I define the instrument as a plural second birth (either by itself or with sex composition). A plural second birth is a rare event, occurring in around 1-in-100 second births in the data that I study. Degeneracy of the estimator is of first-order concern whenever I define the instrument as a plural second birth, and is exacerbated by the additional sample refinement that I perform in this case.

FIGURE E.3
Parental choice over family size given sex of children

	Probabilities (and standard errors) of having another child					
	All women			Married women		
	PUMS 1980	ACS 2010	ACS 2015	PUMS 1980	ACS 2010	ACS 2015
Sex of first child in families with one or more children						
(1.) ♂	0.714	0.743	0.742	0.726	0.752	0.745
	(0.001)	(0.003)	(0.003)	(0.001)	(0.003)	(0.003)
(2.) ♀	0.713	0.749	0.744	0.726	0.756	0.749
	(0.000)	(0.003)	(0.004)	(0.001)	(0.003)	(0.004)
↪(1)-(2)	0.001	-0.006	-0.002	0.000	-0.004	-0.004
	(0.001)	(0.004)	(0.005)	(0.001)	(0.005)	(0.005)
Observations	564,445	60,576	54,146	456,066	50,839	45,983
Sex of first two children in families with two or more children						
(3.) ♂♀ or ♀♂	0.360	0.419	0.423	0.352	0.421	0.419
	(0.001)	(0.005)	(0.005)	(0.001)	(0.005)	(0.006)
♂♂	0.413	0.471	0.498	0.408	0.467	0.490
	(0.002)	(0.006)	(0.007)	(0.002)	(0.007)	(0.008)
♀♀	0.432	0.488	0.490	0.430	0.491	0.494
	(0.002)	(0.006)	(0.008)	(0.002)	(0.007)	(0.008)
(4.) ♂♂ or ♀♀	0.422	0.479	0.494	0.419	0.479	0.492
	(0.001)	(0.004)	(0.005)	(0.001)	(0.005)	(0.005)
↪(3)-(4)	-0.062	-0.060	-0.071	-0.056	-0.058	-0.073
	(0.002)	(0.006)	(0.007)	(0.002)	(0.007)	(0.007)
Observations	360,919	35,892	31,633	297,124	30,258	26,805

The table displays the probability of having an additional child given the sex of existing children in the household; specifically, the probability of having a second child given the sex of the first-born child, and the probability of having a third child given the sex of the first- and second-born child. The sample includes mothers aged between 21 and 35 years with at least one child, whose age or sex (or that of any children) has not been imputed by the Census Bureau. Married couples include only those couples that are married at the date of survey and have only been married to each other.

—| SECTION 5 |—

LOCAL ESTIMATION

Existing empirical evidence as to the effect of children on maternal employment is mixed. ANGRIST and EVANS (1998) finds that additional children reduce the probability that a mother is employed. IACOVOU (2001) and AL-KHAJA (2016) conduct comparable analyses in the context of the British and Egyptian labour markets respectively, instead finding that additional children increase the probability that a mother is employed. Common to all three papers is the parameter that is recovered; specifically, ANGRIST and EVANS (1998), IACOVOU (2001) and AL-KHAJA (2016) recover a Local Average Treatment Effect (LATE; IMBENS and ANGRIST, 1994) that is the effect of children on maternal employment

FIGURE E.4
Differences in means for demographic variables by instrumental variable definition

	Differences in means (and standard errors)					
	PUMS 1980		ACS 2010		ACS 2015	
	Samesex	Plural birth	Samesex	Plural birth	Samesex	Plural birth
Age	-0.021	0.259	0.075	0.632	-0.006	-0.445
	(0.015)	(0.071)	(0.047)	(0.198)	(0.045)	(0.334)
Age at first birth	0.009	0.297	0.110	0.291	0.063	-0.770
	(0.011)	(0.067)	(0.051)	(0.270)	(0.050)	(0.336)
Black	0.000	0.019	0.003	0.014	0.000	-0.020
	(0.001)	(0.008)	(0.004)	(0.029)	(0.004)	(0.021)
White	0.000	-0.019	-0.007	0.089	0.003	-0.017
	(0.001)	(0.009)	(0.006)	(0.038)	(0.007)	(0.050)
Other race	-0.000	-0.000	0.004	-0.102	-0.102	0.037
	(0.001)	(0.004)	(0.005)	(0.029)	(0.006)	(0.047)
Hispanic	-0.001	-0.011	-0.009	-0.115	-0.008	-0.089
	(0.001)	(0.007)	(0.006)	(0.045)	(0.008)	(0.044)
Observations	297,124	171,515	30,258	13,784	26,805	12,232

The sample includes mothers aged between 21 and 35 years with at least two children, whose age or gender (or that of any children) has not been imputed by the Census Bureau. Married couples include only those couples that are married at the date of survey and have only been married to each other. The table displays the difference in the mean of each variable between each the instrument, for each definition of the instrument. The columns titled Samesex *state the difference in mean when the first two children share the same gender versus when they do not. The columns titled* Plural birth *state the difference in mean when a mother experiences a plural second birth versus when she does not. The variables in the left-hand column are demographic and ethnic and racial indicators.*

amongst an unidentifiable subset of the sample population. In Figure E.5 I report comparable estimates for the census and ACS data that I study.[14] Like ANGRIST and EVANS (1998), I find that the local effect of additional children in 1980 is negative and statistically significant. Estimates for 2010 and 2015, however, vary. In particular, I find that the local effect of additional children in 2010 is positive and statistically significant. Furthermore, the absolute magnitudes of these effects vary considerably, from four percentage points up to 25 percentage points.

Although the local effects that I report are empirically interesting, I include these estimates for two reasons. First, to emphasise that local effects may provide mixed evidence. I do not intend this as a criticism. Instead, I view this ambiguity as a natural implication of allowing for heterogeneity of response to treatment. I argue that it is interesting to think about how these local effects fit with evidence about an effect in the wider population. For example, does the sub-population that is affected by a policy represent a group that is particularly responsive to treatment? Second, to provide a numerical benchmark that may be useful for interpreting subsequent estimates.

14 I note that the Two-stage least squares estimates that I report relies on a linear model. I specify this model to include age and age at first birth (not discretised; in years), and a full set of interactions between ethnicity (Hispanic or non-Hispanic) and race (black, white, and other race).

FIGURE E.5
Estimates of local treatment effects

	Estimates and predictions (with standard errors)					
	PUMS 1980		ACS 2010		ACS 2015	
	Samesex	Plural birth	Samesex	Plural birth	Samesex	Plural birth
Wald estimator						
Intercept	0.585	0.612	0.624	0.480	0.638	0.680
	(0.011)	(0.014)	(0.053)	(0.069)	(0.048)	(0.150)
LATE	-0.131	-0.067	0.008	0.257	-0.043	-0.104
	(0.027)	(0.027)	(0.118)	(0.111)	(0.104)	(0.094)
Two-stage least squares estimator						
Intercept	0.581	0.613	0.633	0.515	0.639	0.663
	(0.010)	(0.014)	(0.050)	(0.065)	(0.045)	(0.100)
LATE	-0.121	-0.070	-0.015	0.199	-0.047	-0.080
	(0.026)	(0.027)	(0.111)	(0.104)	(0.099)	(0.159)
Observations	297,124	171,515	30,258	13,784	26,805	12,232

The sample includes mothers aged between 21 and 35 years with at least two children, whose age or gender (or that of any children) has not been imputed by the Census Bureau. Married couples include only those couples that are married at the date of survey and have only been married to each other. Estimates of local treatment effects using the Wald estimator and the Two-stage least squares estimator are stated. The Two-stage least squares estimator includes as covariates all variables in Table E.4 plus interaction terms between the racial and ethnic indicators. Intercept *is reported at the mean of the covariates for the Two-stage least squares estimator.*

—| SECTION 6 |—

ESTIMATION

I focus on the ATE and its primitives, the so-called potential outcome distributions (D. B. RUBIN, 1974), that may inform both *ex-ante* policy design and *ex-post* policy evaluation. I report estimates of these objects in Figure E.6 and in Figure E.7. Each estimate is a non-trivial set, reflecting the fact that the Balke-Pearl model partially identifies the ATE and its primitives. Each estimate is obtained as the sample analogue of a set of bounds that are implied by the model, and are calculated using a half-median unbiased estimator that is described in CHERNOZHUKOV et al. (2013). Focusing on the ATE, the interpretation of an estimate is then as a set that contains the sample analogue of the ATE with probability one; the sample analogue of the ATE may then take any value in this set.

I separate the analysis into two parts, by estimating the ATE and its primitives with and without covariates. I note that the model without covariates implies the generalised Balke-Pearl bounds when the sub-population x is taken to be the population. By separating the analysis into these two parts, I

am able to study the influence that covariates have on estimates.[15] If covariates and the instrument are not statistically independent then it is (theoretically) necessary to include covariates. Such a statement, however, does not preclude the possibility that there is no empirical benefit to including covariates.

I make the following observations. First, that estimates are broadly uninformative. In particular, I fail to recover the sign of the effect of additional children on maternal employment. Effects of 20 percentage points or more cannot be ruled-out in many cases. Second, I cannot reject the hypothesis that the effect of additional children on maternal employment is unchanged over the study period, although I note that I also am unable to reject the competing hypotheses that the effect is increasing or is decreasing. Third, that there is little difference between the estimates that are presented in Figure E.6 versus those that are presented in Figure E.7. One comparison where this observation does not hold is when I define the instrument as a plural second birth and allow age to influence the employment decision and the fertility decision. In this instance, there is considerable difference between the estimates that are delivered under each model. As can be seen in Figure E.4, there is strong correlation between the age of a mother and the probability of a plural second birth. I identified such a case as (theoretically) necessitating the inclusion of covariates, and I find that this is supported empirically. I note that the confidence regions that I report suffer from projection conservatism, and it is possible that this observation is driven simply by the inclusion of an exogenous characteristic with (relatively) many points of support. Fourth, that estimates that use the definition of the instrument as a plural second birth are more informative than those that use the definition of the instrument as the sex composition of a mother's first two children. I note that a plural second birth is a monotonic variable; plural second birth acts mechanically, in that a mother that experiences a plural second birth has three or more children by definition. Put differently, a mother that experiences a plural second birth cannot have two children. It follows that this definition of the instrument precludes some response types and has strong identifying power. Fifth, that the local effects that I report in Figure E.5 do not appear to be extreme in relation to the estimates of the ATE that I report.

—| Appendix E.1 |—

The suitability of an auxiliary model

An interesting question that naturally arises is whether it is necessary to include covariates. More precisely, does omitting relevant covariates constitute a form of model misspecification? Or, is an analysis *à la* Balke and Pearl (1997) sufficient? The short answer is that it depends. If covariates and the instrument are statistically independent then it is not necessary to include covariates, although the

15 I address the question of whether it is necessary to include covariates in an appendix.

FIGURE E.6
Estimates of global treatment effects without covariates

	Half-median unbiased estimates (with 95% confidence sets)					
	PUMS 1980		ACS 2010		ACS 2015	
	$\bar{Y}_{0,n}$	ATE_n	$\bar{Y}_{0,n}$	ATE_n	$\bar{Y}_{0,n}$	ATE_n
Samesex	0.271↔0.629	-0.405↔0.535	0.099↔0.539	-0.407↔0.573	0.113↔0.553	-0.401↔0.570
	(0.269↔0.631)	(-0.410↔0.540)	(0.094↔0.549)	(-0.423↔0.588)	(0.107↔0.564)	(-0.419↔0.588)
Gender decomposition	0.271↔0.629	-0.403↔0.528	0.098↔0.534	-0.403↔0.528	0.113↔0.554	-0.403↔0.569
	(0.270↔0.631)	(-0.406↔0.531)	(0.093↔0.542)	(-0.415↔0.578)	(0.107↔0.562)	(-0.417↔0.583)
Observations	297,124		30,258		26,805	
Plural second birth	0.175↔0.679	-0.252↔0.291	0.064↔0.692	-0.513↔0.180	0.089↔0.714	-0.506↔0.261
	(0.164↔0.681)	(-0.279↔0.331)	(0.056↔0.702)	(-0.530↔0.246)	(0.070↔0.725)	(-0.525↔0.373)
Plural second birth and gender decomposition	0.179↔0.653	-0.224↔0.296	0.066↔0.647	-0.467↔0.186	0.090↔0.681	-0.466↔0.276
	(0.177↔0.656)	(-0.240↔0.314)	(0.059↔0.657)	(-0.486↔0.237)	(0.081↔0.692)	(-0.487↔0.348)
Observations	171,515		13,784		12,232	

The sample includes mothers aged between 21 and 35 years with at least two children, whose age or gender (or that of any children) has not been imputed by the Census Bureau. Married couples include only those couples that are married at the date of survey and have only been married to each other. Half-median unbiased estimates and 95% confidence regions are reported; these estimates are calculated using a method outlined in CHERNOZHUKOV *et al. (2013). Estimates are presented for various definitions of the instrument.* Samesex *is defined as the event that the first two children share the same gender;* gender decomposition *decomposes gender further into the gender of the first child and the gender of the second child;* plural second birth *is defined as the event that a mother experiences a plural second birth; and* Plural second birth and gender decomposition *combines the two definitions of the instrument, giving priority to a plural second birth.*

FIGURE E.7
Estimates of global treatment effects with covariates

	Half-median unbiased estimates and p-values (confidence sets)								
	PUMS 1980			ACS 2010			ACS 2015		
	$\bar{Y}_{0:n}$	ATE_n	χ^2	$\bar{Y}_{0:n}$	ATE_n	χ^2	$\bar{Y}_{0:n}$	ATE_n	χ^2
Samesex									
Hispanic	0.271↔0.628	-0.407↔0.534	0.176	0.100↔0.538	-0.406↔0.571	0.000	0.114↔0.552	-0.399↔0.567	0.000
	(0.270↔0.630)	(-0.410↔0.537)		(0.096↔0.544)	(-0.416↔0.581)		(0.109↔0.559)	(-0.412↔0.579)	
Hispanic, age, age at first birth	0.263↔0.638	-0.426↔0.550	0.018	0.073↔0.542	-0.433↔0.619	0.000	0.088↔0.590	-0.465↔0.629	0.000
	(0.260↔0.638)	(-0.432↔0.555)		(0.069↔0.584)	(-0.482↔0.646)		(0.078↔0.605)	(-0.490↔0.653)	
Observations	297,124			30,258			26,805		
Plural second birth									
Hispanic	0.166↔0.678	-0.246↔0.305	0.000	0.059↔0.689	-0.511↔0.216	0.000	0.073↔0.712	-0.503↔0.311	0.671
	(0.164↔0.680)	(-0.261↔0.325)		(0.056↔0.696)	(-0.522↔0.258)		(0.068↔0.719)	(-0.516↔0.370)	
Age	0.165↔0.679	-0.260↔0.328	0.000	0.060↔0.691	-0.514↔0.222	0.000	Degenerate		0.000
	(0.163↔0.681)	(-0.279↔0.347)		(0.056↔0.698)	(-0.527↔0.279)				
Observations	171,515			13,784			12,232		

The sample includes mothers aged between 21 and 35 years with at least two children, whose age or gender (or that of any children) has not been imputed by the Census Bureau. Married couples include only those couples that are married at the date of survey and have only been married to each other. Half-median unbiased estimates and 95% confidence regions are reported; these estimates are calculated using a method outlined in CHERNOZHUKOV *et al. (2013). Estimates are presented for various definitions of the instrument. For each definition of the instrument, a vector of covariates is specified.* Hispanic *is a dichotomous indicator that a mother is Hispanic;* Hispanic, age and age at first birth *also includes the age of a mother and the age of her first birth (both in years). P-values of Pearson's χ-square test (that the instrument is random across exogenous covariates) are also reported.*

bounds that are stated in Balke and Pearl (1997) are not sharp and instead define an outer region. If many covariates are observed—so that there is a high-dimension vector of covariates—then there may be good reason to prefer to ignore covariates with the loss of some precision. If covariates and the instrument are not statistically independent then it is (theoretically) necessary to include covariates.

To develop a (hopefully) satisfactory explanation, I posit the following simple example. Imagine that h is a non-trivial function of x (i.e., h varies in x) so that covariates are in some sense relevant. I compare what can be learnt from the generalised Balke-Pearl model—Assumptions E.1 to E.3—with what can be learnt from Assumption E.2 combined with the following two assumptions.

Assumption E.4 (Simple exclusion)**.** *The data generating process satisfies*

$$\begin{aligned} s_y &= g_y(t,u) \\ s_t &= g_t(z,u) \end{aligned} \tag{E.10}$$

such that the generalised Balke-Pearl model is triangular in Y, T and Z.

Assumption E.5 (Simple statistical independence)**.** *The data generating process satisfies*

$$U|\{Z\} \sim \text{Uniform}(0,1) \tag{E.11}$$

for which g is appropriately normalised.

I note that Assumptions E.2, E.4 and E.5 are equivalent to the Balke-Pearl model. I intend Assumptions E.2, E.4 and E.5 to constitute an auxiliary model for Assumptions E.1 to E.3. To fix ideas, suppose that Assumptions E.1 to E.3 identify the $\text{ATE}(x)$ and the $\text{ATE}(\tilde{x})$, for some $a, b \in \mathfrak{Z} \mapsto (x, \tilde{x})$, up to the half-lines

$$\left\{ \begin{matrix} q(a) \\ q(b) \end{matrix} \right\} \leq \text{ATE}(x) \text{ and } \left\{ \begin{matrix} \tilde{q}(a) \\ \tilde{q}(b) \end{matrix} \right\} \leq \text{ATE}(\tilde{x}) \tag{E.12}$$

Of course, these bounds are overly simplistic and I do not claim that they reflect all of the identifying information that Assumptions E.1 to E.3 contains—they are a tool for exposition and serve only to convey a point. Given these bounds though, it follows that Assumptions E.1 to E.3 identify the ATE

up to the half-line

$$\left\{\begin{array}{l} q(a)\cdot\Pr(X=x)+\tilde{q}(a)\cdot\Pr(X=\tilde{x}) \\ q(a)\cdot\Pr(X=x)+\tilde{q}(b)\cdot\Pr(X=\tilde{x}) \\ q(b)\cdot\Pr(X=x)+\tilde{q}(a)\cdot\Pr(X=\tilde{x}) \\ q(b)\cdot\Pr(X=x)+\tilde{q}(b)\cdot\Pr(X=\tilde{x}) \end{array}\right\} \leq \text{ATE} \tag{E.13}$$

In contrast, Assumptions E.2, E.4 and E.5 do not identify either the $\text{ATE}(x)$ or the $\text{ATE}(\tilde{x})$ but do identify the ATE—albeit incorrectly—up to the half-line

$$\left\{\begin{array}{l} q(a)\cdot\Pr(X=x|Z=a)+\tilde{q}(a)\cdot\Pr(X=\tilde{x}|Z=a) \\ q(b)\cdot\Pr(X=x|Z=b)+\tilde{q}(b)\cdot\Pr(X=\tilde{x}|Z=b) \end{array}\right\} \leq \text{ATE} \tag{E.14}$$

There are two differences between the half-lines defined by Equations E.13 and E.14. First, the number of unique constraints is different. Second, each constraint assigns (possibly) different weights to each lower bound. If the conditional and unconditional probabilities are equal then it is immediately apparent that the first and fourth constraints of Equation E.13 coincide with the two constraints of Equation E.14. I note that this condition is equivalent to strict randomisation (of Z, independently of X). An obvious consequence is that the auxiliary model identifies an outer region of the identified set. If the conditional and unconditional probabilities are not, however, equal then it is not possible to determine how the two half-lines relate to each other. It is in this specific sense that that the auxiliary model is misspecified—the auxiliary model does not necessarily identify an outer region. I suggest that it is sensible to test for randomisation—something that can be done—before deciding whether to use an auxiliary model.

—| Appendix E.2 |—

Discussion of the generalised Balke-Pearl bounds

Bounding the Average Structural Functions in the context of the generalised Balke-Pearl model is equivalent to constructing a cover in u-space and mapping this to observable probabilities—an insight that Chesher and Rosen (2017) makes and exploits in a more general framework. Even in simple settings though, there are many possible ways to partition a set of latent types. Fortunately, Chesher and Rosen (2017) describes an algorithm that reduces the number of possible sets in u-space that must be considered to a smaller class that Galichon and Henry (2011) names the core-determining sets. Importantly, any undominated cover in u-space can be expressed in terms of the core-determining sets

FIGURE E.8
Latent types and the identification problem

$Y_1(x)$	$Y_0(x)$	$T_1(x)$	$T_0(x)$	Observable probabilities	
1	1	1	1	$Q_{11.1}(x)$	$Q_{11.0}(x)$
1	1	1	0	$Q_{11.1}(x)$	$Q_{10.0}(x)$
1	1	0	1	$Q_{10.1}(x)$	$Q_{11.0}(x)$
1	1	0	0	$Q_{10.1}(x)$	$Q_{10.0}(x)$
1	0	1	1	$Q_{11.1}(x)$	$Q_{11.0}(x)$
1	0	1	0	$Q_{11.1}(x)$	$Q_{00.0}(x)$
1	0	0	1	$Q_{00.1}(x)$	$Q_{11.0}(x)$
1	0	0	0	$Q_{00.1}(x)$	$Q_{00.0}(x)$
0	1	1	1	$Q_{01.1}(x)$	$Q_{01.0}(x)$
0	1	1	0	$Q_{01.1}(x)$	$Q_{10.0}(x)$
0	1	0	1	$Q_{10.1}(x)$	$Q_{01.0}(x)$
0	1	0	0	$Q_{10.1}(x)$	$Q_{10.0}(x)$
0	0	1	1	$Q_{01.1}(x)$	$Q_{01.0}(x)$
0	0	1	0	$Q_{01.1}(x)$	$Q_{00.0}(x)$
0	0	0	1	$Q_{00.1}(x)$	$Q_{01.0}(x)$
0	0	1	0	$Q_{00.1}(x)$	$Q_{00.0}(x)$

The table associates latent types—as defined in Equation E.4—with observable probabilities in a wholly dichotomous framework in which covariates can take only one value.

and their intersection with those latent types that underpin the structural characteristic of interest.[16,17] Moreover, the core-determining sets can be stratified according to covariates, which is especially useful in practice given that the support of covariates is responsible for a curse of dimensionality on the number of latent types. The question that I ask here—and that I attempt to provide an intuitive answer to—is what motivates and rationalises this stratification?

As a preliminary step, I focus on the case in which covariates can take only one value and—for illustration—study a wholly dichotomous framework, which is sufficient to introduce a result that I subsequently rely upon. As is stated in BALKE and PEARL (1997) and its preceding works, this framework is associated with a total of 16 latent types, each defined according to the potential outcomes that they induce for different values of the observable exogenous random variates (in this case, simply different values of the instrument). These latent types are listed in Figure E.8 alongside the observable probabilities that they can map to, as determined by the capacity functional that is defined in MOLCHANOV (2005). What is immediately apparent is that several latent types are associated with the same observable probabilities and it is these collections of latent types—that are disjoint for a single value of the instrument, and overlapping for different values—that form the basis of the core-determining sets. A consequence is that any cover of all 16 latent types must map to observable probabilities that

16 By undominated I mean any cover that maps to a sharp bound on the structural characteristic of interest.

17 Here I emphasise that the core-determining sets themselves are not necessarily proper subsets of those latent types that underpin the structural characteristic of interest. Instead, they are unions of latent types, only some of which need be latent types that underpin the structural characteristic of interest.

sum to one or more, with violation of this a condition for falsification of the Generalised Balke-Pearl model as is explored in Kédagni and Mourifié (2020).

Supposing instead that covariates can take two values, there are a total of 256 latent types even in this simple setting, which serves to illustrate the aforementioned curse. Rather than extend Figure E.8, I instead ask the reader to imagine a matrix comprising 16 rows and 16 columns—with each row corresponding to specific behaviour given one value of the covariates, and each column corresponding to specific behaviour given the other. With this analogy in mind, the problem of bounding one of the Average Structural Functions amounts to constructing a cover over some of the rows or some of the columns. Since each row intersects every column though, the previously stated consequence implies that using columns to cover a row—or *vice versa*—necessarily maps to a trivial unit bound. Such a cover is clearly dominated by using groups of rows to cover rows and groups of columns to cover columns—what amounts to stratification. It is for this reason that each value of the covariates can be considered in isolation.

I remark that this argument does not specifically rely upon the dimensionality of the framework—whether that is with respect to the support of the instrument or the support of the covariates. In particular, the matrix can be expanded in any direction to accommodate the increased number of behaviours that are created by additional points of support of the instrument conditional on a given value of the covariates; the analysis of Beresteanu et al. (2012) can easily be extended to derive Equations E.8 and E.9 in any direction in which the instrument has more than two points of support; and the dimension of the matrix can be increased to accommodate additional points of support of the covariates, with rows and columns transforming into planes and higher-dimensional equivalents.

—| Appendix E.3 |—

Deriving the generalised Balke-Pearl bounds*

This appendix formalises the ideas that are advanced in the preceding appendix, presenting a formal proof of the statement that Equations E.8 and E.9 constitute sharp bounds on the Average Structural Functions.

The proof itself adapts other works. In particular, I lean heavily on results in Beresteanu et al. (2012), Chesher and Rosen (2017), and Richardson and Robins (2014), following the approaches taken in Kédagni and Mourifié (2020) and Richardson and Robins (2024) to proving sharpness. To facilitate this, I exploit the mapping that exists between the structural equation and potential outcome notations

* I thank my examiners—Professors Dennis Kristensen and Karim Chalak—for their helpful suggestions during my oral examination. This appendix is designed to commit their required amendments.

that is discussed in Section 2, using the notation that is more convenient for whichever result I wish to show. I mirror Richardson and Robins (2024) in separating the proof of sharpness of Equations E.8 and E.9 into two parts—establishing that

$$\begin{aligned} P(Y_1(x) = y_1, Y_0(x) = y_0) &\leq Q_{y_1 1.z}(x) + Q_{y_0 0.z}(x) \\ P(Y_t(x) = 1 - y) &\leq 1 - Q_{yt.z}(x) \end{aligned} \tag{E.15}$$

characterises $P(\{Y_1(x), Y_0(x) : x \in \mathfrak{X}\})$, which is Richardson and Robins (2024, §Theorem 1); and then showing that Equations E.8 and E.9 are the vertices of the linear programme that is implied by Equation E.15, which is Richardson and Robins (2024, §Theorem 2). The first part necessitates (i.) showing that the set of structures that are admitted by the generalised Balke-Pearl model and that are compatible with the observable distribution is contained within the set of structures that are implied by Equation E.15, and (ii.) proposing a structure that is compatible with Equation E.15 and that is admitted by the generalised Balke-Pearl model. To do (i.), I apply results from random set theory, as in Beresteanu et al. (2012) and Chesher and Rosen (2017); to do (ii.), I propose a minor adaptation of the marginal distribution that is proposed in Kédagni and Mourifié (2020)—a distribution that exhibits many of the properties that I require—and invoke Richardson and Robins (2024, §Lemma 4) to establish the existence of a joint distribution with the required properties. The second part is somewhat redundant, as Richardson and Robins (2024, §Theorem 2) can always be invoked upon noting that any cover of the Average Structural Functions only uses observable quantities with the same value of covariates—an implication of the idea that is pursued in the preceding appendix that only by using all columns can a row or rows of a matrix be covered.

In what follows, I adopt the following convention. I signify the power set over a collection by including a lemniscate in the superscript position (as I do for more general spaces). For example, $\mathfrak{U}^\infty$ denotes the power set over $\mathfrak{U}$. Where the collection is the image of a correspondence, I position the lemniscate before the parentheses enclosing the arguments.

To facilitate this identification analysis, I sequentially define

$$\text{Contour}(y, t, z, x) \doteq \{u : y = 1(h_y(t, x, u) \geq 0) \text{ and } t = 1(h_t(z, x, u) \geq 0)\} \tag{E.16}$$

$$\text{Level}(\mathbf{u}, z, x) \doteq \{y, t : \text{Contour}(y, t, z, x) \cap \mathbf{u} \neq \emptyset\} \tag{E.17}$$

$$\text{Capacity}(\mathbf{u}, z, x) \doteq Q(\text{Level}(\mathbf{u}, z, x)|\mathbf{z}, \mathbf{x}) \tag{E.18}$$

which I emphasise are functionals that depend implicitly upon the population and structural equations, and where $\mathbf{u} \in \mathfrak{U}^\infty$.

The most general characterisation of those admissible structures that are compatible with a given observable distribution is

$$\text{Artstein}(Q,z,x) \doteq \{P,h : P(\mathbf{u}) \leq \text{Capacity}(\mathbf{u},z,x) \text{ for all } \mathbf{u} \in \mathfrak{U}^{\infty}\} \tag{E.19}$$

which is sharp (Artstein, 1983). If the intersection of Equation E.19 is finite for every observable distribution then the model has complete identification power; and if the intersection of Equation E.19 is empty for some observable distribution then the model is falsifiable.

I recall that heterogeneity is finite, in the sense that it reduces to a finite number of equivalence classes. As such, the power set enumerates all closed sets and unions of closed sets on its support—as is required by the theory of random sets from which Equation E.19 originates (Molchanov, 2005). Given that the capacity functional and population are non-decreasing and increasing in heterogeneity, respectively, there exists a proper subset of the power set that is sufficient to characterise the identification region that is elsewhere labelled the class of core-determining sets (Galichon and Henry, 2011). I let

$$\mathfrak{S}(z,x) \doteq \{\text{Contour}(y,t,z,x) : y,t \in \{1,0\} \times \{1,0\}\} \tag{E.20}$$

and

$$\Theta(Q,z,x) \doteq \{P,h : P(\mathbf{u}) \leq \text{Capacity}(\mathbf{u},z,x) \text{ for all } \mathbf{u} \in \mathfrak{S}(z,x)\} \tag{E.21}$$

for convenience. The generalised Balke-Pearl model identifies

$$\Theta_{Q} \doteq \cap_{x \in \mathfrak{X}} \cap_{z \in \mathfrak{Z} \leftarrow\!\shortmid x} \Theta(Q,z,x) \tag{E.22}$$

as the identification region. That is, Equation E.19 reduces to Equation E.21, where Equation E.20 is the appropriate class of core-determining sets (Chesher and Rosen, 2017). Importantly, the class of core-determining sets contains only images of the contour functional corresponding to the specified values of the observable exogenous random variates, which is a consequence of how the contour functional connects.[25]

The structural equations enact a labelling of heterogeneity (i.e., which realisation of heterogeneity induces which responses and treatments for particular combinations of the observable exogenous random

25 The capacity functional is the complement of the containment functional (Molchanov, 2005). It is straightforward to translate results from one to the other by, for example, substituting intersections for unions. The class of core-determining sets contains only images of the contour functional and their unions (Chesher and Rosen, 2017, §Lemma 1) that cannot be partitioned into disjoint images (Chesher and Rosen, 2017, §Theorem 3).

variates). Fixing this association normalises the structural equations and formulates the identification region as a linear programme over the population—to be precise, over the marginal distribution of heterogeneity. The identification region is, therefore, convex and non-finite due to the abundance of equivalence classes relative to the number of possible realisations of instruments and covariates (i.e., the linear programme has more unknowns than equations). Hence, the generalised Balke-Pearl model has incomplete identification power and is partially identifying since the identification region is non-trivial (Manski, 2003).

I reiterate that Equation E.22 is a sharp characterisation of the identification region. Although the solving of linear programmes typically incur low computational cost, the formulation of the constraint matrix above need not be feasible (the constraint matrix comprising $2^{2\cdot(\mathfrak{r}+1)+\sum_{x\in\mathfrak{X}}\mathfrak{z}\leftarrow\!\shortmid x}$ columns and $\sum_{x\in\mathfrak{X}}\mathfrak{z}\leftarrow\!\shortmid x$ rows). As such, it is of practical and theoretical importance to prove the earlier statement that Equations E.8 and E.9 constitute sharp bounds on the Average Structural Functions.

Given Equation E.22, the first result that I am required to show—the set of structures that are admitted by the generalised Balke-Pearl model and that are compatible with the observable distribution is contained within the set of structures that are implied by Equation E.15—is immediate upon addition of the inequalities. Specifically,

$$P(Y_t(x)=y_1,Y_0(x)=y_0)=P(Y_t(x)=y_1,Y_0(x)=y_0,T_z(x)=1)+P(Y_t(x)=y_1,Y_0(x)=y_0,T_z(x)=0) \tag{E.23}$$

$$\leq P(Y_t(x)=y_1,T_z(x)=1)+P(Y_0(x)=y_0,T_z(x)=0) \tag{E.24}$$

$$=P(u\in\text{Contour}(y_1,1,z,x))+P(u\in\text{Contour}(y_0,0,z,x)) \tag{E.25}$$

$$\leq Q_{y_1 1.z}(x)+Q_{y_0 0.z}(x) \tag{E.26}$$

and

$$P(Y_t(x)=1-y)\leq P(Y_t(x)=1-y,T_z(x)=t)+P(T_z(x)=1-t) \tag{E.27}$$

$$=P(u\in\text{Contour}(1-y,t,z,x))+\textstyle\sum_y P(u\in\text{Contour}(y,1-t,z,x)) \tag{E.28}$$

$$\leq 1-Q_{yt.z}(x) \tag{E.29}$$

which establishes the result. I could otherwise derive Equation E.15 directly via Equation E.19, as in Beresteanu et al. (2012).

The fact that the class of core-determining sets contains only images of the contour functional corre-

sponding to the specified values of the observable exogenous random variates (i.e., there is separation of the marginal distributions) is an ever-present feature of this identification analysis. As a trivial extension of Richardson and Robins (2024, §Lemma 4), if marginal distributions of the form

$$P^*_{z,x}(\{Y_1(x), Y_0(x) : x \in \mathfrak{X}\}, T_z(x)) \tag{E.30}$$

such that

$$P^*_{z,x}(\{Y_1(x), Y_0(x) : x \in \mathfrak{X}\}) = P^*(\{Y_1(x), Y_0(x) : x \in \mathfrak{X}\}) \tag{E.31}$$

can be proposed then there exists a single joint distribution of the form

$$P^*(\{Y_1(x), Y_0(x) : x \in \mathfrak{X}\}, \{T_z(x) : z \in \mathfrak{Z} \nleftarrow x, x \in \mathfrak{X}\}) \tag{E.32}$$

that can be obtained from the marginal distributions. It is this result that motivates continued separation of the marginal distributions in this identification analysis.

Kédagni and Mourifié (2020) proposes a collection of marginal distributions that that exhibit many of the properties that I require; I adapt this collection to suit the framework of the generalised Balke-Pearl model. I let

$$p_{y_1 y_0 t|z}(x) \doteq P^*(Y_1(x) = y_1, Y_0(x) = y_0, T_z(x) = t | z, x) \tag{E.33}$$

and

$$\begin{aligned} q_{i+j.z}(x) &\doteq \min_{z \in \mathfrak{Z} \nleftarrow x} (Q_{i1.z} + Q_{j0.z}) \circ (x) \\ \overline{Q}_{yt.z}(x) &\doteq \max_{z \in \mathfrak{Z} \nleftarrow x} (Q_{yt.z}(x)) \end{aligned} \tag{E.34}$$

for convenience, which is similar notation (although not exactly the same) to that used in Kédagni and Mourifié (2020, in particular, I transpose the position of the first two subscripts relative to that paper). I propose

$$p_{100|z}(x) = \min(A_x, Q_{00.z}(x), L_x - Q_{11.z}(x), U_x + A_x - Q_{11.z}(x) - Q_{10.z}(x)) \tag{E.35}$$

$$p_{101|z}(x) = A_x - p_{100|z}(x) \tag{E.36}$$

$$p_{110|z}(x) = L_x - Q_{11.z}(x) - p_{100|z}(x) \tag{E.37}$$

$$p_{111|z}(x) = Q_{11.z}(x) - A_x + p_{100|z}(x) \tag{E.38}$$

$$p_{010|z}(x) = Q_{11.z}(x) + Q_{10.z}(x) - L_x + p_{100|z}(x) \tag{E.39}$$

$$p_{011|z}(x) = U_x + A_x - Q_{11.z}(x) - Q_{10.z}(x) - p_{100|z}(x) \tag{E.40}$$

$$p_{000|z}(x) = Q_{00.z}(x) - p_{100|z}(x) \tag{E.41}$$

$$p_{001|z}(x) = 1 - U_x - A_x - Q_{00.z}(x) + p_{100|z}(x) \tag{E.42}$$

where

$$\begin{aligned} A_x &\doteq \max(0, 1 - q_{1+0.z}(x) - q_{0+0.z}(x) - q_{0+1.z}(x), \overline{Q}_{00.z}(x) - q_{0+0.z}(x), \overline{Q}_{11.z}(x) - q_{1+1.z}(x)) \\ U_x &\doteq \min(1 - \overline{Q}_{00.z}(x), q_{1+1.z}(x) + q_{0+1.z}(x)) \\ L_x &\doteq \max(\overline{Q}_{11.z}(x), 1 - q_{0+0.z}(x) - q_{0+1.z}(x)) \end{aligned} \tag{E.43}$$

which is a minor modification—the only difference is that I append covariates—of the collection that is proposed in Kédagni and Mourifié (2020). I then propose that

$$P^*(\{Y_1(x) = y_{1x}, Y_0(x) = y_{0x} : x \in \mathfrak{X}\}, T_z(x) = t | z, x) = p_{y_{1x}y_{0x}t}(x) \cdot \prod_{r \in \mathfrak{X} - x} \sum_t p_{y_{1r}y_{0r}t.\mathbf{e}_1^\top \mathfrak{Z} \leftrightarrow r}(r) \tag{E.44}$$

such that the marginals of interest incorporate statistical independence of potential outcomes—specifically, those pertaining to response—for different values of the covariates.[26] By construction, Equations E.35 to E.42 satisfy Equation E.15. It remains for me to show that Equations E.35 to E.42 are compatible with the generalised Balke-Pearl model (i.e., that P^* is a proper distribution and that it satisfies

$$P^*(\{Y_1(x) = y_{1x}, Y_0(x) = y_{0x} : x \in \mathfrak{X}\} | z, x) = P^*(\{Y_1(x) = y_{1x}, Y_0(x) = y_{0x} : x \in \mathfrak{X}\}) \tag{E.45}$$

as is required) and generates the observable distribution.[27]

I reiterate that Equations E.35 to E.42 amount to a minor modification of the collection that is proposed in Kédagni and Mourifié (2020). I am, therefore, able to use several of the results that are stated therein: first, that Equations E.35 to E.42 are non-negative and sum to one (i.e., that the proposed marginal distributions are proper); second, that P^* more generally—being a product of proper marginal

26 Something that the generalised Balke-Pearl model does not impose, but does not exclude either.

27 I emphasise that I do not require that

$$P^*(\{Y_1(x) = y_{1x}, Y_0(x) = y_{0x} : x \in \mathfrak{X}\}, T_z(x) | z, x) = P^*(\{Y_1(x) = y_{1x}, Y_0(x) = y_{0x} : x \in \mathfrak{X}\}, T_z(x)) \tag{E.46}$$

here. I recall that Richardson and Robins (2024, §Lemma 4) means that there exists a joint distribution satisfying a stronger condition than this provided that I can find marginals that align on $\{Y_1(x) = y_{1x}, Y_0(x) = y_{0x} : x \in \mathfrak{X}\}$; I claim that Equations E.35 to E.42 have this property.

distributions—is also proper; and third, that

$$p_{100|z}(x) + p_{101|z}(x) = A_x \tag{E.47}$$

$$p_{110|z}(x) + p_{111|z}(x) = L_x - A_x \tag{E.48}$$

$$p_{010|z}(x) + p_{011|z}(x) = U_x + A_x - L_x \tag{E.49}$$

$$p_{000|z}(x) + p_{001|z}(x) = 1 - U_x - A_x \tag{E.50}$$

which supports Equation E.45.[28] I remark that these properties can all be easily verified—the definition in Equation E.35 corresponds, for instance, to the maximum amount that can be subtracted from Equations E.36, E.37, E.40 and E.41 so that they remain non-negative, contingent upon an appropriate specification of the terms that are defined in Equation E.43. Moreover,

$$P^*(Y = j, T = 1|z, x) = P^*(Y_1(x) = j, T_z(x) = 1|z, x) \tag{E.51}$$

$$= p_{j11.z}(x) + p_{j01.z}(x) \tag{E.52}$$

$$= Q_{j1.z}(x) \tag{E.53}$$

$$= P(Y = j, T = 1|z, x) \tag{E.54}$$

and

$$P^*(Y = j, T = 0|z, x) = P^*(Y_0(x) = j, T_z(x) = 0|z, x) \tag{E.55}$$

$$= p_{1j0.z}(x) + p_{0j0.z}(x) \tag{E.56}$$

$$= Q_{j0.z}(x) \tag{E.57}$$

$$= P(Y = j, T = 1|z, x) \tag{E.58}$$

such that P^* and P are observationally equivalent. Given Equation E.22 and its primitive Equation E.21, however, this is also sufficient to determine that the proposed distribution is compatible with the restrictions that are imposed by the generalised Balke-Pearl model and that link the joint distribution over unobservable variates to the joint distribution over observable variates (i.e., P^* satisfies the inequalities in Equation E.21 with equality).

Having proved that Equation E.15 characterises $P(\{Y_1(x), Y_0(x) : x \in \mathfrak{X}\})$, I now turn my attention to the problem of combining elements of Equation E.15 to derive tight bounds on the Average Structural Functions. If a cover of the Average Structural Functions only uses observable quantities with the same

28 I emphasise that Equation E.44 defines the left-hand side of Equation E.45 as the product of elements of Equations E.47 to E.50 and so is invariant to the exogenous observable variates.

value of covariates then Equations E.8 and E.9 are correct—Richardson and Robins (2024, §Theorem 2) enumerates the vertices of the linear programme in this case. The only thing that I need to show then is that any bound that uses observable quantities with different values of covariates must be slack or is trivial. Clearly, any bound that uses observable quantities with different values of covariates cannot improve upon Equations E.8 and E.9 if it simply adds to Equations E.8 and E.9—all of the constituent types of the Average Structural Functions are already covered by Equations E.8 and E.9 (i.e., it is slack). As such, the only way that a bound that uses observable quantities with different values of covariates might improve upon Equations E.8 and E.9 is if there exists $P(Y_1(x) = y_1, Y_0(x) = y_0)$ that intersects with the Average Structural Function (or its complement) of interest and is not covered. For example, if interest is in $\overline{Y}_1(x)$ then the only possibilities for improvement are if (i.) $P(Y_1(x) = 1, Y_0(x) = 1)$ is covered but $P(Y_1(x) = 1, Y_0(x) = 0)$ is not, (ii.) $P(Y_1(x) = 1, Y_0(x) = 0)$ is covered but $P(Y_1(x) = 1, Y_0(x) = 1)$, or (iii.) if neither $P(Y_1(x) = 1, Y_0(x) = 1)$ nor $P(Y_1(x) = 1, Y_0(x) = 0)$ are covered. Here, by covered, I mean that the probability is bounded in the sense of Equation E.15 by a combination $Q_{11.z}(x) + Q_{y0.z}(x)$. There exists, therefore, some $P(Y_1(x) = y_1, Y_0(x) = y_0, T_z(x) = t)$ that intersects with the Average Structural Function (or its complement) of interest and is wholly uncovered. For example, in (i.) $P(Y_1(x) = 1, Y_0(x) = 0, T_z(x) = 0)$ is wholly uncovered, in (ii.) $P(Y_1(x) = 1, Y_0(x) = 1, T_z(x) = 0)$ is wholly uncovered, and in (iii.) $P(Y_1(x) = 1, Y_0(x) = y, T_z(x) = t)$ is wholly uncovered, which subsumes the other two instances. Extending this argument to observable quantities with different values of covariates though, if a candidate does not cover the entirety of the support then there exists

$$P(\{Y_1(x^*) : x^* \in \mathfrak{X} - x\} = \mathbf{y}_1, \{Y_0(x^*) : x^* \in \mathfrak{X} - x\} = \mathbf{y}_0, \{T_z(x^*) : z \in \mathfrak{Z} \mapsto x^*, x^* \in \mathfrak{X} - x\} = \mathbf{t}) \quad \text{E.59}$$

that is wholly uncovered. This implies, however, that

$$P(\{Y_1(x^*) : x^* \in \mathfrak{X}\} = \{\mathbf{y}_1, y_1\}, \{Y_0(x^*) : x^* \in \mathfrak{X}\} = \{\mathbf{y}_0, y_0\}, \{T_z(x^*) : z \in \mathfrak{Z} \mapsto x^*, x^* \in \mathfrak{X}\} = \{\mathbf{t}, t\}) \quad \text{E.60}$$

(i.e., the intersection of the event in Equation E.59 with the event that is wholly uncovered in (i.), (ii.) or (iii.), say) is wholly uncovered, and so such a candidate does not constitute a proper cover of the Average Structural function. A proper cover involving observable quantities with different values of covariates must therefore cover the entirety of the support of heterogeneity, and so must be at least one (i.e., it is trivial); such a bound is clearly not an improvement upon Equations E.8 and E.9.

—| References |—

Bibliography

Abadie, A., J. Angrist, and G. Imbens (2002).

Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings.

In: *Econometrica*.

Abramowitz, M. and I. A. Stegun (1964).

Handbook of mathematical functions with formulas, graphs, and mathematical tables.

US Government printing office.

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015).

Estimating conditional average treatment effects.

In: *Journal of Business & Economic Statistics*.

Adjaho, C. and T. Christensen (2022).

Externally Valid Treatment Choice.

In: *arXiv preprint arXiv:2205.05561*.

Alquier, P., J. Ridgway, and N. Chopin (2016).

On the properties of variational approximations of Gibbs posteriors.

In: *The Journal of Machine Learning Research*.

Amos, D. E. (1974).

Computation of modified Bessel functions and their ratios.

In: *Mathematics of Computation*.

Andrews, D. W. and G. Soares (2010).

Inference for parameters defined by moment inequalities using generalized moment selection.

In: *Econometrica*.

Angrist, J. D. and W. N. Evans (1998).

Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size.

In: *The American Economic Review*.

Arias, J., J. Rubio-Ramírez, and D. Waggoner (2018).

Inference Based on Structural Vector Autoregressions Identified with Sign and Zero Restrictions: Theory and Applications.

In: *Econometrica*.

Artstein, Z. (1983).

Distributions of random sets and random selections.

In: *Israel Journal of Mathematics*.

Athey, S. and S. Wager (2021).

Policy learning with observational data.

In: *Econometrica*.

Aumann, R. J. (1965).

Integrals of set-valued functions.

In: *Journal of Mathematical Analysis and Applications*.

Balke, A. and J. Pearl (1997).

Bounds on treatment effects from studies with imperfect compliance.

In: *Journal of the American Statistical Association*.

Bégin, L., P. Germain, F. Laviolette, and J.-F. Roy (2014).

PAC-Bayesian theory for transductive learning.

In: *Artificial Intelligence and Statistics*.

– (2016).

PAC-Bayesian bounds based on the Rényi divergence.

In: *Artificial Intelligence and Statistics*.

Beresteanu, A., I. Molchanov, and F. Molinari (2012).

Partial identification using random set theory.

In: *Journal of Econometrics*.

Best, D. and N. I. Fisher (1979).

Efficient simulation of the von Mises distribution.

In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Beygelzimer, A. and J. Langford (2009).

The offset tree for learning with partial labels.

In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*.

ACM.

Bhattacharya, D. and P. Dupas (2012).

Inferring welfare maximizing treatment assignment under budget constraints.

In: *Journal of Econometrics*.

Bhattacharyya, A. (1943).

On a measure of divergence between two statistical populations defined by their probability distributions.

In: *Bull. Calcutta Math. Soc.*

Bingham, C. (1974).

An antipodally symmetric distribution on the sphere.

In: *The Annals of Statistics*.

Bissiri, P. G., C. C. Holmes, and S. G. Walker (2016).

A general framework for updating belief distributions.

In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Bloom, H. S., L. L. Orr, S. H. Bell, et al. (1997).

The benefits and costs of JTPA Title II-A programs: Key findings from the National Job Training Partnership Act study.

In: *Journal of human resources*.

Blundell, R. W. and J. L. Powell (2004).

Endogeneity in semiparametric binary response models.

In: *The Review of Economic Studies*.

Bretagnolle, J. and C. Huber (1978).

Estimation des densités: risque minimax.

In: *Séminaire de probabilités de Strasbourg*.

Bugni, F. A. (2010).

Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set.

In: *Econometrica*.

Catoni, O. (2007).

PAC-Bayesian Supervised Classification.

In: *Lecture Notes-Monograph Series. IMS*.

Chamberlain, G. (2011).

Bayesian aspects of treatment choice.

In: *The Oxford Handbook of Bayesian Econometrics*.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013).

Intersection Bounds: Estimation and Inference.

In: *Econometrica*.

Chernozhukov, V. and H. Hong (2003a).

An MCMC approach to classical estimation.

In: *Journal of Econometrics*.

– (2003b).

An MCMC approach to classical estimation.

In: *Journal of Econometrics*.

Chesher, A. and A. M. Rosen (2013).

What do instrumental variable models deliver with discrete dependent variables?

In: *The American Economic Review*.

– (2017).

Generalized instrumental variable models.

In: *Econometrica*.

– (2020).

Generalized instrumental variable models, methods, and applications.

In: *Handbook of Econometrics*.

Elsevier.

Csaba, D. and B. Szoke (2020).

Learning with misspecified models.

Tech. rep.

mimeo.

D'Adamo, R. (2021).

Policy Learning Under Ambiguity.

In: *arXiv preprint arXiv:2111.10904*.

Dawid, A. P. (2008).

Beware of the DAG!

In: *Proceedings of the 2008 International Conference on Causality: Objectives and Assessment*.

Dehejia, R. H. (2005).

Program evaluation as a decision problem.

In: *Journal of Econometrics*.

Derbeko, P., R. El-Yaniv, and R. Meir (2004).

Explicit learning curves for transduction and application to clustering and compression algorithms.

In: *Journal of Artificial Intelligence Research*.

Dhillon, I. S. and S. Sra (2003).

Modeling data using directional distributions.

Tech. rep.

University of Texas at Austin, Department of Computer Sciences.

Diethe, T. (2015).

A Note on the Kullback-Leibler Divergence for the von Mises-Fisher distribution.

In: *arXiv preprint arXiv:1502.07104*.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2012).

Fairness through awareness.

In: *Proceedings of the 3rd innovations in theoretical computer science conference*.

Fisher, R. A. (1953).

Dispersion on a sphere.

In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*.

Galichon, A. and M. Henry (2011).

Set identification in models with multiple equilibria.

In: *The Review of Economic Studies*.

Germain, P., A. Lacasse, F. Laviolette, and M. Marchand (2009).

PAC-Bayesian learning of linear classifiers.

In: *Proceedings of the 26th Annual International Conference on Machine Learning*.

ACM.

Giacomini, R. and T. Kitagawa (2021a).

Robust Bayesian Inference for Set-identified Models.

In: *Econometrica*.

Giacomini, R. and T. Kitagawa (2021b).

Robust Bayesian inference for set-identified models.

In: *Econometrica*.

Gronau, R. (1977).

Leisure, home production, and work–the theory of the allocation of time revisited.

In: *Journal of political economy*.

Guedj, B. (2019).

A primer on PAC-Bayesian learning.

In: *arXiv preprint arXiv:1901.05353*.

Gunsilius, F. (2019).

A path-sampling method to partially identify causal effects in instrumental variable models.

In: *arXiv preprint arXiv:1910.09502*.

– (2020).

A path-sampling method to partially identify causal effects in instrumental variable models.

In: *arXiv preprint arXiv:1910.09502v2*.

Haavelmo, T. (1944).

The probability approach in econometrics.

In: *Econometrica*.

Han, S. (2022).

Optimal dynamic treatment regimes and partial welfare ordering.

In: *unpublished manuscript*.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997).

Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme.

In: *The review of economic studies*.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2006).

Understanding instrumental variables in models with essential heterogeneity.

In: *The review of economics and statistics*.

Hillen, T., K. J. Painter, A. C. Swan, and A. D. Murtha (2017).

Moments of von Mises and Fisher distributions and applications.

In: *Mathematical Biosciences & Engineering*.

Hirano, K. and J. R. Porter (2009).

Asymptotics for statistical treatment rules.

In: *Econometrica*.

Hornik, K. and B. Grün (2013).

On conjugate families and Jeffreys priors for von Mises–Fisher distributions.

In: *Journal of statistical planning and inference*.

Iacovou, M. (2001).

Fertility and female labour supply.

Institute for Social and Economic Research, University of Essex.

Imbens, G. W. and J. D. Angrist (1994).

Identification and estimation of Local Average Treatment Effects.

In: *Econometrica*.

Imbens, G. W. and D. B. Rubin (1997).

Estimating Outcome Distributions for Compliers in Instrumental Variables Models.

In: *The Review of Economic Studies*.

Ishihara, T. and T. Kitagawa (2021).

Evidence Aggregation for Treatment Choice.

In: *arXiv preprint arXiv:2108.06473*.

Kaido, H., F. Molinari, and J. Stoye (2019).

Confidence intervals for projections of partially identified parameters.

In: *Econometrica*.

Karlin, S. and H. Rubin (1956).

The theory of decision procedures for distributions with monotone likelihood ratio.

In: *The Annals of Mathematical Statistics*.

Kasy, M. (2016).

Partial identification, distributional preferences, and the welfare ranking of policies.

In: *Review of Economics and Statistics*.

Kédagni, D. and I. Mourifié (2020).

Generalized instrumental inequalities: testing the instrumental variable independence assumption.

In: *Biometrika*.

Kent, J. T., A. M. Ganeiber, and K. V. Mardia (2013).

A new method to simulate the Bingham and related distributions in directional data analysis with applications.

In: *arXiv preprint arXiv:1310.8110*.

Kent, J. T. (1982).

The Fisher-Bingham Distribution on the Sphere.

In: *Journal of the Royal Statistical Society. Series B (Methodological)*.

Al-Khaja, A. (2016).

Essays on female empowerment and women's status.

PhD thesis. University Colege London.

Kido, D. (2022).

Distributionally Robust Policy Learning with Wasserstein Distance.

In: *arXiv preprint arXiv:2205.04637*.

Kitagawa, T. (2021).

The identification region of the potential outcome distributions under instrument independence.

In: *Journal of Econometrics*.

Kitagawa, T., S. Lee, and C. Qiu (2022a).

Treatment Choice with Nonlinear Regret.

In: *arXiv preprint arXiv:2205.08586*.

Kitagawa, T., H. Lopez, and J. Rowley (2022b).

Stochastic Treatment Choice with Empirical Welfare Updating.

In: *arXiv preprint arXiv:2211.01537*.

Kitagawa, T., S. Sakaguchi, and A. Tetenov (2021).

Constrained classification and policy learning.

In: *arXiv preprint arXiv:2106.12886*.

Kitagawa, T. and A. Tetenov (2018a).

Who should be treated? empirical welfare maximization methods for treatment choice.

In: *Econometrica*.

– (2018b).

Who should be treated? empirical welfare maximization methods for treatment choice.

In: *Econometrica*.

– (2021).

Equality-Minded Treatment Choice.

In: *Journal of Business Economics and Statistics*.

Kline, B. and E. Tamer (2016).

Bayesian inference in a class of partially identified models.

In: *Quantitative Economics*.

Kock, A. B., D. Preinerstorfer, and B. Veliyev (2022).

Treatment recommendation with distributional targets.

In: *Journal of Econometrics*.

Kurz, G. and U. D. Hanebeck (2015).

Stochastic sampling of the hyperspherical von Mises–Fisher distribution without rejection methods.

In: *2015 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*.

IEEE.

Lattimore, T. and C. Szepesvári (2020).

Bandit algorithms.

Cambridge University Press.

Lin, L., V. Rao, and D. Dunson (2017).

Bayesian nonparametric inference on the Stiefel manifold.

In: *Statistica Sinica*.

Liu, Y. (2022).

Policy Learning under Endogeneity Using Instrumental Variables.

In: *arXiv preprint arXiv:2206.09883*.

Manski, C. F. (1975).

Maximum Score Estimation of the Stochastic Utility Model of Choice.

In: *Journal of Econometrics*.

– (1985).

Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator.

In: *Journal of Econometrics*.

– (2000a).

Identification problems and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice.

In: *Journal of Econometrics*.

– (2000b).

Using Studies of Treatment Response to Inform Treatment Choice in Heterogeneous Populations.

– (2003).

Partial identification of probability distributions.

Springer Science & Business Media.

– (2004a).

Statistical treatment rules for heterogeneous populations.

In: *Econometrica*.

– (2004b).

Statistical treatment rules for heterogeneous populations.

In: *Econometrica*.

– (2005).

Social choice with partial knowledge of treatment response.

Princeton University Press.

– (2007a).

Identification for prediction and decision.

Harvard University Press.

– (2007b).

Minimax-regret treatment choice with missing outcome data.

In: *Journal of Econometrics*.

– (2009).

The 2009 Lawrence R. Klein Lecture: Diversified treatment under ambiguity.

In: *International Economic Review*.

– (2022).

Identification and Statistical Decision Theory.

In: *arXiv preprint arXiv:2204.11318*.

Manski, C. F. and A. Tetenov (2007).

Admissible treatment rules for a risk-averse planner with experimental data on an innovation.

In: *Journal of Statistical Planning and Inference*.

Manski, C. F. (1989).

Anatomy of the Selection Problem.

In: *The Journal of Human Resources*.

Mardia, K. V. (1975).

Statistics of Directional Data.

In: *Journal of the Royal Statistical Society. Series B (Methodological)*.

Mardia, K. V. and S. El-Atoum (1976).

Bayesian inference for the von Mises-Fisher distribution.

In: *Biometrika*.

Mardia, K. V., G. Hughes, C. C. Taylor, and H. Singh (2008).

A multivariate von Mises distribution with applications to bioinformatics.

In: *Canadian Journal of Statistics*.

Mardia, K. V. and P. E. Jupp (2009).

Directional statistics.

John Wiley & Sons.

Maurer, A. (2004).

A note on the PAC Bayesian theorem.

In: *arXiv preprint cs/0411099*.

Mbakop, E. and M. Tabord-Meehan (2021).

Model selection for treatment choice: Penalized welfare maximization.

In: *Econometrica*.

McAllester, D. A. (1999).

Some PAC-Bayesian Theorems.

In: *Machine Learning*.

– (2003).

PAC-Bayesian stochastic model selection.

In: *Machine Learning*.

Molchanov, I. (2005).

Theory of Random Sets.

Springer verlag.

Moon, H. R. and F. Schorfheide (2012).

Bayesian and frequentist inference in partially identified models.

In: *Econometrica*.

Mourifié, I. (2015).

Sharp bounds on treatment effects in a binary triangular system.

In: *Journal of Econometrics*.

Nie, X., E. Brunskill, and S. Wager (2021).

Learning When-to-Treat Policies.

In: *Journal of the American Statistical Association*.

NIST (2021).

Digital Library of Mathematical Functions, Release 1.1.1 of 2021-03-15.

F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

OECD (2007).

Mothers in Paid Employment.

In.

Pearl, J. (2009).

Causality.

Cambridge University Press.

Pellatt, D. F. (2022).

PAC-Bayesian Treatment Allocation Under Budget Constraints.

In: *arXiv preprint arXiv:2212.09007*.

Pentina, A. and C. H. Lampert (2015).

Lifelong learning with non-iid tasks.

In: *Advances in Neural Information Processing Systems*.

Pinsker, M. S. (1964).

Information and information stability of random variables and processes.

Holden-Day.

Racine, J. S., L. Su, and A. Ullah (Feb. 2014).

The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics.

Oxford University Press.

Rai, Y. (2019).

Statistical Inference for Treatment Assignment Policies.

In: *unpublished manuscript*.

Richardson, T. S. and J. M. Robins (2014).

ACE bounds; SEMs with equilibrium conditions.

In: *Statistical Science*.

– (2024).

Assumptions and bounds in the instrumental variable model.

In: *arXiv preprint arXiv:2401.13758*.

Rosenbaum, P. R. and D. B. Rubin (1983).

The central role of the propensity score in observational studies for causal effects.

In: *Biometrika*.

Rubin, D. B. (1974).

Estimating causal effects of treatments in randomized and nonrandomized studies.

In: *Journal of Educational Psychology*.

Ruggles, S., K. Genadek, R. Goeken, J. Grover, and M. Sobek (2015).

Integrated Public Use Microdata Series: Version 6.0 [dataset].

Minneapolis, MN: University of Minnesota.

Sabelfeld, K. K. (2018).

Application of the von Mises–Fisher distribution to Random Walk on Spheres method for solving high-dimensional diffusion–advection–reaction equations.

In: *Statistics & Probability Letters*.

Sakaguchi, S. (2019).

Estimating Optimal Dynamic Treatment Assignment Rules under Intertemporal Budget Constraints.

Tech. rep.

cemmap working paper.

Sasaki, Y. and T. Ura (2020).

Welfare analysis via marginal treatment effects.

In: *arXiv preprint arXiv:2012.07624*.

Schlag, K. H. (2006).

ELEVEN-Tests needed for a Recommendation.

In: *Economics Working Papers*.

Seeger, M. (2002).

Pac-bayesian generalisation error bounds for gaussian process classification.

In: *Journal of machine learning research*.

Sei, T., H. Shibata, A. Takemura, K. Ohara, and N. Takayama (2013).

Properties and applications of Fisher distribution on the rotation group.

In: *Journal of Multivariate Analysis*.

Shaikh, A. M. and E. J. Vytlacil (2011).

Partial identification in triangular systems of equations with binary dependent variables.

In: *Econometrica*.

Shawe-Taylor, J. and R. C. Williamson (1997).

A PAC analysis of a Bayesian estimator.

In: *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*.

Splawa-Neyman, J. (1923).

On the application of probability theory to agricultural experiments. Essay on principles.]

In: *Roczniki Nauk Rolniczych*.

Stoye, J. (2009).

Minimax regret treatment choice with finite samples.

In: *Journal of Econometrics*.

– (2012).

Minimax regret treatment choice with covariates or with limited validity of experiments.

In: *Journal of Econometrics*.

Strotz, R. H. and H. O. Wold (1960).

Recursive vs. nonrecursive systems: An attempt at synthesis (part i of a triptych on causal chain systems).

In: *Econometrica: Journal of the Econometric Society*.

Sun, L. (2021).

Empirical welfare maximization with constraints.

In: *arXiv preprint*.

Swaminathan, A. and T. Joachims (2015).

Counterfactual risk minimization: Learning from logged bandit feedback.

In: *International Conference on Machine Learning*.

Tetenov, A. (2012).

Statistical treatment choice based on asymmetric minimax regret criteria.

In: *Journal of Econometrics*.

Thompson, W. R. (1933).

On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.

In: *Biometrika*.

Tsagris, M., G. Athineou, C. Adam, et al. (2022).

Directional: A Collection of Functions for Directional Data Analysis.

R package version 5.5.

Uhlig, H. (2005).

What are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure.

In: *Journal of Monetary Economics*.

Valiant, L. G. (1984).

A theory of the learnable.

In: *Proceedings of the sixteenth annual ACM symposium on Theory of computing*.

ACM.

van Erven, T. and P. Harremoës (2014).

Rényi divergence and Kullback-Leibler divergence.

In: *IEEE Transactions on Information Theory*.

Vapnik, V. N. (1998).

Statistical Learning Theory.

Wiley.

Viviano, D. (2021).

Policy targeting under network interference.

In: *arXiv preprint*.

von Mises, R. (1918).

Uber die'Ganzzahligkeit'der Atomgewicht und verwandte Fragen.

In: *Phys. Z.*

Vytlacil, E. J. (2002).

Independence, monotonicity, and latent index models: An equivalence result.

In: *Econometrica*.

Wainwright, M. J. and M. I. Jordan (2008).

Graphical Models, Exponential Families, and Variational Inference.

In: *Foundations and Trends in Machine Learning*.

Watson, G. S. (1984).

The theory of concentrated Langevin distributions.

In: *Journal of multivariate analysis*.

Wood, A. T. (1994).

Simulation of the von Mises Fisher distribution.

In: *Communications in statistics-simulation and computation*.

Wooldridge, J. M. (2015).

Control function methods in applied econometrics.

In: *Journal of Human Resources*.

Yata, K. (2021).

Optimal Decision Rules Under Partial Identification.

In: *arXiv preprint arXiv:2111.04926*.

Zadrozny, B. (2003).

Policy mining: Learning decision policies from fixed sets of data.

PhD thesis. University of California, San Diego.

Zhang, B., A. A. Tsiatis, M. Davidian, M. Zhang, and E. Laber (2012).

Estimating optimal treatment regimes from a classification perspective.

In: *Stat*.

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012).

Estimating individualized treatment rules using outcome weighted learning.

In: *Journal of the American Statistical Association*.

—| Statement |—

Division of labour

This statement—regarding the division of labour for multi-authored work—supplements the preceding declaration.

—| Declaration 1 |—

The f-divergence of a von Mises-Fisher distribution from some reference distributions

Chapter C extracted from a paper co-authored with Toru Kitagawa.

The idea for the paper was arrived at jointly by Toru and I. We recognised that there was an absence of results concerning the statistical divergence of spherical distributions—in particular, von Mises-Fisher distributions—having already identified a use for such results. I conducted preliminary research into the topic and we agreed that it was feasible to proceed with the project.

Work on the project was divided as follows. I was responsible for obtaining the main theoretical results, and for drafting and editing the resulting paper; Toru was responsible for reviewing the resulting paper, and for providing helpful discussion and perspective throughout the project.

Given this allocation of responsibility, I find it appropriate to describe only Toru's influence on the project and its output. First, Toru suggested the use of the moment-generating function—with its well-known form for exponential families—to derive the moments of a von Mises-Fisher distribution.[1] This approach was simpler than the approach that I had taken up until that point, which was to perform integration by substitution involving products of trigonometric functions. This simplicity extended to the resulting expressions of the moments. Second, Toru shaped the introduction by linking the subject-matter to a common economic model—the canonical binary choice model—and then to several other economic problems—optimisation of maximum score or empirical welfare criteria, and underidentifying linear

1 The moments of the distribution—specifically, the first moment—are necessary to obtain an closed-form expression for the Kullback-Leibler divergence. We originally intended that the paper would also present the moments of a von Mises-Fisher distribution as a result in itself, but Toru identified that some of these results were available elsewhere.

simultaneous equation models. Third, Toru read-through the paper and suggested minor changes.

As a reflection of this attribution of work and output, I include the paper in its entirety herein.

—| Declaration 2 |—

Stochastic treatment choice with empirical welfare updating

Chapter D extracted from a paper co-authored with Toru Kitagawa and Hugo Lopez.

The idea for the paper was arrived at jointly by Toru and Hugo. I was assigned to the project to provide research assistance and later upgraded to a co-author in recognition of my contribution to it. At the time I was assigned to the project, some theoretical results had been obtained and organised into an initial draft. I attribute these results to both Toru and Hugo since I am unaware of how work on the project was divided prior to my involvement (although I presume that Hugo prepared the initial draft).

The initial draft included an introduction, a framework, and several of the main theoretical results (first theorem and its proof, second theorem and its proof, third theorem and its proof, and first lemma and its proof). I was directed to apply these theoretical results to data and to rewrite the initial draft, which was incomplete and a work in progress. My influence on the paper can be seen throughout—due to my preparation of it—but my contribution to its substance is apparent in the proof of the second theorem, the third theorem and its proof, and parts that were not included in the initial draft.

Work on the project subsequent to my involvement was divided as follows. I was responsible for verifying the existing main theoretical results and for correcting any issues with these, for applying the aforementioned results to data, for rewriting the initial draft, and for editing the resulting paper; Toru was responsible for correcting any issues with the existing main theoretical results, and for providing helpful discussion and perspective throughout the project; Hugo was responsible for correcting any issues with the existing main theoretical results, and for editing the resulting paper.

Given this allocation of responsibility, I find it appropriate to describe the issues that I identified with the initial draft before discussing how these issues were corrected. First, the proof of the second theorem was incorrect and somewhat incomplete—although the theorem itself was correct. Second, the third theorem was stated for a von Mises—rather than a von Mises-Fisher—distribution and was, therefore, quite different.[2] Third, the proof of the third theorem was also incomplete in those parts that remained applicable.

2 Although the idea underpinning the third theorem remains the same—find an expression for the constant that appears in the first lemma in terms of the sample size such that the right-hand side of the inequality converges—the substance of this calculation is different.

The issue with the second theorem was due to the non-convexity of the objective function, which made establishing the veracity of the derivative as a solution for the case of a continuous prior difficult. Toru, Hugo and I all worked on this issue for several months, holding regular meetings to discuss our progress, with approaches ranging from contraction mapping to functional derivatives all rejected. The eventual proof is the culmination of our collective effort—although I note that the key insight that the derivative is monotone in the Lagrange multiplier is due to Toru.

The issue with the third theorem was that a von Mises distribution was—as I understand it—initially selected for two reasons—one valid, and one based upon an erroneous assumption. I persuaded Toru and Hugo to instead work with von Mises-Fisher distributions and provided the theoretical results that were necessary to do this. Toru, Hugo and I worked together to adapt the initial proof of the third theorem to incorporate these results. The eventual proof is the culmination of our collective effort.

The issue with those parts of the proof of the third theorem that remained applicable were that they were rather vague as to the rate of convergence. I worked independently to derive an expression for the constant and in doing so provided a concise expression of this rate.

As a reflection of this attribution of work and output, I include a partial reproduction of the paper herein—I omit several proofs (proof of the first theorem and proof of the first lemma) but include the associated theorems for the reader's comprehension. My influence on the substance of the paper is most apparent in the third theorem and its proof, and in the empirical application and its associated appendix.