COMPRESSION-DECOMPRESSION OF MULTIVARIATE DATA VIA MAXIMUM ENTROPY RESAMPLING AND APPLICATIONS TO INFERENCE

Dissertation zur Erlangung des Doktorgrades

vorgelegt von

Federico Bonofiglio

an der Fakultät für Mathematik und Physik
der Albert-Ludwigs-Universität Freiburg

Dekan: Prof. Dr. Gregor Herten
Physikalisches Institut,
Albert-Ludwigs-Universität Freiburg
Hermann-Herder-Straße 3
79104 Freiburg, Deutschland

1. Referent: Prof. Dr. Martin Schumacher
   Institut für Medizinische Biometrie und Statistik,
   Universitätsklinikum Freiburg, Medizinische Fakultät,
   Albert-Ludwigs-Universität Freiburg
   Stefan-Meier-Straße 26
   79104 Freiburg, Deutschland

2. Referent: Prof. Dr. Hein Putter
   Biomedical Data Sciences (Divisie 4),
   Leids Universitair Medisch Centrum, Faculteit Geneeskunde,
   Universiteit Leiden
   Einthovenweg 20
   2333 ZC Leiden, Nederland

Datum der mündlichen Prüfung: 13. September 2018

Erklärungen

Name, Vorname: _____

Adresse: _____

1. Ich erkläre hiermit, dass ich die vorliegende Arbeit **ohne** unzulässige Hilfe Dritter und **ohne** Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Insbesondere habe ich hierfür **nicht** die entgeltliche Hilfe von Vermittlungs bzw. Beratungsdiensten (Promotionsberater/-beraterinnen oder anderer Personen) in Anspruch genommen. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer Prüfungsbehrde vorgelegt.

2. Ich habe mich **nicht** bereits an einer in- oder ausländischen Hochschule um die Promotion beworben.

3. Die Bestimmungen der Promotionsordnung der Universität Freiburg für die Fakultät für Mathematik und Physik sind mir bekannt; insbesondere weiß ich, dass ich vor der Aushändigung der Doktorurkunde zur Führung des Doktorgrades **nicht** berechtigt bin.

Freiburg, den . . . . . .

. . . . . .

Unterschrift

Funding

This work was mostly supported by funding from the European Community's Seventh Framework Programme FP7/2011: Marie Curie Initial Training Network MEDIASRES ("Novel Statistical Methodology for Diagnostic/Prognostic and Therapeutic Studies and Systematic Reviews"; www.mediasres-itn.eu) with the Grant Agreement Number 290025.

Special thanks

I must express special thanks to Professor Martin Schumacher, and Professor Harald Binder, respectively former and current Head of the Institute of Medical Biometry and Statistics (IMBI), Medical Center Freiburg, Albert Ludwig University Freiburg (Germany), for their kind support and trust, without which any of this work could be done.

I express dear thanks to Professor Jan Beyersmann, Head of the Institute of Statistics, University of Ulm (Germany), for according me trust as appointee to the MEDIASRES Project, to Dr Guido Schwarzer (IMBI, Freiburg) for his keen support, to Professor Hein Putter from Leiden University Medical Center, Leiden (Netherland), for the constructive conversations and clarifying inspirations, and to Professor Ludger Rschendorf from the Department of Mathematical Stochastic, Albert Ludwig University Freiburg, for his interest and very helpful suggestions.

The list of people I must thank is long, and to all those who could not fit here, I express my best thanks. A special thank to all colleagues, internal and external mentors, and friends, who contributed to the good advancement of my work.

Ai miei Nonni, Nicoletta e Nandino

Preface

Roughly speaking this work starts with the scope to find new strategies of summary information appraisal, allowing for implementation of non-standard meta-analytic models. For instance, in order to perform general event-history meta-analysis we should recover time-dependent information about all cause-specific cumulative events counts across studies, that is typically a non-standard summary data input in classic meta-analysis. We also face the additional problem to retrieve summary information on all relevant study factors and confounders as well as to be able to flexibly model it. Hence a meta-analyzer only working with summary data would actually like to have as much original Individual Person Data (IPD) information as possible instead. This is because there seems to be no good work-around to gain relevant insight into original IPD information except actually having the IPD.

Here we try to argue for the opposite: there is an IPD compression that can retain and yield much of the original information if sensibly decompressed. Although counter-intuitive, it turns out such criterion is well founded in statistical physics and mechanics where entropy, the chaos or incertitude in a system, plays a key role. This and further connections to copula and inference theory slowly and strenuously helped to identify, and better clarify, a comprehensive method of IPD and IPD inference reconstruction when only certain compressed IPD items are available. This method could naturally fits into applications such as statistical disclosure control, research reproduction or syntheses.

For this work I shall be extremely thankful to the Institute of Medical Biometry and Statistics (IMBI) of the University Medical Center Freiburg, University of Freiburg (Germany), that hosted the project, and to the adjoining European Commission grant, MEDIASRES, that provided the starting means for it. Especially thanked shall be Dr. Nadine Binder who put the initial indispensable stimulus to outset this project. To ease the exposition I first present main concepts and results and put the rest into Appendix. I strove for simple notation and clear as possible notions. Despite the effort to address typos, lapses, and notation inconsistencies, something might still gone unchecked and I apologize for this in such case. In Chapter 2 and 3, which deal with theory/methods and results, each main section is followed by a summary that should help the reader speeding through the text. Most of the produced code used for the experiments is currently not freely available. The scope was not production of an open software package. While most of the proposed algorithms and data should not be difficult to reproduce, I'm available for possible code sharing upon personal request (`bono@imbi.uni-freiburg.de`).

> Briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.
>
> Fisher (1922); Section 2, page 278.

Abbreviations

IPD	Individual Person Data, a collection of p-dimensional dependent records, independently observed over n statistical units.
SDC	Statistical Disclosure Control.
rv	random variable
pm	probability measure
e.d.	empirical distribution
MaxEnt	Maximum Entropy
NORTA	NORmal To Anything (transformation)
NORTAmax	MaxEnt NORTA transformation (typically used to denote IPD simulation)
MaxEntBoot	Maximum Entropy Bootstrap (typically used to denote IPD inference simulation)
MLE	Maximum Likelihood Estimate
r.f.i.d.	reciprocal Fisher Information diagonal
c.h.e.	cumulative hazard estimate
MC	Monte Carlo
s.p.d.	semi positive definite
CIs	Confidence Intervals (typically 95% wide)
GLM	Generalized Linear Model
HR	Hazard Ratio
PH	Proportional Hazards
OR	Odds Ratio
p.s.s.	partially sufficient statistic (log-likelihood numerator); here the log-likelihood denominator is not reducible.

Notation

$\ell(\cdot)$	log-likelihood function
θ	a parameter, or a quantity of inferential interest
x	an observed $n \times p$ matrix, or IPD, with row $x_{i\cdot}$, and column $x_{\cdot j}$, $\quad i = 1, \ldots, n; \quad j = 1, \ldots, p$
$\bar{m}_j^k$	empirically observed moment of degree $k = 1, 2, \ldots$
$\bar{R}_x$	empirically observed correlation matrix
$\bar{C}_x$	abbreviation to denote both empirical moments and correlation matrix
X	random equivalent of x, with row $X_{i\cdot}$ and marginal $X_{\cdot j}$
$X_{i\cdot}$	a p-dimensional vector with dependent elemnts $(X_{i1}, X_{i2}, \ldots, X_{ip}) \sim Q, \quad \forall i = 1, \ldots n$
$X_{\cdot j}$	a n-dimensional vector with i.i.d. elements, $X_{1j}, X_{2j}, \ldots, X_{nj}, \quad X_{1j} \sim Q_j, \quad \forall j = 1, \ldots p$
E_P	expectation relative to a distribution P, also written as $E(\cdot)$
F_n	p-dimensional empirical distribution with marginal $F_{n,j} \quad j = 1, \ldots p$
$D(P\|\|Q)$	Kullback-Leibler divergence between P and Q
P^*	the p-dimensional MaxEnt distribution with marginal $P_j^* \quad j = 1, \ldots p$
$K(\cdot)$	a copula function
$\mathcal{N}_p$	the p-dimensional Normal distribution, also written as the function $\Phi(\cdot)$
$\mathscr{M}(\cdot)$	a generic inferential function
$\mathcal{I}_n^*$	the MaxEnt bootstrap estimator for the distribution of $\mathscr{M}(X)$, given plug-in P^*
θ^*	alternative notation for $\mathscr{M}(X^*)$

Abstract

Individual Person Data (IPD) typically consists of repeated independent observations of a multi-dimensional dependent record. Imagine collection of a multi-variate medical record (age, height, health-status, etc ...) on several unrelated patients. IPD sharing is crucial for scientific advancement, that is, for experimental validation, evidence pooling, and reliable statistical inferences. While IPD disclosure is feasible it is sometimes difficult or impossible.

If IPD is not available researchers still try to recover original information from disclosed IPD syntheses. For instance in meta-analysis we often focus on appraisal and combination of disclosed regression slopes. This is sometimes equivalent to perform the original pooled IPD regression but generally it is not. The implicit question is how much information about the original IPD, and IPD inference, the IPD syntheses do convey. The general opinion is that non negligible information loss should occur.

Here we propose a new paradigm by which appraisal of certain IPD summaries, that is IPD marginal moments and correlation matrix, seems to generally entail small information loss at both the data and inferential level. The idea is to reconstruct original IPD from the above summaries only, and to recover an original IPD inference from such reconstructed IPD. We argue this approach is well founded in an information theoretic sense which seems not fully acknowledged in the literature so far.

The reconstruction method is based on maximum entropy (MaxEnt) resampling where the basic MaxEnt formalism is extended to include record dependence by the aide of copula theory. We argue the Gaussian copula with given moment-based MaxEnt marginals and correlation matrix equals the multi-variate MaxEnt distribution from which stochastic simulations of the original IPD are drawn. By an extension of the renowned Gibbs conditioning principle there are strong hints the used Gaussian copula is asymptotically equal to the true IPD generating mechanism, given summaries on its empirical distribution. We verify such claims experimentally. So far this seems one of the strongest arguments for an objective method of IPD reconstruction from IPD summaries only.

Next we build a MaxEnt bootstrap estimator by using the proposed MaxEnt joint distribution as plug-in approximation for the IPD generating process, under conditions on its empirical summaries. We give hints of MaxEnt bootstrap consistency and argue for good predictive properties of a bootstrap average. Experimental assessments suggests the MaxEnt bootstrap does recover key features of an IPD inference distribution, or ensemble. We practically show this for commonly performed IPD inferences like Generalized Linear Models and proportional hazards Cox regression parameters, or Breslow/Nelson-Aalen type cumulative hazard, estimates.

The proposed method could find natural applications in IPD anonymization, distributed network computing, research reproduction and synthesis (meta-analysis), where no original IPD but only key IPD summaries can be made available. This work seems to suggest a new standard for IPD summary reporting and general IPD inference recovery, by which an important limitation of IPD information loss is possible.

Keywords

Individual Person Data, Statistical Disclosure Control, Marginal Moments, Correlation Matrix, Meta-analysis, Reproduction, Distributed Network, Maximum Entropy, Multivariate, Copula, NORTA, Generative model, Bootstrap, Generalized Linear Model, Cox Regression, Nelson-Aalen, Breslow.

Contents

Chapter 1

Introduction

1.1 Background and overview

Individual Person Data (IPD) consists of original information recorded at the study source. Here, we begin with considering non-longitudinal records. Each person, or statistical unit, contributes with several measurements at one single point in time. For example, we record medical measurements like age, height, clinical status, etc, once on a cohort of independent patients. The independence condition is usually enforced by design, or it is generally assumed as we do here.

While IPD is the empirical basis for scientific research, due to privacy, legal, or technical issues it is not always publicly available. Instead, critical IPD findings are usually only disclosed via journal publications, in form of a limited set of summaries. These summaries are typically the object of an intense recovering activity, and of further analytic processing. IPD summaries are priced proxy for original IPD information. Typically the focus is on some IPD inferential quantity, like the effect of a drug on some clinical outcome. For instance, in meta-analysis the focus is on collection and pooling of IPD regression slope estimates. This practice is not void of logic. It is the natural consequence of certain model parameter estimates being preferentially more highlighted and reported, thus, more easily recoverable, than other syntheses. However, when considering such summaries, the implicit attention in the literature is often framed in respect to IPD information, or individual level data.

Olkin and Sampson (1998) shows that pooling of linear regression slopes for a single categorical effect is in fact equivalent to the corresponding regression on the original IPD. Lin and Zeng (2010) extends this argument to vector-valued linear regression slopes, provided the slopes covariance matrix is available. Liu et al. (2015) further extend the claim by accounting for study heterogeneity. In modeling frameworks yet more general than those considered above it is typically accepted that some IPD information loss occurs via compression, further biasing inference (Debray et al., 2013a,b; Crowther et al., 2012; Stewart et al., 2012; Abo-Zaid et al., 2012; Cai et al., 2011; Jackson et al., 2011; Bowden et al., 2011; Riley and Steyerberg, 2010; Riley et al., 2010; Jones et al., 2009; Katsahian et al., 2008; Moodie et al., 2004; Stewart and Tierney, 2002). Analogous

issues are also relevant in fields like research synthesis, distributed computing and Statistical Disclosure Control (SDC).

Here we propose another paradigm with a procedure to recover original IPD from its empirical marginal moments and correlation matrix only. We argue this approach is well founded in an information theoretic sense. The Maximum Entropy Principle (Jaynes, 1957, 1996; Grünwald and Dawid, 2004) is a rational approach to decompress empirical moments into the associated raw variable. Such basic formalism is necessary but not sufficient for our goal of IPD reconstruction, because dependence information must also be factored in. The standard approach would be to impose empirical correlation knowledge via constrained optimization. This can be technical and we note a key connection between maximum entropy and copula theory seems to yield a more automatic solution. To the best of our knowledge this machinery was never used with the explicit goal to reconstruct original IPD. In fact the idea of IPD reconstruction is itself rather novel with only some historic application in SDC.

Using the above IPD summaries, our method can reproduce IPD emulations that convey quite generic inferential content. That is, we reconstruct original IPD inferences from the reconstructed IPD. We show this approach can be seen as a type of bootstrap, evocatively called the MaxEnt bootstrap (MaxEntBoot), that roughly retains key characteristics of the original IPD inference distribution, but where empirical IPD marginal moments and correlation structure is the only input data. We experimentally validate this via reproduction of commonly used IPD multi-variate parametric, semi-parametric, and non-parametric statistical models. Hence we give a broader answer to the question of Olkin and Sampson (1998) considering a wider class of multi-variate inferential problems (not only linear regressions).

The method can depend on some recovery assumptions about the required IPD empirical summaries. Marginal variable moments, at least to a certain degree, can be typically recovered through basic descriptive IPD summaries, often reported. Reporting of a correlation matrix may be less frequent, although is recently recognized as a relevant feature for generic recovery of inferential information (Lin and Zeng, 2010; Becker and Wu, 2007; Yoneoka and Henmi, 2016). Our work seems to further motivate and justify usage of such summaries for information reporting and retrieval. This could find further application in SDC, research reproduction or synthesis, and related disciplines.

Hence one main message for the practice is the following: there seems to be a rational and standardizing method to summarily report IPD information without incurring in both IPD and IPD inference serious information loss. Hopefully this might also help making reporting a bit more fit for purpose (Altman, 2015).

1.2 Further motivations

A speedy and easy access to IPD is ideal in the scientific practice. Claims must be quickly corroborated, empirical evidence must be rapidly appraised, pooled, and flexibly processed to confirm information or gain new insights. IPD access is perceived as a gold standard in meta-analysis (Chalmers, 1993), and it is particularly relevant when it comes to transparency issues or mere results reproduction (Peng, 2015; Stodden, 2015; Chan et al., 2014, 2013; Hrynaszkiewicz et al., 2010; Vickers, 2006). Recently, special editorials are regularly issued on the importance of IPD availability, and results reproducibility (Nature, 2017).

While IPD access may be feasible, it is not always possible, especially across different data sources. The most economic and efficient alternative may be IPD proxies collection, in the typical form of high level inferences made on IPD, like regression parameter or probability estimates. While such IPD proxies are typically highly valued, they may only carry indirect and limited information about the original IPD. Nevertheless, disclosed IPD proxies are becoming the object of intense exploitation, in order to recover as much IPD information as possible. For instance, Guyot et al. (2012); Liu et al. (2014) recently attempted to reconstruct survival IPD from published survival probability curves.

Standard meta-analytic procedures for survival data (Parmar et al., 1998) have evolved into more articulated methods (Williamson et al., 2002; Moodie et al., 2004; Arends et al., 2008; Fiocco et al., 2009, 2012; Commenges and Hejblum, 2013; Dias et al., 2013; Bonofiglio et al., 2015; Cafri et al., 2015), reflecting the desire to perform more flexible and realistic analyses, that depends on less restrictive and more comprehensive summary reporting. A richer and more reliable summary reporting can help recover various and more detailed aspects of IPD allowing for broader scope inferences. Equally, we see the attempt to shift from basic meta-analytic methods (DerSimonian and Laird, 1986) to less standard pooling procedures (Goodman, 1989; Eddy et al., 1990; Raftery et al., 1995; Schweder and Hjort, 1996; Higgins and Whitehead, 1996; Efron, 1996; ORourke et al., 2001; ORourke, 2001; O'Rourke and Altman, 2005; Baker and Kramer, 2005; Ades and Sutton, 2006; O'Rourke, 2007; Guolo, 2012, 2013; Madan et al., 2014), which often requires more adequate summaries as input data.

Distributed computing (Dean and Ghemawat, 2008; Gaye et al., 2014; Chamandy et al., 2015) is a similar field heavily relying on IPD proxies. Here due to privacy restrictions IPD cannot be shared across a network, but the goal is to pool anonymous IPD syntheses between web-connected hubs to gain as much as possible pooled IPD knowledge. Here the advantage is that the IPD summary format can be standardized across the network, allowing information consistency. For instance commonly used IPD pooled analyses can be performed by sharing anonymous regression scores and their covariance matrix, similar to Lin and Zeng (2010), across the network. However such approach does yet not extend to inclusion of random effects, which could be important to account for network heterogeneity. Similarly, most IPD inferential procedures cannot be reproduced via such pooling approach.

In Statistical Disclosure Control (SDC) we try to limit access to sensible IPD information when we make IPD publicly available. This is done by producing aggregated, altered, or entirely

simulated proxies of the original IPD, such to mask original sensible records (Hundepool et al., 2012; Templ, 2017). Attention is paid on the *utility* of the anonymized IPD, that is, its capacity to still yield inferential content similar to the original IPD. Some authors show that entirely simulated IPD (microdata in their jargon) can increase utility. Under the requirement to access original IPD census information, they propose IPD reconstruction schemes based on multiple imputation via Bayesian posterior distribution prediction (Raghunathan et al., 2003; Drechsler and Reiter, 2010), or random forests techniques (Reiter, 2005). Their IPD generation implementation has some vague conceptual resemblance to the approach we propose here, and that we broadly sketch in Figure 2.1, page 28. However, in more detail our approach is entirely different and does also eliminate the need to access original (census) IPD. The modes of IPD inference recovery are also different. While we focus on estimation of the IPD inference distribution in a broad sense, they focus on construction of point and variance estimators and, via combination, obtain interval estimators for certain statistics. Beaulieu-Jones et al. (2017); Huang et al. (2017) propose generative adversary neural networks to simulate anonymous raw IPD proxies, but their focus is not on IPD inference recovery, and the simulations' utility is not entirely clear.

Duchi et al. (2018) more formally study the trade-off between privacy preservation and utility optimization, working on conditional IPD generative models and on a number of information theoretic bounds. Their set-up requires access to original raw IPD. They assess utility on GLM and density estimation, similar to us. They conclude the cost of increased anonymization could be decreased utility, or the need to adjust IPD inferential procedures on privacy constraints, especially in high dimensional settings. They do not exclude the possibility of better trade-offs.

For $p < n$, $\text{IPD}_{n\times p}$ first dimension compression is equivalent to masking individual records information. By considering this type of aggregation, it is easy to think to classic statistical sufficiency. For $p = 1$ the log-likelihood with exponential form

$$\ell(\theta) = \theta s - \psi(\theta) + c, \tag{1.1}$$

allows for information compression $s = \sum_i^n y_i$, $\theta \in \mathbb{R}$, ψ a function, and c constant. Here vector y_n is the IPD and its scalar proxy, s, suffers no information loss. In this ideal uni-parametric example by only reporting and recovering s it would allow relatively ample model flexibility (in the sense of disposing of a wide family of distribution), guaranteeing full IPD information appraisal.

Inclusion of a covariate, x, is important. If x is categorical, the sufficiency principle still holds, and one only needs reporting/recovery of additional $2(m-1)$ summary features, where m is a contained number of categories, $\theta \in \mathbb{R}^m$, and n is given. Interestingly, even this latter trivial example shows that pooled IPD GLMs, based on (1.1), could be used for popular fixed-effects meta-analyses of treatment-control effects, if all sufficient summaries were recoverable. These trivial situations are conform to classic notions of evidential equivalence between an aggregate-based and an original IPD inference (Birnbaum, 1962, 1964, 1972; Kalbfleisch, 1975; Basu, 1975; Berger et al., 1988).

Another widely used model, the Cox regression, has partial log-likelihood similar to (1.1), where $\psi(\theta)$ is a risk-set denominator also depending on x, and θ is an Hazard Ratio vector. Here,

imagine survival data had only few events and x was a binary treatment-control covariate. Then, it is instructive to see we could in principle perform a model of the type

$$\ell(\theta) = \theta s - \sum_t \log\{n_{1t}\psi(\theta) + n_{0t}\}, \quad t = 1, 2, \dots T, \tag{1.2}$$

by only using the summaries s, n_{1t}, and n_{0t}, where T is a contained number of event-times, and n_{1t}, and n_{0t} are exact counts of at risk units in each arm, at time t. If T grows, exact reporting/recovery of (n_{1t}, n_{0t}), $\forall t$, becomes prohibitive, and (1.2) unrealistic (see Appendix B.5.1, page 95, for further development on this topic).

In general, if x is continuous, or multi-dimensional by inclusion of confounders and interactions, the log-likelihood denominator, $\psi(\theta, x)$, is irreducible and classical principles of inferential equivalence no longer hold. Here if raw x is not available we face a problem of missing information. One idea could be to simulate $\psi(\theta, x)$ by mimicking the unavailable x. Borgan and Keogh (2015); Keogh et al. (2018) do something similar in the context of case-control studies where some covariates value must be imputed. In Appendix B.5.2 we propose a similar approach when the log-likelihood compressed numerator is known. Otherwise, the main theme of this work is to simulate the entire IPD to also account for situations where no numerator compression is available.

Broadly speaking, if researchers can only access limited IPD information, and for the various application purposes introduced above, it emerges the importance of

- reliable and generic IPD proxies, that holds most of original IPD information.
- freedom and flexibility to model IPD proxies, in order to pool, reproduce, or *de novo* produce generic IPD inferences.

These two aspects are surely inter-related and here we confront both of them. For the first point we see that an intuitive and compact IPD proxy is the set of its empirical distributional descriptors, such as k-degree marginal moments and correlation matrix. It turns out these empirical IPD compressions have justification by unifying formal aspects of entropy maximization and copula theory. This leads to a decompressed IPD proxy that tend to satisfy the second point. That is, the generated IPD can be processed with the *same* analytic procedure we would use on the original IPD. This tends to show good utility by yielding IPD inferences quite similar to the original ones.

This process can be naturally seen through a classic information flow scheme (encoder-channel-decoder). We encode original IPD into a transportable anonymous summary and channel it through a certain MaxEnt distribution, whose samples are decoded IPD reconstructions. This transmission can be quite reliable. Moreover, from the inferential point of view, it is maybe easier to look at this compression-decompression program as a bootstrap procedure with a certain plug-in approximation for the IPD generating mechanism, given IPD empirical summaries. Here inferential recovery is understood as a form of IPD inference distribution estimation.

1.3 Outline

Figure 1.3, page 8, shows the main connection among the following Sections, with an emphasis on topic and goal. Section 2.1.2 to 2.1.5 have a rather theoretical tone, however, the scope remains quite practical. Section 2.1.2 deals with the theoretical reconstruction of a marginal IPD variable, by usage of basic MaxEnt formalism. Here, inter-marginals dependence is disregarded. Section 2.1.3 provides new contributions, extending the basic MaxEnt formalism, with the goal to reconstruct all IPD marginals jointly, also accounting for their inter-dependence structure. In Section 2.1.4 we argue that the MaxEnt IPD reconstruction is close to a realization from the joint IPD generating mechanism, under empirical IPD constraints. Section 2.1.5 studies connections to the bootstrap. We show MaxEnt resampling is akin to bootstrapping the IPD inference, leading to a good IPD inference reconstruction in some distributional sense. From Section 2.2 to 2.3 we implement all formalism into a practical methodology of IPD and IPD inference reconstruction. Section 3.1 to 3.3 empirically show that we can produce satisfactory IPD and IPD inference reconstructions, if enough prior IPD summary data is available. It follows a discussion, and some conclusions.

1.4 Extended outline

Section 2.1 introduces the mathematical notions to develop a method of IPD and IPD inference reconstruction. The starting point is a body of well established information theoretical results. Section 2.1.2 reviews basic definitions, and results, about the distribution with maximum entropy. Here, the Conditional Limit Theorem and Concentration Theorem seem to justify IPD marginal reconstruction, by sampling from a distribution with high entropy. That is, an IPD marginal variable, given its first k moments, converges to the maximum entropy distribution (Conditional Limit Theorem), or to a distribution of approximately equally high entropy (Concentration Theorem), that comply to such moment constraints. Here the limitation is that, typically, IPD must be handled jointly, and dependence between IPD marginals is important. We denote with P^* the multi-variate MaxEnt distribution, with arbitrary marginals, and dependence structure. To the best of our knowledge, two aspects yet lack enough attention in the literature:

to conveniently sample from P^*.

to verify that P^* is conditional limit of a IPD generating mechanism.

In Section 2.1.3 we first propose a solution to sample from a multi-variate MaxEnt distribution, based on Gaussian copulas, that avoids variational calculus altogether. Our resampling proposal is based on two relatively trivial premises. First, the NORmal To Anything (NORTA) transformation (Definition 2.1.3, page 17), is a Gaussian copula inversion allowing to sample from arbitrary multivariate distribution, with known marginals, and correlation matrix. Second, a Gaussian copula has maximum entropy among all joint distributions with given marginals and correlation structure (Proposition 2.1.9, page 2.1.9). It follows a NORTA transform with MaxEnt marginals must be a

draw from the multi-variate MaxEnt distribution with fixed correlation structure (Theorem 2.1.1 and Corollary 2.1.1.1, page 18 and 18). We conveniently denote such operation as NORTAmax. In Section 2.1.4 we argue NORTAmax resampling is asymptotically equal to draw from the IPD generating mechanism, given IPD empirical distributional summaries are available. The claim is based on a mode of convergence in information (Theorem 2.1.2 and Conjecture 2.1.1, page 19 and 20), that is our key justification for reconstruction of original IPD. Similarly, NORTAmax can be seen as a tool for IPD anonymization.

In Section 2.1.5, we acknowledge NORTAmax resampling is akin to IPD bootstrapping where the MaxEnt distribution is used as plug-in approximation for the IPD generating mechanism, under empirical IPD features constraints. We then define a MaxEnt bootstrap estimator for the distribution of an original IPD inference that is our key tool for IPD inference reconstruction, given empirical IPD summaries as only input data. We crudely assess MaxEnt bootstrap consistency (Conjecture 2.1.2 and Proposition 2.1.10, page 22), and conjecture the MaxEnt bootstrap average is a good predictor for the original IPD inference (Conjecture 2.1.3, page 23).

In Section 2.1.6, Algorithms 2.1.1 and 2.1.2 (page 25) give simple instructions to implement IPD and IPD inferences reconstruction, from the given IPD summaries. Sections 2.2–2.4 give further methodological implementations. In particular, we want to assess the reconstruction method under different amount of the input IPD summary data, especially, under varying completeness of the empirical IPD dependence structure. We further assess the method robustness, by using a MaxEnt surrogate for a continuous IPD marginal, that is given by the Johnson distributions system. We also assess the method on real IPD examples and design four IPD batches by different amount and type of included covariates. For simplicity, and without loss of generality, we decide to work with mostly four variables per IPD. To assess IPD inference reconstruction, we select commonly used inferential procedures among General Linear and survival models. We focus on GLMs with Gaussian, Binomial, and Poisson family, in addition to Cox regression followed by Breslow, or Nelson-Aalen estimation, depending on the number and type of covariates involved.

In Section 3.1 and 3.2 we give results on IPD and IPD inference reconstruction that generally confirm our arguments and conjectures. We show NORTAmax resampling produces honest simulations of the original IPD. From those IPD reconstructions we show we can well recover IPD MLEs, their 95% empirical intervals, as well as cumulative hazard estimates, by taking appropriate syntheses of the MaxEnt bootstrap sample for these IPD inferences. A good IPD inference reconstruction is generally due more to complete IPD dependence information, and less to marginal moments of degree greater than two. In Section 3.3 we give further practical examples on IPD inference reconstruction, when, for instance, the original IPD inference has no interpretation. In particular, in Section 3.3.2 we graphically show the MaxEnt bootstrap of the cumulative hazard estimate can typically result in a close approximation of the original IPD curve.

These results suggest the method could be meaningfully applied in disciplines like research reproduction and synthesis, given empirical IPD marginal moments and correlation matrix as only input data. In particular, the method seems readily suited for distributed network computing, upon compliance of the network to the above proposed IPD summaries. Chapter 4 is devoted to interpretations and discussions of the produced methodological and empirical materials.

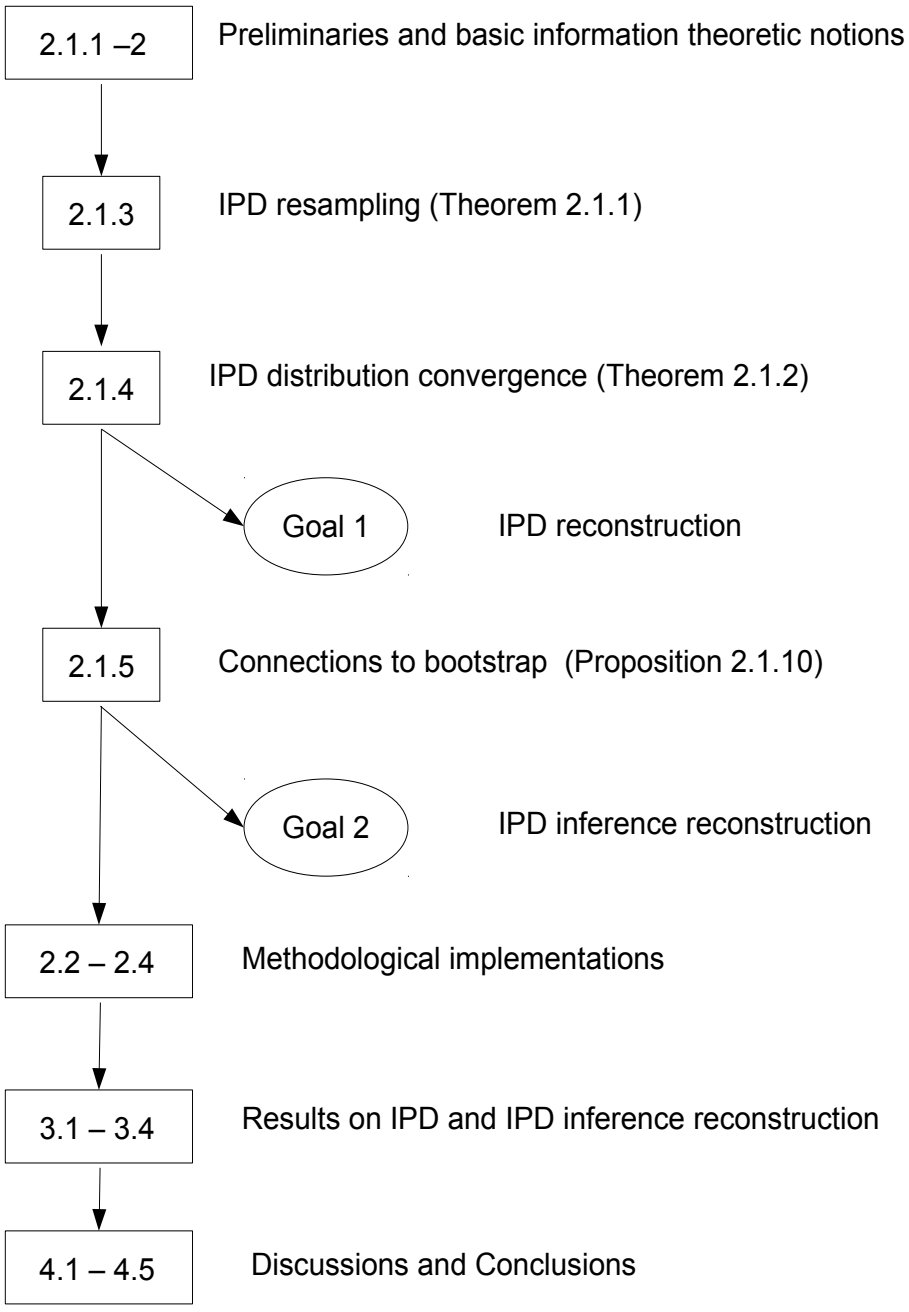

Figure 1.1: Flowchart displaying a graphical outline of the remaining parts of the manuscript. Boxes represent Sections with a short description of their content. In parentheses are some of the main statements. Ovals represent the goals.

Chapter 2

Theory and Methods

2.1 Mathematics

Consider an IPD, defined as n independent observations of a p-dimensional dependent record. For example, we collect a multi-variate medical record (age, height, health-status, etc ...) on several unrelated patients. Imagine a situation where we have no IPD, nor IPD inference, but only key summaries of the original IPD, namely its empirical marginal moments and correlation matrix. Here we propose a formal argument for a method of IPD and IPD inference reconstruction, when only such empirical IPD summaries are available. Further applications are discussed later. In the sequel x denotes a $n \times p$ matrix, $x_{i\cdot}$ a p-dimensional vector, $x_{\cdot j}$ a n-dimensional vector, x_{ij} a scalar, and capital letters are used to denote random counterparts. Throughout the text convergence is understood for $n \to \infty$. I provide a summary of all relevant notions of this Section on page 27.

2.1.1 Preliminaries

Consider an observed numerical $n \times p$ matrix, otherwise referred to as IPD,

$$x = \begin{bmatrix} x_{1\cdot} \\ \vdots \\ x_{i\cdot} \\ \vdots \\ x_{n\cdot} \end{bmatrix} = \begin{bmatrix} x_{\cdot 1} & \dots & x_{\cdot j} & \dots & x_{\cdot p} \end{bmatrix}, \tag{2.1}$$

with row record $x_{i\cdot} = (x_{i1}, x_{i2}, \dots, x_{ip})$, $i = 1, 2, \dots n$, and column vector $x_{\cdot j} = (x_{1j}, x_{2j}, \dots, x_{nj})^\top$, $j = 1, 2, \dots p$. Let stress that $p < n$. The sign $\top$ indicates transpose. There is nothing special about matrix (2.1): it arises after collection of p (dependent) variables, independently observed across n statistical units. Trivially, we allow $y \geq 1$ variables to represent an outcome, whose relation with the remnant $p - y$ covariates is the scope of further statistical investigation. For instance, (2.1) could arise after observation of y clinical outcomes along, or after, collection of $p - y$ medical records, on

n unrelated individuals. We stress again observations on the individual statistical unit are typically dependent. Generic examples in other social and scientific disciplines are licit.

For the time being, we focus on the situation where IPD (2.1) is not available. Instead, assume only a certain compression of IPD (2.1), denote it the IPD signature, $C_n(x) = \bar{C}_x$, is available. Let

$$\bar{C}_x \coloneqq \bar{m}_j^k \wedge \bar{R}_x; \qquad j = 1, \ldots, p; \quad k = 1, 2, \ldots, \tag{2.2}$$

where $\bar{m}_j^k$ and $\bar{R}_x$ denote empirical marginal moment of degree k and correlation matrix of IPD (2.1). Symbol $\wedge$ denotes logical conjunction. Our goal is to reconstruct the unavailable IPD (2.1), when only its signature (2.2) is available. We justify this idea more formally.

Let assume IPD (2.1) is the realization of a random matrix

$$X = \begin{bmatrix} X_{1\cdot} \\ \vdots \\ X_{i\cdot} \\ \vdots \\ X_{n\cdot} \end{bmatrix} = \begin{bmatrix} X_{\cdot 1} & \ldots & X_{\cdot j} & \ldots & X_{\cdot p} \end{bmatrix}, \tag{2.3}$$

where $X_{i\cdot} = (X_{i1}, X_{i2}, \ldots, X_{ip})$ is a p-dimensional random vector with $\mathrm{Cov}(X_{ij}, X_{ik})$ not necessarily zero, $\forall\, j \neq k$, and $X_{\cdot j} = (X_{1j}, X_{2j}, \ldots, X_{nj})^\top$ is a n-dimensional random vector with $X_{ij} \perp\!\!\!\perp X_{\ell j}$, $\forall i \neq \ell$. Let fix an arbitrary measurable space $(\mathcal{X}_j, \mathscr{B}_j)$ and let $(\mathcal{X}, \mathscr{B})$ be its p-product, $j = 1, \ldots, p$. We consider $X_{\cdot j}$ as a n-series of $\mathcal{X}_j$-valued i.i.d. rv's with common fixed distribution Q_j in some some sample space $(\mathcal{X}_j, \mathscr{B}_j)^n$. We consider $X_{i\cdot}$ as a p-dimensional $\mathcal{X}$-valued rv with fixed distribution Q, and we regard X as a n-series of p-dimensional rv's, i.i.d. according to Q from some sample space $(\mathcal{X}, \mathscr{B})^n$. Denote with $\mathcal{P}_j$ and $\mathcal{P}$ the set of all pm's on $(\mathcal{X}_j, \mathscr{B}_j)$ and $(\mathcal{X}, \mathscr{B})$ respectively. The series $F_{n,j}(X_{\cdot j}) \in \mathcal{F}_j \subset \mathcal{P}_j$ and $F_n(X) \in \mathcal{F} \subset \mathcal{P}$ is the empirical distribution (e.d.) of $X_{\cdot j}$ and X, respectively.

In the next section we review known stochastic properties of the n-series $X_{\cdot j}$ when $F_{n,j}$ is in some convex subset $C_j \subset \mathcal{P}_j$. Additional to convexity set C_j can be made to satisfy certain moment conditions of the form $m_{n,j}^k = \sum_i^n X_{ij}^k$, $k = 1, 2, \ldots$. This seems particularly useful in the framework we operate. Results we first present are well known in the information theoretic community, and often find application in the financial, physical, and engineering sciences, but not so much in the broader statistical community, especially in the bio-sciences. I show these information theoretic results suggest a sound approach to reconstruct a marginal IPD variable, $x_{\cdot j}$, if only some of its first empirical k-moments are observed. Here, the joint structure of x is ignored, and a good reconstruction of the joint IPD is not possible, unless all IPD marginals are independent. As a further and main contribution we consider the non zero dependence structure of the joint IPD (Section 2.1.3), and address the stochastic properties of the latter (Section 2.1.4). This leads to a practical approach for reconstruction of the joint IPD, as described in Section 2.1.7. As a further contribution we investigate the connections of this reconstruction procedure to the bootstrap in Section 2.1.5, which leads to a method of IPD inference reconstruction (Section 2.1.6)

with application to some empirical statistical disciplines, such as SDC, research reproduction, and synthesis .

2.1.2 Marginal IPD distribution given IPD moments

Let the n-series $X_{\cdot j}$ be i.i.d. according to Q_j. The main known result from information theory is that

$$\Pr\left(X_{ij} \in B_j \quad | \quad X_{\cdot j} \in \mathcal{A}_j\right) \to P_j^*, \quad B_j \subset \mathcal{X}_j, \tag{2.4}$$

the probability distribution of X_{ij}, given $X_{\cdot j}$ is in $\mathcal{A}_j = \left\{X_{\cdot j} : F_{n,j} \in C_j\right\}$, $\mathcal{A}_j \in \mathscr{B}_j^n$, converges to the Gibbs measure P^* in information. Since probability (2.4) must satisfy conditions of set C_j, let us focus on all pm's in such set $F_j \in C_j$.

A result like (2.4), sometimes known as Conditional Limit Theorem, or Gibbs Conditioning Principle, is the formal basis for our proposed approach to reconstruct IPD (2.1). In the sequel, we shall further elaborate on (2.4).

For a pm $P_j \in \mathcal{P}_j$ and Q_j (fixed) with density dP_j and dQ_j we need the following.

Definition 2.1.1. (Information divergence). The Kullback-Leibler (KL) divergence of P_j relative to fixed Q_j, $P_j << Q_j$ is

$$D(P_j \| Q_j) = E_{P_j} \log(dP_j/dQ_j). \tag{2.5}$$

Equation (2.5) is a convex function in P_j, always positive, zero iff $P_j = Q_j$, where E denotes expectation. Minus (2.5) is also known as relative entropy, or information. Let $C_j \subset \mathcal{P}_j$ be a convex set of pm's intersecting a "neighbourhood" of Q_j of possibly infinite radius, $\left\{P_j : D(P_j \| Q_j) \leq \infty\right\}$.

Definition 2.1.2. (D-projection of Q_j on C_j). By strict convexity of (2.5)

$$P_j^* = \arg\min_{C_j} D(P_j \| Q_j), \tag{2.6}$$

is the closest pm to Q_j and is unique. It follows $D(P_j^* \| Q_j) = \min_{C_j} D(P_j \| Q_j)$.

Remark. A sufficient condition for existence of P_j^* is that subset C_j is closed to the variational distance; see Csiszár (1975), Theorem 2.1, page 148 (there the term I-projection is used analogously). We shall more precisely refer to a generalized D-projection (Csiszár, 1984) below, but for the technical level adopted here we hereinafter neglect this detail.

Divergence (2.1.1) is not a metric but it plays the role of squared Euclidean distance between P_j and Q_j.

Proposition 2.1.1. (Csiszár (1975)). If $D(P_j^* \| Q_j)$ and $D(P_j \| P_j^*)$ are both finite, we have

$$D(P_j \| Q_j) \geq D(P_j \| P_j^*) + D(P_j^* \| Q_j), \tag{2.7}$$

where $P_j \in C_j \cap \left\{P_j : D(P_j \| Q_j) < \infty\right\}$.

Proof. See proofs of Lemma 2.1, Theorem 2.2, and Theorem 2.3 of Csiszár (1975). A simple proof is given in Cover and Thomas (2006), Theorem 11.6.1, page 367. □

Remark. The usage of Proposition 2.1.1 is the following. If it can be found a series $D(P_{n,j}\|Q_j) \to D(P_j^*\|Q_j)$, then 2.7 holds with equality and $D(P_{n,j}\|P_j^*) \to 0$.

The next important notion relates the probability of observing a series in C_j to the distance of the closest element to Q_j.

Proposition 2.1.2. (Sanov Theorem).

$$(1/n)\log \Pr(F_{n,j} \in C_j) \to -D(P_j^*\|Q_j), \tag{2.8}$$

where probability is measured relative to Q_j.

Proof. See for instance Sanov (1958), or Csiszár (2006), or Theorem 11.4.1, page 362 (Cover and Thomas, 2006) for a simplified proof. □

Remark. The meaning of Proposition 2.1.2 is that pm's far from P_j^* seem highly unlikely in C_j.

In the sequel let adopt the following specification

$$C_j = \left\{ F_j : \int_{\mathcal{X}_j} x^k dF_j \geq a_j^k, \quad k = 1, 2, \ldots \right\}. \tag{2.9}$$

Hence we make set C_j to have specific moment features. Accordingly $\mathcal{A}_j = \left\{ X_{\cdot j} : \frac{1}{n}\sum_i^n X_{ij}^k \geq a_j^k \right\}$.

Remark 2.1.1. For specification (2.9) the D-projection (2.6) of Q_j has form

$$dP_j^*/dQ_j = \exp\left(\sum_k \lambda_{jk} x_j^k \right) c, \quad k = 1, 2, \ldots, \tag{2.10}$$

where c is a normalizing constant. For Q_j equal the Uniform measure, density (2.10) is the general form of a maximum entropy (MaxEnt) distribution, where parameter vector λ must be chosen according to $E_{P_j^*} X_j^k = a_j^k$ on domain $\mathcal{X}_j$. We shall hereinafter assume a solution for λ_j exists, implying existence of (2.10).

Vasicek (1980) first proves convergence in probability for a simplified version of (2.4). His result can be read as "the empirical frequencies of a finite support rv $X_{ij} \sim Q_j$ converge to the MaxEnt frequencies having form (2.10), for a given mean value of the n-series $X_{\cdot j}$". The author calls this result a conditional law of large numbers as compared to the more famous unconditional version, as another mode of interpreting limiting frequencies as probabilities. This result is in some respect simple but clear. Van Campenhout and Cover (1981) are apparently the first to prove (2.4) in probability, for both discrete and continuous rv's, under a condition comparable to (2.9). In the words of the authors the given empirical mean is a sufficient statistic for the n-series $X_{\cdot j}$. The conditional limit argument is greatly generalized by Csiszár (1975, 1984). Here, we mainly use his

results. Particularly easy proofs of these results are later formulated via the so called method of types (Csiszár, 1998) under the more restrictive premise of finite alphabets. Dembo and Zeitouni (1996, 1998) recast the conditional limit phenomenon under rigorous measure theoretic arguments, and under the name of Gibbs conditioning principle. A later proof of (2.4) in less general settings is given by Grünwald (2001).

We shall now restate (2.4) as the following simplified statement from Csiszár (1984).

Proposition 2.1.3. (Conditional Limit Theorem). Let the n-series $X_{\cdot j}$ be drawn i.i.d. $\sim Q_j \notin C_j$, with e.d. $F_{n,j} \in C_j$. Let P_j^* be as in Definition 2.1.2. Then

$$D(F_j \| P_j^*) \to 0, \tag{2.11}$$

All pm's $F_j \in C_j$ converge to the Gibbs distribution in information. Otherwise said, a sample X_{ij} with e.d. in C_j has quasi-independent elements with common limit distribution P_j^*.

Remark. With C_j as in 2.9 Proposition 2.1.3 means all pm's with given moment features converge to the MaxEnt distribution P_j^* having density (2.10) and with Q_j being the Uniform distribution. Interestingly, the effect of conditioning seems to be negligible on the limit (see comments of Van Campenhout and Cover (1981) in proof of their Theorem 1, page 484), since (2.4) converges to the unconditional density form (2.10).

Proof. It shall mainly follow from Proposition 2.1.1 and 2.1.2. See Csiszár (1984) (in particular Theorem 1 and 4 with corresponding proofs). An accessible and intuitive proof is given in Cover and Thomas (2006) (Theorem 11.6.2, page 371), although on simplified premises, namely countable finite support for the rv. □

Remark. KL convergence is convergence in information, and implies convergence in total variation (DasGupta (2008), Chapter 2).

Remark 2.1.2. (IPD marginal reconstruction) Usage of Proposition 2.1.3 is the following. We want to retrieve an inaccessible IPD marginal variable, $x_{\cdot j}$, knowing only some empirical moments of it, $\bar{m}_j^k = a_j^k$, $k = 1, 2, \ldots, \forall j$. To do so we can sample from the MaxEnt distribution, P_j^*, that is the limit of the IPD marginal distribution, given its empirical moments. The input data for this program is the observed constraint a_j^k only. For a marginal IPD variable putting non negligible information at $k > 1$, knowledge of higher order moments should improve reconstruction of $x_{\cdot j}$. For a continuous not normally shaped series, knowledge of moments up to the fourth degree should be sufficient. This approach would reconstruct each IPD marginal variable, with no regard to inter-variables dependence. This is unsatisfactory in most cases.

The so called concentration phenomenon is an intuitive and useful result, although weaker than Proposition 2.1.3. The premise for the following is the rv having finite support, and notation below is to be understood in such context.

Proposition 2.1.4. (Concentration Theorem). In the set C_j, for $n > n_0$, such that $\bar{m}^k_{n_0,j} \doteq \hat{k}_j$, the following holds

$$\Pr\left(F_{n,j} \in \mathscr{H}(P^*_j)\right) = 1 - e^{-nc}, \tag{2.12}$$

where $\mathscr{H}(F^*_j)$ is an open neighbourhood of P^*_j, and c is some constant. That is, with high probability, a realization of the e.d. is very close to the MaxEnt distribution, for increasing n.

Remark. Differently said, there are exponentially more ways for empirical frequencies to resemble those of the MaxEnt distribution, for growing n.

Proof. A first version of Proposition 2.1.4 is given and proved by Jaynes (1982) (see Rosenkrantz (2012) for archived material). Grünwald (2001) strengthens the Proposition by proving it for arbitrarily large $|\mathcal{X}_j|$, provided its ratio to n goes to 0. Another proof is given by Robert (1990). □

Remark. Proposition 2.1.4 mainly borrows notation from Robert (1990). Oikonomou and Grünwald (2016) give a stronger assertion about the concentration phenomenon, by providing explicit non-asymptotic bounds, while allowing for sample error in the linear constraint, and, in some instances, independence from $|\mathcal{X}_j|$.

Remark. The usage of Proposition 2.1.4 is the following. In order to reconstruct a marginal IPD series as described in Remark 2.1.2, one does not need to use the MaxEnt solution (2.10), but, instead, any other distribution of approximately equal entropy, for the same given constraints. We might use this statement if the rv's domain can be approximated to a finite one.

Both the Concentration and Conditional Limit Theorem seem to provide an eminent justification for usage of the marginal MaxEnt distribution for reconstruction of each IPD column, given corresponding IPD column summary only. However this does not allow for a joint IPD reconstruction that also accounts for between columns dependence. This latter aspect is addressed in the next Section.

2.1.3 Joint IPD reconstruction given IPD moments and correlation

In the previous Section, we see how basic information theoretic results justify usage of the MaxEnt marginal distribution to recover each marginal column of IPD (2.1). However, to obtain a reasonable joint IPD reconstruction we must account for between IPD marginals dependence. In this Section we offer one such joint IPD reconstruction, and in Section 2.1.4 we assess its stochastic properties.

Basic definitions of Section 2.1.2 extend to joint pm's. Denote the KL-divergence of any joint pm $P \in \mathcal{P}$ from a fixed joint Q as $D(P\|Q)$, strictly convex in P, and zero iff $P = Q$. The closest distribution to Q must be the p-dimensional D-projection P^* that is unique by convexity of $D(\cdot\|\cdot)$. If Q is the Uniform measure, P^* must be the p-dimensional MaxEnt distribution. We shall extend (2.9) and consider the intersection

$$C = \left(\cap_j^p C_j\right) \cap \mathcal{D}, \tag{2.13}$$

$C \subset \mathcal{P}$, where

$$\mathcal{D} = \left\{ F : \int_{X_j} \int_{X_\ell} x_j x_\ell dF \propto r_{jl}, \quad \forall j \neq \ell \right\}. \tag{2.14}$$

is a convex set made to satisfy specific pairwise dependence conditions. Accordingly we might consider all series X with correlation at least R. We consider the following probability

$$\Pr\left(X_{i\cdot} \in B \quad | \quad X \in \mathcal{A}\right), \qquad B \subset \mathcal{X}, \quad i = 1, \ldots n, \tag{2.15}$$

where $\mathcal{A} = \{X : F_n \in C\}$, $\mathcal{A} \in \mathscr{B}^n$. Accordingly let us focus on all p-dimensional pm's $F \in C$ (with given marginal and dependence features). One first challenge here is to identify a convenient form for P^*. Second we stress that a result analogous to (2.4) is not available for (2.15).

To first determine the form of P^* one could try KL-minimization by variational calculus (Hiai and Petz, 1998; Ebrahimi et al., 2008; Mansoury and Pasha, 2008; Larralde, 2012). Here we can by-pass this step by adopting an alternative mode of dependence modeling. The idea is to use a copula (Nelsen, 2006) to link the MaxEnt marginals into a joint distribution. If the copula is Gaussian, we show this mode of construction must yield the MaxEnt joint distribution, which seems not fully recognized in the literature. We first need the following.

Let $V_{ij} \sim U(0,1)$ be independent of $X_{ij} \sim G_j$, with one-dimensional distribution function G_j and generalized inverse

$$G_j^{-1} = \inf\left\{x_{ij} : G_j(x_{ij}) > u_{ij}\right\}, \quad u_{ij} \in (0,1). \tag{2.16}$$

Using notation from Oertel (2015), define the modified distribution function as

$$G_j(x, \lambda) \coloneqq (1-\lambda) G_j(X_{ij} < x_{ij}) + \lambda G_j(X_{ij} = x_{ij}), \tag{2.17}$$

for $\lambda \in [0,1]$. Define the distributional transform of X_{ij} as

$$U_{ij} \coloneqq G_j(X_{ij}, V_{ij}). \tag{2.18}$$

Proposition 2.1.5. (Rüschendorf transform). Consider the transform (2.18), we have

$$U_{ij} \overset{d}{=} U(0,1), \qquad X_{ij} = G_j^{-1}(U_{ij}) \quad \text{a.s.} \tag{2.19}$$

Proof. See Rüschendorf (2009). See Oertel (2015) for a more detailed approach. □

Remark. From a probabilistic point of view, the usage of Proposition 2.1.5 seems that of producing a "smoothing" of an arbitrary rv X_{ij}. In particular we have useful results on the distribution of the transformed G_j and its inverse. Sometimes (2.18) is called the copula representer.

The fundamental copula identity is due to Sklar, see for instance Schweizer (1991) (page 17). We shall put emphasis on the distributional and probabilistic aspect of Sklar statement, that is sometimes given as a separate Lemma (see for instance de Amo et al. (2012), page 105).

Proposition 2.1.6. (Sklar identity). Let $X_{i\cdot}$ be a p-dimensional random vector with marginal and common joint distribution respectively G_j, $j = 1, \dots, p$, and G. Then for a p-dimensional copula K,

$$G(x_{i1}, \dots, x_{ip}) = K(G_1(x_{i1}), \dots, G_p(x_{ip})) \tag{2.20}$$

is a p-dimensional joint distribution function with marginal G_j, $\forall j$.

That is, a copula $K : [0,1]^p \mapsto [0,1]$ can be written as

$$K(u_{i1}, \dots, u_{ip}) = G(G_1^{-1}(u_{i1}), \dots, G_p^{-1}(u_{ip})), \tag{2.21}$$

or as a Uniform p-dimensional vector $U_{i\cdot} = G(X_{i\cdot})$ for which it holds $X_{i\cdot} = G^{-1}(U_{i\cdot})$, that is relevant for random variate generation. From the latter we see (2.21) cannot be a copula if G is non-continuous (Genest and Nešlehová, 2007).

Remark. If G_j, $\forall j$, is continuous then K is unique, otherwise it is unique only on $\prod_j^p \mathrm{Ran}(G_j)$, that is on the product image grid induced by the discrete support. Outside the range there are many K satisfying (2.20).

The importance of Proposition 2.1.5 is to extend the copula identity to non-continuous rv's, such that (2.21) holds for arbitrary rv's. The following is a convenient property of (2.20).

Proposition 2.1.7. (Copula density factorization). If marginal G_j has density g_j, it holds

$$g(x_{i1}, \dots, x_{ip}) = k(G_1(x_{i1}), \dots, G_p(x_{ip})) \cdot \prod_j^p g_j(x_{ij}), \tag{2.22}$$

where g is the joint density and k is the first derivative of copula K.

Proof. Compute $dK(G_1(x_{i1}), \dots, G_p(x_{ip}))/dx_{i1} \cdots dx_{ip}$, applying chain rule. See also Kotz and Seeger (1991) (section 4, page 120). □

Remark. Also see Fang et al. (2002) for copulas admitting (2.22).

In (2.22) g_j is defined only iff $dG_j(x_{ij})/dx_{ij}$, $\forall j$, is defined. This raises an issue if G_j is non-continuous. For the relevant role Proposition 2.1.7 later plays, we shall introduce two other continuation statements which might further help generalize identity (2.21). To ease exposition we shall omit unnecessary detail. Consider a smoothing transformation $\tilde{G}_m$ of an otherwise arbitrary joint distribution G, $m \in \mathbb{N}$. In other words $\tilde{G}_m$ consists of a convolution after application of a kernel (mollifier) on G. We refer to Durante et al. (2012) for more details. It can be shown that both $\tilde{G}_m$ and its marginals are continuous, and admit derivatives of any order. Moreover for $m \to \infty$ $\tilde{G}_m$ tends to G. Thus, a non-continuous G can be approximated by a continuous version $\tilde{G}_m$. Through identity (2.20), the copula associated with $\tilde{G}_m$ is also continuous which can be convenient especially in respect to Proposition 2.1.7. In essence, we are interested in the following aspect from Durante et al. (2012).

Proposition 2.1.8. (Copula smoothing). Let $\tilde{G}_m$ be a continuous approximation of an arbitrary G, as described above. Then an associated copula identity of the form (2.20) holds.

Proof. See proof to Theorem 4.7 of Durante et al. (2012). □

Remark. We have in mind the following use of Proposition 2.1.8. For some non-continuous G, apply a smoothing transformation making the associated copula continuous and invariant to the original one. Then exploit properties of a continuous copula.

Remark 2.1.3. A similar continuation technique is given in a probabilistic setting by Faugeras (2013), who considers the smoothed random vector $Y_m = Y + h_m T$, with Y, T arbitrary and continuous random vectors respectively, and $h_m \in \mathbb{R}$ tending to zero. Such setting amount to fix an arbitrary kernel in Durante et al. (2012), hence shall need less assumptions. From Y_m Faugeras (2013) shows the associated copula rapresenter $U_m \to U$ in distribution, with U non necessary continuous. In fact, using a specific construction (2.18) with randomizer $V_{ij} \perp\!\!\!\perp V_{ik}$, $\forall j \neq k$, the author can resolve indeterminacy issues related to a discrete copula (see Faugeras (2017), section 2.3, page 124, and Faugeras (2015)).

Our main focus is on to generate a random IPD, X, by resampling i.i.d. rows from a common multivariate distribution with certain marginal and dependence structure – see Tiit (2002); Iman and Conover (1982); Cuadras (1992); Rüschendorf (1985) for a sparse selection of theoretical and applied related works. An easy method to sample from an arbitrary joint distribution is to use the copula inversion identity, $X_{i\cdot} = G^{-1}(U_{i\cdot})$ when the linker is Gaussian. This procedure is also known as NOrmal To Anything (NORTA) transformation (Li and Hammond, 1975; Cario and Nelson, 1997). Assume a p-variate model $X_{i\cdot} \sim G$, with marginal $X_{ij} \sim G_j$, $\forall j$, and $\mathrm{Cor}_G(X_{i\cdot}) = R_x \in \mathcal{R}_x$. Let $Z_{i\cdot}$ be a p-variate standard normal (s.n.) vector, with marginal Z_{ij}, and $\mathrm{Cor}(Z_{i\cdot}) = R_z \in \mathcal{R}_z$. Define marginal $X_{ij} = \Psi(Z_{ij})$, where $\Psi = G_j^{-1}(\Phi(\cdot))$, and Φ is the cumulative s.n. distribution. This is similar to inverse probability weighting (i.p.w.), since $\Phi(Z_{ij}) \sim U(0,1)$.

Definition 2.1.3. (NORTA transformation). If G_j^{-1} is defined, we have that

$$E(X_j, X_\ell) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Psi_j(z_1)\Psi_\ell(z_2)\phi(z_1, z_2|\rho_{j\ell})dz_1 dz_2, \tag{2.23}$$

is proportional to the $j\ell$-th off-diagonal entry of R_x, where $\phi(z_1, z_2)$ is the bivariate s.n. density with correlation $\rho_{j\ell}$, the off-diagonal element of R_z, $\forall j \neq \ell$. Given marginal G_j, $\forall j$, and matrix R_x, do:

1) map R_x to R_z by solving (2.23) for $\rho_{j\ell}$, $\forall j \neq \ell$. Obtain R_z.

2) draw $Z_{i\cdot} \sim \mathcal{N}_p(0, R_z)$, where N_p is the p-variate 0 - R_z Normal distribution.

3) obtain $X_{i\cdot} = \Big(\Psi_1(Z_{i1}), \ldots, \Psi_j(Z_{ij}), \ldots, \Psi_p(Z_{ip})\Big)$.

Row vector $X_{i\cdot}$ is drawn $\sim G$, with marginal, G_j, $\forall j$, and correlation R_x. Independently repeat step 2) to 3) for n trials.

Remark. The NORTA transformation is a quantile inversion technique based on a Normal copula with parameter R_z. It is a straightforward method to sample from an arbitrary multi-variate distribution with given marginals and correlation structure.

Remark 2.1.4. If G_j, $\forall j$, is continuous we have analytic solution $\rho_{j\ell} \equiv 2 \sin \pi r_{j\ell}/6$. If any G_j is discrete, an analytic solution is no longer available and (2.23) must be solved numerically. In this latter case a solution for $\rho_{j\ell}$ is unique, because there is only one Gaussian copula taking value $\dot{\rho}_{j\ell}$ at fixed point $\dot{r}_{j\ell}$. That is, we can constrain the Gaussian copula to be unique by fixing the dependence structure. In this latter case it is not assured R_z is semi-positive definite (Ghosh and Henderson, 2002, 2003), that seems related to instances of model impossibility raised by Faugeras (2015) (section 3.3, page 126). We shall further elaborate on this issue in later methodological sections.

The following is a relevant fact for our exposition.

Proposition 2.1.9. (Jansen, 1997) Let $X_{i\cdot} \sim G$, for some $G \in \mathcal{G}$, with continuous invertible marginal G_j, $\forall j$, and correlation $R_{z(x)} = R_z$. Then, the transformation $\Psi\left(\mathcal{N}_p(0, R_{z(x)})\right)$ has maximal entropy among all p-dimensional distributions with marginal G_j, $\forall j$, and correlation $R_{z(x)}$.

Proof. See Jansen (1997), which is basically a consequence of Cover and Thomas (2006) (Theorem 8.6.5, page 254). □

Remark. The importance of Proposition 2.1.9 is to recognize the Gaussian copula is a joint distribution of maximum entropy among those with same marginals and correlation structure (just invert the quantile identity). This is relevant because it suggests we could define a joint MaxEnt distribution with dependence constraint with no need of variational calculus altogether.

To the best of our knowledge, the following is yet not fully acknowledged.

Theorem 2.1.1. (Joint MaxEnt distribution, with dependence constraint). Let $u^*_{ij} = P^*_j(x_{ij})$ the marginal distribution defined on (2.10) with Q_j Uniform, $\forall j$. Then,

$$\Phi_{R_z}\left(\Phi^{-1}(u^*_{i1}), \ldots, \Phi^{-1}(u^*_{ip})\right) = P^*(x_{i1}, \ldots, x_{ip}), \tag{2.24}$$

the Gaussian copula Φ_{R_z} with dependence parameter R_z is the p-dimensional joint MaxEnt distribution P^*.

Proof. It shall follow from Proposition 2.1.5, 2.1.6, 2.1.9, and Definition 2.1.3. The Rüschendorf transform applied to P^* here ensures extension of Proposition 2.1.5 to discrete marginals. By Sklar identity and Proposition (2.1.9), the Gaussian copula linking MaxEnt marginals must define the joint distribution of maximum entropy.

□

Corollary 2.1.1.1. (Multi-variate MaxEnt resampling: NORTAmax) Let $X^*_{ij} = P^{*-1}_j\left(\Phi(Z_{ij})\right)$ and $\mathrm{Cor}_Q(X_{i\cdot}) = \bar{R}_x$ be given. Apply steps 1) to 3) of Definition 2.1.3, where the map is $\bar{R}_x \mapsto R_z$. Then, $X^*_{i\cdot}$ is drawn from P^*, the p-variate MaxEnt distribution with marginal P^*_j and correlation $\bar{R}_x$, $\forall j = 1, \ldots, p$.

Proof. Just consider the probabilistic version of (2.21) and the following quantile inversion identity. □

Remark. Alternatively said, the NORTA transformation of Definition 2.1.3 with MaxEnt marginal P_j^* and correlation $\bar{R}_x$ yields a sample $X_{i\cdot}^*$ from the p-variate MaxEnt distribution. We denote this transformation as NORTAmax.

Remark. See Clemen and Reilly (1999) for more advanced usages of the Gaussian copula Φ_{R_z}.

In the next Section we shall assess stochastic properties of (2.24). The goal is to reach a statement analogous to (2.4) in the joint non-independent case (2.15). If this occurs, it would suggest a NORTAmax n-draw X^* should tend to recover most information about an unobserved IPD x (plus-minus a resampling error) under an empirical condition of form (2.2).

2.1.4 Joint IPD distribution given IPD moments and correlation

We shall be here concerned with stochastic properties of the joint MaxEnt distribution (2.24), hence of the NORTAmax quantile inverse of Corollary 2.1.1.1. We shall first consider the case of continuous marginals. Let first identify (2.15) with all pm's F in set (2.13), that is all p-dimensional distributions of $X_{i\cdot}$ with given marginal and dependence features. The main idea is to represent all such pm's via their corresponding Gaussian copula.

Theorem 2.1.2. (Limit of (2.15) with continuous marginals). Consider the joint D-projection $P^* = K(P_1^*, \dots, P_p^*)$ as in (2.24), with $K(\cdot)$ Gaussian and P_j^* as in (2.6), continuous, with density p_j^*, and Q_j fixed Uniform. Similarly consider $F = K(F_1, \dots, F_p)$, $F \in C$, $F_j \in C_j$, assumed continuous with density f_j, $\forall j = 1, \dots p$. Then we have

$$D(F\|P^*) \to 0. \tag{2.25}$$

All p-dimensional pm's $F \in C$ (with given marginal and dependence features) converge to the joint p-dimensional MaxEnt distribution (2.24) in information.

Proof. By continuity of all marginals, F and P^* are uniquely identified by the respective Gaussian copula K. We further consider first order derivable marginal distributions. Then by Proposition 2.1.7 we have

$$D(F\|P^*) = D(\Phi\|\Phi^*) + \sum_j^p D(F_j\|P_j^*), \tag{i}$$

where $\Phi = K(F_1, \dots, F_p)$ and $\Phi^* = K(P_1^*, \dots, P_p^*)$. The right summand in (i) goes to zero by Proposition 2.1.3, $\forall j$. The Gaussian copula K has maximum entropy by Proposition 2.1.9, and, since $F_j \to P_j^*$, $\forall j$, we have $\Phi \to \Phi^*$, thus $D(\Phi\|\Phi^*) \to 0$, which makes (i) equal zero. □

Remark. Let X be i.i.d. according to unknown joint pm Q. An approximation of the latter is given by P^* via copula (2.24) under condition $F_n \in C$.

Corollary 2.1.2.1. (limit to NORTAmax sample). We have $X \to X^*$, where X^* is a NORTAmax n-draw according to Corollary 2.1.1.1.

Proof. Again, this follows by the copula quantile inversion identity. □

Remark. The message of Corollary 2.1.2.1 is quite strong. A data matrix X with a certain description (2.13) converges to a NORTAmax draw in information.

Remark 2.1.5. (Joint IPD reconstruction). The usage of Theorem 2.1.2 is the following. We can recover most of unobserved IPD information, x, (plus-minus a white error) via NORTAmax resampling, using the observed IPD marginal moments and correlation matrix only. That is, we can reconstruct the IPD by resampling from the limit IPD joint distribution under the above summary constraint. Similar considerations to Remark 2.1.2, page 13, are here holds as well. Convergence is proved when all marginals are continuous but there are strong hints this should hold generally for any marginal.

Remark 2.1.6. (Extension to mixed discrete-continuous marginals). Since there are issues with discrete copulas (Section 2.1.4), a natural question is if Theorem 2.1.2 holds when some marginals are not continuous. This question maybe requires a careful and extensive answer that cannot be the scope here. However, using Proposition 2.1.8 a possible solution could be based on approximating a discrete copula with a smoothed version.

Conjecture 2.1.1. (Extension of Theorem 2.1.2 to non-continuous marginals). Relation (2.25) holds if not all marginals are continuous.

Following from Proposition 2.1.8, this Conjecture could be formally proved by application of a smoothing transformation upon arbitrary F and P^*.

Remark. For the time being we shall adopt Conjecture 2.1.1 as a statement of reasonably approximate validity. Empirical validation of this statement follows in later Sections.

MacKenzie (1994); Dempster et al. (2007); Zhao and Lin (2011); Chu (2011); Piantadosi et al. (2012); Bedford and Wilson (2014); Butucea et al. (2018) variably study maximization of copulas via classic constrained optimization, seemingly not considering the role of the Gaussian copula as we do here. Manomaiphiboon et al. (2008); Singer (2010) focus on non copula based joint density estimation. While not explicitly considering IPD reconstruction, Ponomareva et al. (2015) and Miller and Liu (2002) focus on data simulation but respectively use a sort of deterministic method and an information theoretic approach different than ours. None of the mentioned authors seem to focus on limiting properties of the MaxEnt distribution, and some of them use factorization (2.22) for other purposes. Singh and Zhang (2018) seem one of the few to recognize a link between MaxEnt copulas and a concentration phenomenon, but do not consider this aspect thoroughly, and not in respect to the conditioning limit principle. Thus, Theorem 2.1.2 seems one of the first to acknowledge the possibility to prove a Conditional Limit Theorem in the multivariate dependent case, by using copulas and property (2.22) which is of independent interest here.

In this Section we tried to give a probabilistic and information theoretic justification for usage of a MaxEnt based resampler to retrieve an unobserved IPD, given a summary description of it is

available. In the next Section we see how the MaxEnt representation of an unobserved IPD can be used to retrieve statistical inferences from that IPD.

2.1.5 Connections to the bootstrap

NORTAmax resampling is akin to permuting with repetition from the MaxEnt distribution (2.24). Hence, MaxEnt quantile inversion of Corollary 2.1.1.1 can be seen as the basis for a bootstrap procedure. Here we build a bootstrap estimator for the distribution of an unavailable IPD inference based on P^* as defined in (2.24), given IPD empirical signature (2.2) as only input data, and we try to assess its consistency. First, we briefly review some key bootstrap ideas.

Consider the sampling distribution of a functional $\mathscr{M}$ on the random n-series X, i.i.d. according to a fixed unknown Q,

$$\mathcal{I}_n = \Pr\left(\mathscr{M}(X) \leq t|Q\right), \tag{2.26}$$

that can be otherwise seen as the distribution of a random quantity $\mathscr{M}_n(X, Q)$. Distribution (2.26) is unknown because Q is not known. A typical way to approximate (2.26) is by asymptotic theory. For instance, $\mathcal{I}_n$ might be approximated using the Central Limit Theorem for a certain choice of $\mathscr{M}$. A more general approach is to estimate (2.26) with

$$\mathcal{I}^*_n = \Pr_*\left(\mathscr{M}(X^*) \leq t|Q_n\right), \tag{2.27}$$

that is the *ordinary*, or classic, bootstrap estimator of $\mathcal{I}_n$, where Q_n is the e.d. estimate of Q and X^* is a resample from Q_n. Estimate (2.27) is based on the principle of substitution, where Q_n is the plug-in substitute of Q. The idea behind (2.27) is that the generating triplet $(Q, X, \mathscr{M}(X))$ is well mimicked by the (observable) substitute $(Q_n, X^*, \mathscr{M}(X^*))$. In fact, any sensible substitute of Q builds a bootstrap estimator for (2.26), but deserves attribute 'ordinary' only when the plug-in substitute is Q_n. To compute (2.27) would require generating all possible n^n resamples, with replacement, from Q_n and then count the relative frequencies of transform $\mathscr{M}(X^*)$. For n large this program is replaced by the Monte Carlo (MC) approximation

$$\mathcal{I}^*_B = \frac{1}{B}\mathbb{1}\left\{\mathscr{M}(X^*_1, \ldots, X^*_B) \leq t|Q_n\right\}, \tag{2.28}$$

for some B and indicator function $\mathbb{1}$. Hence, (2.28) is subject to two estimation errors, the first is due to the choice of a plug-in substitute for Q, and the second is a MC error. The latter is typically negligible since is controllable for $B \to \infty$. Thus, the major source of error in (2.28) is due to approximation of Q. The better we approximate Q, the closer a bootstrap estimate should be to (2.26). As customary we study bootstrap probabilities using form (2.27), not (2.28).

Performance of a bootstrap estimator is generally measured by the distance, $\delta(\mathcal{I}^*_n, \mathcal{I}_n)$, between $\mathcal{I}_n$ and $\mathcal{I}^*_n$, for some metric δ. A bootstrap estimator is said to be respectively strongly or weakly consistent, if $\delta \to 0$ respectively almost surely or in probability. These ways of convergence have been extensively studied and proved for (2.27) – for a compendium see Shao and Tu (1995), chapter 3, or DasGupta (2008), chapter 29.

We now want to consider the bootstrap estimator

$$\mathcal{I}_n^* = \Pr_*\left(\mathscr{M}(X^*) \leq t|P^*\right), \tag{2.29}$$

where P^* is the Gaussian copula (2.24) and X^* is a NORTAmax n-draw upon inversion of P^*. We shall naturally denote (2.29) as the MaxEnt bootstrap estimator. In some sense (2.29) looks as a parametric estimator, since the data model P^* is defined by the exponential model (2.10). On the other side P^* is not arbitrary and Theorem 2.1.1–2.1.2 suggest it as a reasonable substitute for Q, given empirical IPD summaries as only input data.

An extensive assessment of the consistency of (2.29) is out of scope here. We should focus on bootstrap consistency for the mean to get a clue of the performance of (2.29) in simple cases. Suppose $X_{1\cdot}, \ldots, X_{n\cdot} \overset{\text{i.i.d.}}{\sim} Q$ and $E(X_{1\cdot}^2) < \infty$. Let $\mathscr{M}_n(X, Q) = \sqrt{n}(\bar{X} - \mu)$ define distribution (2.26), and let $\mathscr{M}_n(X^*, P^*) = \sqrt{n}(\bar{X}_i^* - \bar{X}^*)$ define MaxEnt bootstrap $\mathcal{I}_n^*$ given in (2.29). Since IPD is unavailable we use bootstrap average $\bar{X}^*$ as an estimator for μ.

Consistency in uniform metric intuitively equals showing both the bootstrap estimator and its estimand statistic converge to the same distribution. For the standardized sample mean this a standard Normal distribution, but we can work with the centered sample mean as just introduced.

Proposition 2.1.10. (Strong consistency of (2.29) in $\mathcal{L}_\infty$ – Kolmogorov – metric).

$$\mathcal{L}_\infty(\mathcal{I}_n^*, \mathcal{I}_n) \overset{\text{a.s.}}{\to} 0.$$

Proof. Consider the last summand of the Kolmogorov distance given in DasGupta (2008), page 465. Adapting notation, we shall replace the term $E_{Q_n}|X_1^* - \bar{X}|^3$ with $E_{P^*}|X_1^* - \bar{X}^*|^3$, where $\bar{X}^* = \sum_i^n X_i^*$, under P^*, since the original average, $\bar{X}$, is not available under our setting. We only need to show that $\bar{X}^* \approx \bar{X}$, which indeed is, since by definition we have $E_{P^*}(X_1^*) = \bar{X}$. Hence the proof proceeds as for the ordinary bootstrap. □

A popular alternative is considering consistency in $\mathcal{L}_2$ metric. This requires some considerations on the plug-in estimator that can be more difficult under our setting.

Conjecture 2.1.2. (Approximate consistency of (2.29) in $\mathcal{L}_2$ – Mallow-Wasserstein – metric).

$$\mathcal{L}_2(\mathcal{I}_n^*, \mathcal{I}_n) \approx 0.$$

A proof of the above could be based on the following. By Theorem2.1.2, the MaxEnt copula (2.24) is the limit of the distribution of $X_{1\cdot}$ under condition C. That is, sample $X_{1\cdot}, \ldots, X_{n\cdot}$ given C has quasi-independent elements with common distribution P^*. Considering condition C is negligible on the limit, we have sample frequencies $P^* \approx Q_n$ in some sense, where Q_n is the e.d. estimate of Q. From that point a proof would proceed as for the ordinary bootstrap.

Remark. One needs a quantification of the approximation $P^* \approx Q_n$ to make the argument rigorous. This effort cannot be the scope here. Conjecture 2.1.2 stands on ground of a reasonable similarity between P^* and Q_n.

Remark. Bootstrap consistency generally implies consistency of the bootstrap CI at a pre-specified level (van der Vaart (2007), Lemma 23.3, page 329). If the bootstrap percentile is not studentized, symmetry of the bootstrap estimator is typically required.

Proposition 2.1.10 does not imply consistency of the bootstrap variance, which we shall deduce from Conjecture 2.1.2, if true. Hints on how to assess variance consistency in the former case might be found in Shao and Tu (1995), section 3.3.3, page 86, following the same mimicking approach used above. It is known the ordinary bootstrap (2.27) tends to underestimate the variance of $\mathscr{M}_n(X, Q)$. Intuitively, this occurs because all available permutations with repetition are constrained by the given observed data. In some respect, the MaxEnt bootstrap (2.29) should shrink the variance even further, because each resampled data-set must obey the same constraint on marginal moments and correlation matrix, which does not need to occur in (2.27).

The Delta method in conjunction with consistency of moment statistics is typically used to prove ordinary bootstrap consistency of other moment-based statistics, like the sample correlation. Functional versions of the above are exploited similarly. It is commonly accepted that, as a rule of thumb, consistency of a generic statistic should hold if the latter accepts a Central Limit Theorem. In our setting we have the additional difficulty that the original IPD inference is not available, and even arguing for the above rule seems difficult. Relevant for us is to show that our bootstrap average is generally close enough to the original IPD inference, in order for the former to substitute the latter. This also serves the more practical reconstruction purpose but it seems not trivial.

Remark 2.1.7. (IPD inference reconstruction). We see NORTAmax resampling as a form of bootstrapping procedure, by which we can reconstruct an unavailable IPD inference from the reconstructed unavailable IPD. We stress, again, the whole procedure takes simple IPD summaries as input data only. Proposition 2.1.10 and Conjecture 2.1.2 provide hints of MaxEnt bootstrap consistency in simple cases. Consistency might hold practically in more general situations, but this needs experimental verification. The use of (2.29) is that to estimate the distribution of an IPD inference from a generated ensemble of IPDs. In particular some central distributional measures seem important to recover an original IPD inference when the latter is not available. Imagine (2.29) is reconstructing the distribution of some unavailable IPD estimate of a Cox regression parameter. The role of the bootstrap distribution average is to predict the original IPD estimate. We shall further elaborate on this aspect below.

One of our main goals is to recover some original IPD inference value using the generated MaxEnt bootstrap distribution (2.29). Computation of the bootstrap average in a spirit similar to Breiman (1996) seems relevant.

Conjecture 2.1.3. (MaxEnt bootstrap average) Consider a specific IPD inference $\mathscr{M}(x)$ with IPD signature $C_n(x) = \bar{C}_x$, and the MaxEnt bootstrap (2.29). We have

$$E(\mathscr{M}(X^*_{1\cdot}, \ldots, X^*_{n\cdot})|P^*) \overset{n\to\infty}{\approx} \mathscr{M}(x), \tag{2.30}$$

the MaxEnt bootstrap average is roughly equal to the IPD inference.

Proof. (Sketch). Under condition C, defined by a constraint of type (2.2), $X_{1\cdot}^*, \ldots, X_{n\cdot}^*$ are quasi-independent with common limit distribution P^*. Let x be a realized value of $X_{i\cdot}^*$. By exchangeability, $E(\mathscr{M}(X^*)|P^*) \approx \mathscr{M}(x)$ as n grows, since any $X_{i\cdot}^*$ accumulates toward a common MaxEnt set, and x is a value from such set. Conversely one can see that x given $C_n(x) = \bar{C}_x$ tends to be a realized value from the MaxEnt accumulation set by means of Proposition 2.1.3, or more intuitively, but under more restricting setting, by Proposition 2.1.4. □

Remark 2.1.8. Using information theoretic arguments, $X^* \sim P^*$ tends to be a configuration almost indistinguishable from x under constraint $C_n(x)$. On the long run this amounts to (2.30). An exception to this argument could be when x is on the boundary of the MaxEnt accumulation set. For instance, an original IPD displays an unusual relative risk between a treatment and control group variable disproportionally larger in favor of the former. Then an average Hazard Ratio of the form (2.30) would represent a more stable, predictive, value relative to the ill-behaving original IPD estimate.

In practice, one uses the bootstrap approximation (2.28) and all arguments made so far apply to such estimate, since the MC error is controllable. In virtue of the constrained nature of (2.29) we might guess that a choice for B does not need to be large here. Intuitively this is because one resampled data-set cannot differ by much from the next one under (2.29), although any two realizations are "independent". In the next Section we give practical instructions to reconstruct IPD and IPD inferences.

2.1.6 IPD inference reconstruction: the MaxEnt bootstrap

In the following let refer to empirical matrix (2.1) as the observed IPD x with compressed signature $C_n(x) = \bar{C}_x$ as in (2.2). Hereinafter it helps to think at the IPD x as not (always) available. We assume only its signature $\bar{C}_x$ is available. In the following we use θ to denote a point or point-wise inference. A full MaxEnt bootstrap procedure would practically work as follows. As a first step we bootstrap the data, with the premise that the marginal MaxEnt distribution family $P_j^*(\lambda_j)$ is identified via constrained entropy maximization. Identification is always possible and the unique KL-solution has form (2.10), page 12, $\forall\, j = 1, \ldots, p$.

Algorithm 2.1.1. (Bootstrap the data, or IPD reconstruction – NORTAmax)

1) Solve $E_{P_j^*(\lambda_j)} X_j^k = \bar{m}_j^k$, for unknown vector λ_j, and fix P_j^*, $\forall\, j$.

   By Proposition 2.1.3 P_j^*, is the limit distribution of a marginal IPD given $\bar{m}_j^k$.

2) Invert copula rapresenter (2.24) to get IPD simulation $X_b^*, \quad \forall b = 1, \ldots, B$.

   By Theorem 2.1.1, (2.24) is the joint MaxEnt distribution P^* with marginal P_j^*, and correlation R.

   By Theorem 2.1.2 and Conjecture 2.1.1 X_b^* is a n-draw from the limit IPD p-distribution given constraint $\bar{C}_x$.

As a second step we bootstrap the IPD statistic by transformation of the bootstrapped IPD.

Algorithm 2.1.2. (Bootstrap the inference, or IPD inference reconstruction)

1) Transform the NORTAmax sample to obtain a statistic $\mathscr{M}(X_b^*) = \theta_b^*, \quad \forall b = 1, \ldots, B.$

   $\theta_1^*, \ldots, \theta_B^*$ is a MaxEnt bootstrap MC sample from (2.29) for statistic θ.

2) Compute a distributional index on MC sample $\theta_1^*, \ldots, \theta_B^*$.

   Proposition 2.1.10 and Conjecture 2.1.2 suggest 95% θ^*-CIs should be roughly comparable to those of a classic IPD bootstrap (2.27).

   By Conjecture 2.1.3 $\bar{\theta}^* \approx \mathscr{M}(x)$ where $\bar{\theta}^* = \frac{1}{B}\sum_b^B \theta_b^*$ and $\mathscr{M}(x)$ is the (unavailable) IPD inference value.

For the time being let conveniently refer to the MaxEnt bootstrap procedure of Algorithm 2.1.1 and 2.1.2 as the MaxEntBoot . Similarly to classic bootstrap MaxEntBoot is also collectively defined by two steps: first sample the data, second obtain samples of a statistic computed on the sampled data. A few differences, though, are in order.

- MaxEntBoot does not permute the original IPD.
- Instead each permutation is a draw from the limit conditional IPD distribution.

Based on MaxEnt convergence arguments the method should be relatively stable if $n \geq 100$. Permuting via the limit IPD distribution we expect MaxEntBoot to display less variation relative to classic bootstrap. Then $100 \leq B \leq 300$ could suffice, but this argument is informal.

2.1.7 MaxEnt bootstrap: possible applications

Now I cast the MaxEntBoot algorithm in light of the statistical practice. Figure 2.1, page 28, shows the general MaxEntBoot rationale.

Full IPD recovery and IPD analysis (re)production

The basic application of Algorithms 2.1.1 and 2.1.2 is when original IPD access is not possible but IPD summary disclosure is allowed. IPD summaries can be shared – for instance through a journal-report or its supplementary material – while protecting privacy. Here we purposely underlie the connection between evidence reproduction and SDC. The procedure of Figure 2.1 enables a summary data holder to recover full IPD information with good utility. Statistical results can be replicated and validated, new models and hypotheses can be formulated.

Meta-analysis, research synthesis, distributed computing

To generalize the scheme of Figure 2.1 from a single IPD to several ones is trivial, provided the IPD sources are mutually independent. This is the typical case when we consider unrelated study-reports or study-centers. Then application of Algorithms 2.1.1 for each study source seems potentially relevant for meta-analysis, research synthesis, and analysis across a distributed network. Here random or source-specific effects can be easily modeled after tagging the reconstructed IPD by source provenance. The data is then pooled and analyzed in one block. This procedure implicitly requires that each source has information on the same set of variables. But this requirement is hardly met if the study-sources do not follow a common protocol, that is typical if the study purposes are unrelated. In this case one collects only IPD information commonly shared across sources or try to impute missing variables under certain assumptions, and maybe borrowing strength from the available variables.

Missing data imputation

Imagine IPD is available, but it has Missing At Random (MAR) records. Denote with n_0 the original records size of the data. and with $n << n_0$ the actual records size after standard deletion of the missing units. To fill in missing data we apply Algorithms 2.1.1 substituting n with n_0. Then, missing values are directly imputed by the limit joint IPD distribution with given marginals and correlation structure.

2.1.8 Experimental goals

Via implement of Algorithm 2.1.1 and 2.1.2, page 24, our main practical goals are to

A. reproduce IPD from its marginal moments and correlation matrix,

B. reproduce IPD inferences from the reproduced IPD.

We want to assess how good our reproductions are relative to the original IPD information. That is, we assess the utility of our IPD reconstructions similar in spirit to Joshua et al. (2018).

SUMMARY. We argue for a method of IPD and IPD inference recovery, when only IPD marginal moments and correlation matrix is available (IPD summary constraints). Theorem 2.1.1 defines a method of IPD reconstruction (NORTAmax), which uses the joint MaxEnt distribution as key resampling mechanism. The latter is based on a Gaussian copula representation (2.24) that directly draws from the limit joint IPD distribution with given IPD summary constraints (Theorem 2.1.2 and Conjecture 2.1.1). NORTAmax resampling operates akin to bootstrapping the IPD. We introduce a MaxEnt bootstrap estimator (2.29) and give hints of its consistency (Proposition 2.1.10 and Conjecture 2.1.2). The average MaxEnt bootstrap works as an expected IPD inference value and it is close to the original IPD inference under certain conditions (Conjecture 2.1.3). Algorithm 2.1.1 and 2.1.2 implement these formal arguments and conjectures into instructions for IPD and IPD inference reconstruction. We give an immediate practical example of this procedure and briefly consider some of its possible applications.

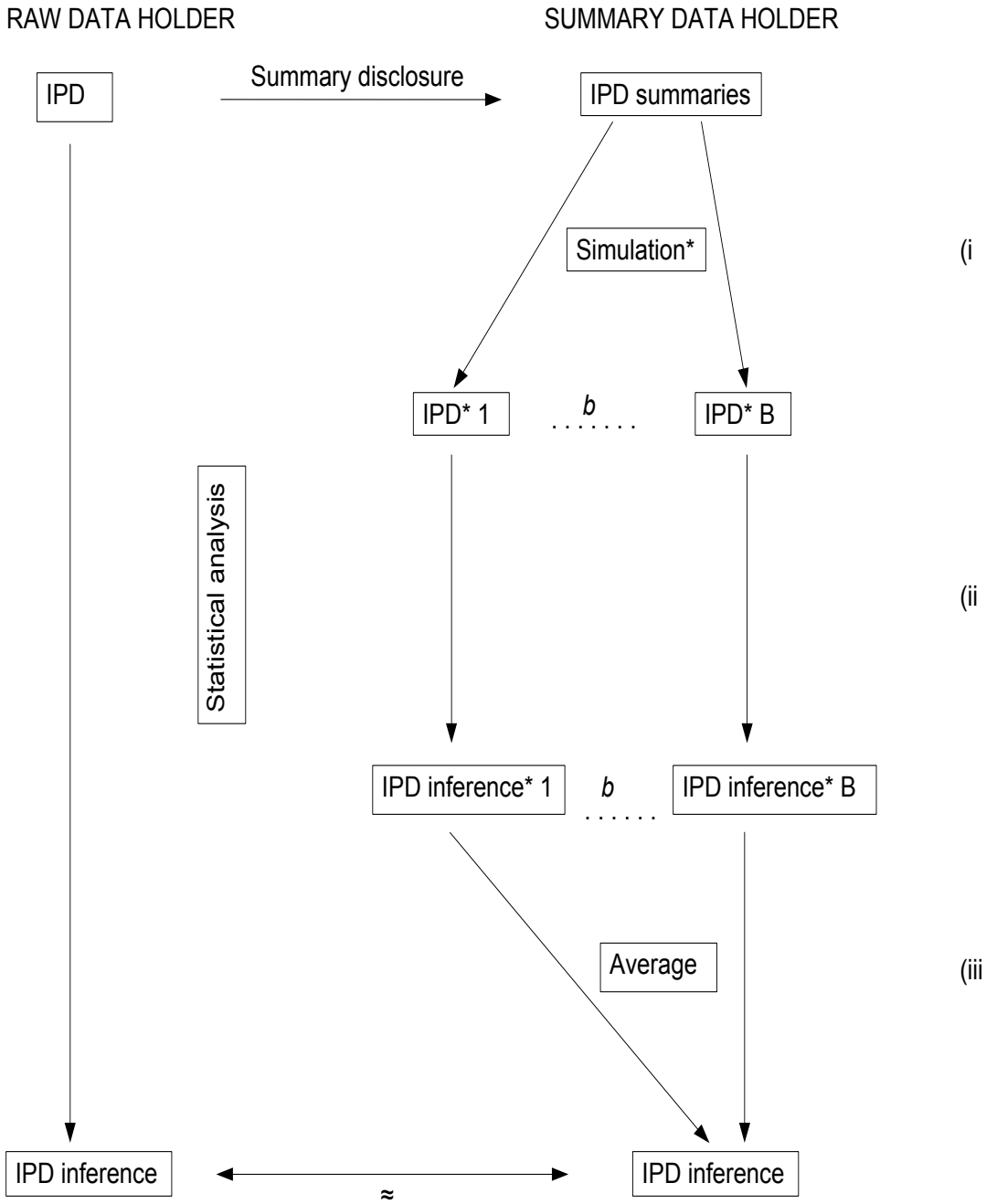

Figure 2.1: MaxEntBoot (general situation): on the left someone has the original raw IPD and discloses only some IPD summaries, like marginal moments and a correlation matrix. On the right someone has only access to the IPD summaries. At step i) the raw data is simulated ('*') into a b-th copy $b = 1, \ldots, B$. At step ii) a statistical analysis performable on the original IPD is performed on each IPD simulation to get a corresponding IPD inference simulation. At step iii) some sort of synthesis is taken on the IPD inference simulations, say an average. Bottom: the average IPD inference simulation is well comparable to the original IPD inference.

2.1.9 Example: IPD and IPD inference recovery from IPD summaries only.

To get a better understanding of the arguments so far introduced, we now give a practical example. We further elaborate on this example in Section 3.3 (page 59). Following Figure 2.1, page 28, a so called raw data holder has full access to some otherwise undisclosed IPD. An excerpt of the undisclosed IPD is showed in Table 2.1. This IPD is a selection of 4 variables from data `diabetes`, that records time to retinopathy in either eye of a diabetic patient. For each patient, treatment is randomly assigned to only one eye while the other eye is kept as control. Under such design we make the approximation that each patient eyes is an independent observation. The study cohort is composed of 197 patients and thus the data has 394 records.

Table 2.1: First and last six records of original IPD `diab.2`, a subset of four variables (time and status outcomes with two binary/continuous covariates) of data `diabetes` (see Section 2.4, page 44). This information is ideally not disclosed.

True IPD				
Record Nr.	*Time*	*Status*	*Treatment*	*Age*
1	46.250	0	1	28
2	46.276	0	0	28
3	42.507	0	1	12
4	31.341	1	0	12
5	42.301	0	1	9
6	42.274	0	0	9
389	50.010	0	1	33
390	2.911	1	0	33
391	45.957	0	1	3
392	1.488	1	0	3
393	41.973	0	1	32
394	41.986	0	0	32

Ideally the raw data holder discloses only a reduced version of the IPD of Table 2.1, in the form of IPD marginal moments (Table 2.2) and correlation matrix (Table 2.3). Through this compression phase all original 1576 IPD points are shrunk to only 22 points which makes the original information anonymous. We call someone having access to the summaries of Table 2.2 and 2.3 the summary data holder. The disclosed summaries convey some coarse distributional information on the observed joint IPD distribution. From Table 2.2 we see time to retinopathy is slightly shifted to higher values (negatively skewed, see 3rd moment) with a rather flat-shaped distribution (platykurtic, see 4th moment), while the condition is observed in about 40% of the cases. Treatment assignment is balanced while patients age is shifted to lower values (positively skewed, see 3rd moment) with a rather normal kurtosis. From Table 2.3 we see the condition is developed relatively

earlier in time (moderately strong negative correlation between status and time), whereas treated eyes seem to be less prone to develop the condition (moderate negative correlation beetween status and treatment).

Table 2.2: Known IPD constraints: Marginal empirical moments of the original data `diab.2`. This information is ideally disclosed.

	Given IPD Moments			
Moment's degree	*Time*	*Status*	*Treatment*	*Age*
1st	35.610	0.393	0.500	20.782
2nd	21.357	0.489	0.501	14.812
3rd	-0.113	0.436	0.000	0.808
4th	1.749	1.190	1.000	2.538

Table 2.3: Known IPD constraints: Empirical correlation lower triangular of original data `diab.2`. This information is ideally disclosed.

Given IPD Correlation				
	Time	Status	Treat.	Age
Time				
Status	-0.638			
Treat.	0.154	-0.244		
Age	-0.002	0.036	0.000	

The starting step for IPD recovery is access to summaries of Table 2.2 and 2.3. These summaries play the role of initial empirical constraints on the definition of a joint probability distribution for the IPD. The method of distribution reconstruction uses the principle of maximum entropy, that is a generalization of the Laplacian principle of indifference. In a sense, then, we follow a minimally presumptive path to information recovery.

Corollary 2.1.1.1 establishes an identity between NORTAmax resampling and drawing from the joint MaxEnt distribution. We use near maximum entropy densities (see Section 2.2.1, page 35, and Section 3.3.2, page 61) to describe the generating marginal law of variable 'Time' and 'Age', constrained on four moments. This yields the approximations of Figure 2.2, page 31, where we see the MaxEnt principle guides toward a good description of the data marginal distributions (see Proposition 2.1.4 and 2.1.3). All remnant binary variables ('Status' and 'Treatment') are appropriately described by a Bernoulli law with given mean that is the analytic MaxEnt solution here.

From Theorem 2.1.2 it follows that a NORTAmax sample asymptotically looks like a draw from the joint IPD distribution for given moments and correlations constraints. In order to explore enough data sample space, Algorithm 2.1.1 draws $1, \ldots, B$ IPD realizations from the joint MaxEnt

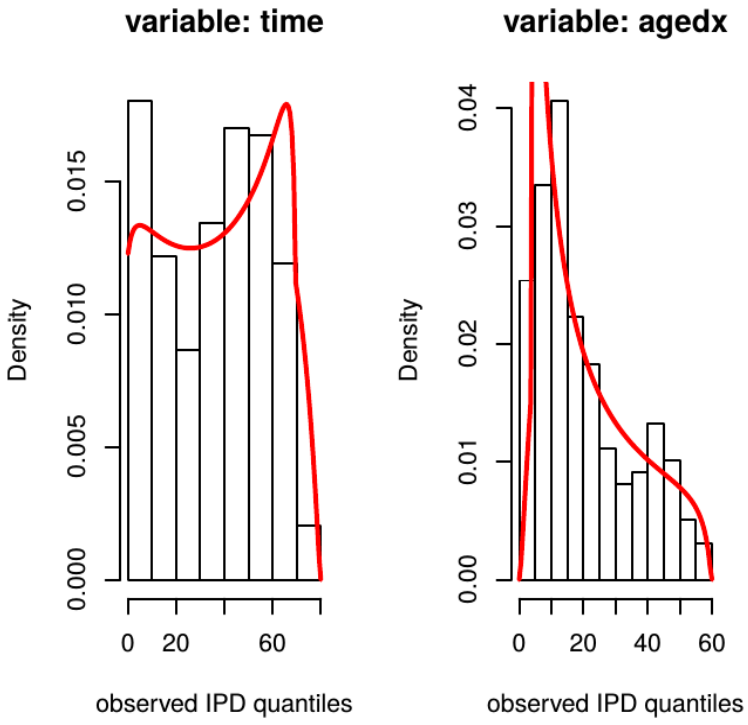

Figure 2.2: Agreement of the `diab.2` IPD continuous marginal, 'Time' and 'Age' (Table 2.1), to its respective near maximum entropy density (red line) constrained on four moments

distribution. To say we draw IPD from a distribution can be slightly confusing. The performed operation is more precisely akin to bootstraping the IPD.

We resample $B = 300$ IPD realizations with replacement. In our experiments, a simple diagnostic to verify if this bootstrap sample does well capture the original IPD features is to compare the original IPD summaries to the simulated ones. Here, we practically use MaxEnt bootstrap to estimate the distribution of the IPD first four sample moments and of each pairwise correlation. We compare the expectation estimate of each distributions to the respective original IPD summary value as shown in Table 2.4 and 2.5. We see good agreement everywhere that shows NORTAmax resampling can recover all original IPD distributional features on average (MC mean). This fact alone seems to support Proposition 2.1.10 and Conjecture 2.1.2. Later we further experimentally verify these arguments more systematically.

We can try to exploit the B bootstrap realizations not only to gain insights on simple statistics, like the IPD marginal moments and correlation, but in principle on any IPD inference. Here the next statistic of interest is the IPD log HR as estimated via a Proportional Hazards Cox model. The MaxEnt bootstrap tries to mimick the distribution of the log HR and other relevant log HR

Table 2.4: Moments for each marginal variable of data `diab.2` as estimated on true IPD or as an average (MC mean) across 300 simulated IPDs. The method of IPD simulation is that of Algorithm 2.1.1, page 24, that implements NORTAmax resampling.

		Variables Moments			
Moment's degree	Estimate	*Time*	*Status*	*Treatment*	*Age*
1st	IPD	35.610	0.393	0.500	20.782
	MC mean	35.613	0.393	0.502	20.831
2nd	IPD	21.357	0.489	0.501	14.812
	MC mean	21.313	0.489	0.500	14.846
3rd	IPD	-0.113	0.436	0.000	0.808
	MC mean	-0.099	0.439	-0.007	0.807
4th	IPD	1.749	1.190	1.000	2.538
	MC mean	1.736	1.203	1.009	2.538

Table 2.5: Lower triangular of `diab.2` correlation matrix as estimated on true IPD or as an average (MC mean) across 300 simulated IPDs. The method of IPD simulation is that of Algorithm 2.1.1, page 24, that implements NORTAmax resampling.

	Variables Correlations					
Estimate	*Time / Status*	*Time / Treat.*	*Time / Age*	*Status / Treat.*	*Status / Age*	*Treat. / Age*
IPD	-0.638	0.154	-0.002	-0.244	0.036	0.000
MC mean	-0.649	0.152	-0.007	-0.242	0.029	-0.000

functions like the log HR variance. In Table 2.6 we compare the expectation estimate of these log HR functions against their original IPD value.

The MaxEnt bootstrap average (MC mean) for the log HR of 'Treatment' and 'Age' well recovers the original IPD value. Here variable 'Treatment' is highly protective aginst the onset of retinopathy while 'Age' seems to play no role. Taking the square of the MC log HR standard deviation (MC sd) approximately recovers the original IPD log HR variance, indicating the the log HR MaxEnt bootstrap estimate might be approximately normally distributed. The IPD log HR variance is also accurately recovered on average indicating good preservation of likelihood information on the long run. Other recovered point estimates are the normally approximated 95% CIs, the log-likelihood maximum, and the AIC, all showing a good correspondence with the original IPD value. Here this seems to generally confirm Conjecture 2.1.3. In Table 2.7 we compare the 2.5th and 97.5th quantile of the log HR MaxEnt bootstrap estimate against different types of orid-

Table 2.6: Proportional Hazards Cox regression for the MLE of the log HR of 'Treatment' (β_1) and 'Age' (β_2), as estimated on the true `diab.2` IPD or as an average (MC mean) across 300 simulated MLEs. The method of MLE simulation from the simulated IPDs is that of Algorithm 2.1.2, page 25, based on NORTAmax resampling. HR: Hazard Ratio; Info. proxy: inverse of Fisher Information diagonal. Lower/Upper CI: $\hat{\beta} \pm 1.964 \sqrt{\text{Var}(\hat{\beta})}$; Log Lik.: $\ell(\hat{\beta})$. MC sd: standard deviation across all simulated MLEs

	Point Estimate									
	log HR		Info. proxy		Lower CI		Upper CI			
Estimate	$\hat{\beta}_1$	$\hat{\beta}_2$	$\text{Var}(\hat{\beta}_1)$	$\text{Var}(\hat{\beta}_2)$					Log Lik.	*AIC*
IPD	-0.782	0.004	0.029	0.00003	-1.113	-0.007	-0.451	0.015	-856.5	1717.0
MC mean	-0.768	0.003	0.029	0.00003	-1.100	-0.008	-0.435	0.013	-857.7	1719.4
MC sd	0.174	0.006	0.002	0.00000	0.182	0.006	0.166	0.005	50.3	100.6

inary bootstrap 95% CIs. The latter are computed on a 10000 repetitions of the original IPD. We see a tolerable agreement everywhere. To sum up we show NORTAmax resampling can recover IPD and IPD inferential information from disclosed key IPD summaries only. One key question is how complete the IPD summary information must be in order to yield good recovery performance, which we investgate in later sections.

Table 2.7: Empirical quantiles for the log HR of Table 2.6. MC denotes the HRs generated from the 300 simulated IPDs. The method of HR simulation from the simulated IPDs is that of Algorithm 2.1.2, page 25, based on NORTAmax resampling. Results compare the empirical quantiles of the MaxEnt bootstrap of size 300 against several types of non-parametric bootstrap CIs, as computed each on a 10000 sample of the original IPD. B: bootstrap sample size.

			Empirical Quantile			
			2.5th		97.5th	
Estimate	B	Bootstrap	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
MC	300	MaxEnt	-1.121	-0.008	-0.454	0.013
IPD	10000	normal	-1.115	-0.007	-0.433	0.015
		basic	-1.107	-0.007	-0.427	0.015
		percent	-1.138	-0.007	-0.457	0.015
		Bca	-1.123	-0.007	-0.443	0.015

2.2 Methods: IPD reconstruction

We give here methodological implementation of point A of Section 2.1.8, page 26, that is IPD reconstruction. We focus on implementation of Algorithm 2.1.1, page 24, while introducing different IPD reconstruction scenarios, based on the amount and completeness of the given IPD summaries. On page 39 I summary all relevant notions of this section.

Marginal binary information is always assumed fully recoverable (first moment is always given). Marginal continuous information can vary in amount by varying the empirical moment degree. Correlation information can vary in completeness. Two opposite scenarios are:

near-optimal: moments up to fourth degree and complete correlation.

sub-optimal: moments up to second degree and incomplete correlation.

While complete correlation knowledge allows an implementation of Algorithm 2.1.1 (NORTAmax), incomplete correlation knowledge forces us to devise an alternative reconstruction procedure. We derive some additional intermediate scenarios by combinations of the two above.

2.2.1 Near-optimal implementation of NORTAmax scheme

I give implementation details about NORTAmax resampling (Definition 2.1.3, page 17), especially about point 3) of Algorithm 2.1.1. Here the main challenge is to draw from each marginal MaxEnt distribution, $\sim P_j^*$, $j = 1, \ldots, p$, and to convert the empirical matrix into s.n. space.

NORTAmax: surrogates for MaxEnt marginals

In order to draw from the MaxEnt marginal distribution one has to first solve the general expression (2.10), page 12, for all IPD marginals involved. For $k < 2$ moment constraints (Tagliani (1993) claims for $k < 3$) the solution is analytic and otherwise must be numerical. In the latter case a solution for (2.10) should be guaranteed and at most be ϵ-achievable (Cover and Thomas (2006), page 415) with error margin $\epsilon \gtrsim 0$. A number of optimization schemes are proposed in Basu and Templeman (1984); Zellner and Highfield (1988); Ormoneit and White (1999); Rockinger and Jondeau (2002); Wu (2003); Holly et al. (2011).

Interestingly, by the phenomenon of empirical entropy accumulation (Proposition 2.1.4) usage of (2.10) does not have to be mandatory in practice. We resort to an alternative moment-based sampler, the Johnson distribution (Johnson, 1949), for practical generation of real-valued IPD marginals. The Johnson distribution is defined as a transformation from a standard Normal variate (see Appendix A.1.1, page 79) to a variable with the wished moment features. Since the standard Normal variate is the MaxEnt distribution in the real line, the Jonson transform should intuitively retain relatively high entropy. Denote the Johnson distribution with $\sim J(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$, where the first four parameters must be solved for given empirical moments constraints up to degree four (Hill et al., 1976; Hill, 1976), and α_5 specify the variable domain (log, unbounded, or bounded). Both Johnson density and the general MaxEnt solution (2.10) share an exponential

form related to a certain type of Pearson family distributions. Members of such family practically show very similar behaviours (Johnson, 1949; Siddall and Diab, 1975; Zellner and Highfield, 1988; Rockinger and Jondeau, 2002) that agrees with Proposition 2.1.4. See also Zografos (1999) for a more focused work on Pearson distributions entropy.

Under compliance to a common set of moment constraints, and for n large, a sample from the Johnson distribution should hardly be distinguishable from one from (2.10). Hence,

Stipulation 2.2.1. (Marginal MaxEnt distribution surrogate) By Proposition 2.1.4, page 14, we hereinafter use $\sim J_j$ as near surrogate for the MaxEnt distribution, $\sim P_j^*$, for all $j = 1, \ldots, p_0$ continuous marginal variables, with $p_0 \leq p$.

We stress all remnant non continuous variables (always binary) are drawn $\sim B(\pi_j)$, the Bernoulli distribution with proportion π_j that is the analytic MaxEnt solution here, $\forall = j = 1, \ldots, (p - p_0)$. The entropy of J_j shall only be at most near-maximal and often sub-maximal. Then usage of a Johnson marginal serves to assess the robustness of NORTAmax when some marginals might have less than maximal entropy.

NORTA: implementation of correlation conversion in s.n. space

The NORTA transformation is based on conversion of the IPD empirical correlation matrix $\bar{R}_x$ into s.n. space. Since the copula is Gaussian we let $\bar{R}_x$ be moment-based, which seems to improve numerical stability. We shall adopt this convention hereinafter but with no loss of generality. Clemen and Reilly (1999) describe one-to-one conversions from common correlation indexes to the product-moments one.

Xiao (2014) gives analytic correlation conversion in s.n. space when all marginals are Johnson. However since we mostly use mixed binary-continuous variables we must pursue a numerical solution based on (2.23), page 17, for each entry ρ_l of the s.n. correlation matrix lower triangular, $l = 1, \ldots, \frac{p(p-1)}{2}$. A Newton-Raphson scheme is enough to solve (2.23) as also suggested in Li and Hammond (1975); Emrich and Piedmonte (1991). In Appendix A.1.2, page 80, we give details about this procedure.

If the marginal inverses in (2.23) are defined a solution for ρ_l is unique. However the resulting s.n. matrix is not guaranteed to be semi positive definite (s.p.d.). In order to guarantee a s.p.d. result some authors suggest non-linear optimization routines, or semi-definite programming (Ghosh and Henderson, 2002). We decide to resort to a simpler approach, since non s.p.d. results seem not too common, and often require only minimal tweaking in order to be corrected. A description of the steps taken to avoid non s.p.d. results is in Appendix A.1.3, page 83.

Occasionally numerical integration of (2.23) may fail. As a stable alternative we can always resort to stochastic Monte-Carlo integration which is easy here since (2.23) has form

$$E_\phi \left[\Psi_j(z_{ij}), \Psi_\ell(z_{i\ell}) \right],$$

that is the double expectation of the transforms Ψ_j and Ψ_ℓ relative to the bivariate Normal density ϕ. This operation is described in more detail in Appendix A.1.2.

2.2.2 Sub-optimal and non NORTA resampling schemes

Imagine we have for a continuous marginals only limited empirical moments information. If only marginal mean and variance are available we use the Gamma distribution $\Gamma_j(a_j, b_j)$. The latter can be showed to be the closed-form MaxEnt solution in the positive real line given $k = 2$. Using moment estimation we set $a_j = (\bar{m}_j)^2/\bar{\sigma}_j^2$ and $b_j = \bar{m}_j/\bar{\sigma}_j^2$ where $\bar{m}_j$ and $\bar{\sigma}_j^2$ are the empirical mean and variance respectively. We can always use the Gamma marginal in the NORTAmax routine.

Furthermore imagine we can recover only few IPD empirical correlation pairs but not all of them. An incomplete correlation matrix forces us to impute missing correlation entries, or to implement a procedure different than NORTAmax. We adopt the latter strategy to assess IPD reconstruction robustness under incomplete IPD dependence information. To this end denote first-order correlation ς' that is the first row of the $p-1$ off-diagonal elements of correlation matrix $\bar{R}_x$ lower triangular. Let introduce the transformation

$$\mathscr{R}_{\varsigma'} : X^*_{\cdot 1}, \ldots, X^*_{\cdot j}, \ldots, X^*_{\cdot p} \mapsto X^*_{-}$$

where $\mathscr{R}_{\varsigma'}$ is a function parametrized in ς' taking uncorrelated MaxEnt marginals $X^*_{\cdot 1}, \ldots, X^*_{\cdot p}$ and returning an incompletely correlated $n \times p$ matrix X^*_{-}. Here we can impose at most first-order correlations, letting other entries freely vary. Hence for growing p the p-variate sample X^*_{-} it is more incompletely correlated.

We stress the map $\mathscr{R}_{\varsigma'}$ is not a NORTAmax transformation and the generated dependency structure does not need to give maximal entropy configuration to X^*_{-}. Next introduce the $n \times 2$ matrix $(Z)_b = \left(X^{*(0)}_{\ell}, X^{*(0)}_{j}\right)_b$ of completely uncorrelated elements, that is such that $\mathrm{Cor}(Z)_b \approx 0$ where $(X^{*(0)}_{\ell})_b$ is a fixed reference variable, for $\ell \neq j$ fixed, $j = 1, \ldots, (p-1)$, and $b = 1, \ldots, B$. Accordingly introduce the $n \times 2$ matrix $(U)_b = \left(X^{*(\max)}_{\ell}, X^{*(\max)}_{j}\right)_b$ of maximally correlated elements. We construct $(U)_h$ by sorting both variables in ascending order if $\mathrm{sign}(\bar{r}_{x_j x_\ell}) > 0$ and by sorting them in opposite order otherwise. We denote such extremal correlation as $\mathrm{Cor}(U)_b \approx \Re^{(b)}_{j\ell}$.

We design a choice for $\mathscr{R}_{\varsigma'}$ resulting in the $n \times 2$ matrix

$$(X^*_{-,j\ell})_b = (I_{\alpha_b} \circ Z)_b + (I_{1-\alpha_b} \circ U)_b \qquad j \neq \ell \quad , \tag{2.31}$$

where $\circ$ denotes row-wise Schur (element-wise) multiplication, I_{α_b} is a $n \times 1$ vectors of elements drawn i.i.d. $\sim B(\alpha_b)$, the Bernoulli distribution with parameter $\alpha_b = 1 - (\bar{r}_{x_j x_\ell}/\Re^{(b)}_{j\ell})$, and $I_{1-\alpha_b} = 1 - I_{\alpha_b}$, $\forall b$. In (2.31) matrix $(X^*_{-,\ell j})_b = \left(X^*_{-,\ell}, X^*_{-,j}\right)_b$ is the pair of re-ordered marginal columns such that their correlation is on average roughly equal to $\bar{r}_{x_j x_\ell}$. Next, column $X^*_{-,j}$ is appropriately re-merged with the remnant transformed $p-1$ columns and along with the fixed ℓ-reference, $\ell \neq j$. A more detailed description of this procedure is given in Appendix A.2.2, page 86, while derivation of (2.31) is described in Appendix A.2.1, page 85. We mention that in (2.31) we construct $(Z)_b$ via a brute-force search of zero correlation, $\forall b$. Hence we may refer to this method as a permutation search of an incomplete correlation structure.

2.2.3 IPD reconstruction: all simulation options

We can control the following simulation factors.

- Marginal distribution (continuous variable): Gamma ($k = 2$ moments) or Johnson ($k = 4$ moments).
- Type of correlation index: Pearson moment-based or Spearman rank-based.
- Type of correlation matrix: complete (NORTAmax transformation) or incomplete (based on permutation search (2.31)).

A binary variable is always drawn from a Bernoulli distribution with corresponding first moment ($k = 1$).

We will only focus on the following four simulation options:

1. Incomplete correlation matrix, Gamma marginal, rank correlation.
2. Incomplete correlation matrix, Johnson marginal, rank correlation.
3. Complete correlation matrix, Gamma marginal, moment correlation.
4. Complete correlation matrix, Johnson marginal, moment correlation.

Only the last two options define a NORTAmax transformation. In point 1-2 we use rank correlations for an easier solution of (2.31). Differently in point 3-4 we use moment-based correlations, the natural choice for the bivariate standard Normal distribution used in (2.23).

2.2.4 Comparison with IPD: bias definition

We want to assess how well the IPD simulation X_b^*, for $b = 1, \ldots, B$, overall retrieves information on the original IPD x. To this end we can check how the simulated IPD summaries compare to the original IPD constraint on average. We refer to the constraint (2.2), page 10.

Denote with $m_j = a_j/a_j^2$ and $\bar{m}_j^* = \bar{a}_j^*/\bar{a}_j^{2*}$ the normalized original and average first moment respectively, where $\bar{a}_j^{k*}$ is the average marginal moment simulation and a_j^k is the original moment, $k = 1, 2, \ldots$. Denote with $\bar{r}_l^*$ the average correlation simulation relative to original IPD correlation, r_l, $l = 1, \ldots, \frac{p(p-1)}{2}$. We adopt the following convention $\forall j$. An IPD sample is a good overall approximation of the original IPD if all the following criteria are met:

1. $\|\bar{m}_j^* - m_j\| \leq 0.7$,
2. $\|(1/\bar{m}_j^*) - (1/m_j)\| \leq 0.5$,
3. $\|\bar{a}_j^{3*} - a_j^3\| \leq 1$,
4. $\|\bar{a}_j^{4*} - a_j^4\| \leq 1$,

5. $\|\bar{r}_l^* - r_l\| \leq 0.05$, and $\text{sign}(\bar{r}_l^*) = \text{sign}(r_l)$, $\quad \forall l$.

In our experiments we practically use sample standard deviation, skewness, and kurtosis for second, third, and fourth moment. Hence point 1 and 2 consider the coefficient of variation and its reciprocal. As an alternative approach we can also graphically inspect how well the used MaxEnt marginal density describes the original IPD marginal empirical frequencies.

SUMMARY. In this section we give practical methods to implement IPD reconstruction from its summaries only. We define four different reconstruction scenarios based on the amount and completeness of given IPD summaries. Best and worst scenarios are respectively:

moments up to fourth degree and complete correlation knowledge.

moments up to second degree and incomplete correlation knowledge.

In the first point we use a so-called Johnson distribution that we use as surrogate for the MaxEnt distribution based on four moments. In the second point we use the Gamma distribution with given mean and variance. While complete correlation knowledge allows NORTA implementation, incomplete knowledge does not and we propose an alternative permutation-based procedure that does not guarantee maximum entropy configuration. We define criteria to assess if the IPD reconstruction well approximates the original IPD.

2.3 Methods: IPD inference reconstruction

Here we give methodological implementation of point B of Section 2.1.8, page 26, especially on point 1) and 2) of Algorithm 2.1.2, page 25, that is IPD inference reconstruction from reconstructed IPD. To avoid confusion we sometimes use a dot cap to distinguish between an IPD inference, or variable, and its simulation. For the reader who wish to skip the following material we provide a summary of this Section on page 43.

We should generally denote with X_b^*, $b = 1, \dots, B$, an IPD simulation from Algorithm 2.1.1, page 24, or Algorithm A.2.1, page 86, as reflecting simulation options Section 2.2.3, page 38. Where needed we stress the distinction. Here we compute a generic statistic $\theta_b = \mathscr{M}(X_b^*)$ on each IPD simulation. The main scope of this section is the overall comparison between IPD inference simulation $\mathscr{M}(X_b^*)$, $\forall b$, and the original IPD value $\mathscr{M}(x)$. As explained in Section 2.3.2 and 2.3.3 we can exploit a partial sufficiency property for a specific definition of $\mathscr{M}$.

2.3.1 Inferences considered in our experiments

We now specify the type of IPD statistical inference $\mathscr{M}(\cdot)$. We consider the following Maximum Likelihood Estimate (MLE).

> GLM (family): linear regression slope (Gaussian), risk ratio (Poisson), Odds Ratio (OR, Binomial).
>
> Cox model: Hazard Ratio (HR).

Time adjustment in Poisson regression can occur by introduction of a log-time offset. We also focus on the MLE variance, which we use as a proxy for the Fisher Information. The MLE variance is the diagonal of minus the inverse of the Hessian, as simply Fisher Information. See Efron and Hinkley (1978) for usage of minus the Hessian as a working Fisher Information. Translation from MLE variance to Fisher Information is one-to-one when the MLE is scalar. Otherwise, we clearly imply a simplification of the Fisher Information appraisal, by considering only its diagonal.

Remark. Fisher Information or its proxy can inform on the amount of IPD information retained by the log-likelihood. See Barron (1986); Mukherjee and Ratnaparkhi (1986); DasGupta (2008) for more connections to entropy.

For what concerns interval estimation of the MLE, a primary focus is on the 2.5th and 97.5th quantile of the MLE simulation as compared to an ordinary bootstrap for it. For what concerns Nelson-Aalen or Breslow estimates the primary endpoint remains acquisition of the cumulative hazard graph, or of key summaries of it, like inter-quartile indexes of each axis.

We also consider Nelson-Aalen or Breslow estimates for the cumulative hazard accompanying the Cox HR. If there are competing time-events outcomes, the reconstruction focus remains on the cause-specific cumulative hazard that entirely specify other functionals of interest, like transition probabilities. For all chosen models except Nelson-Aalen estimation data-reduction properties enable us to devise different resampling schemes, based on the amount of knowledge about the log-likelihood numerator.

2.3.2 Model-induced partial sufficiency

All statistical models introduced in Section 2.3.1 except Nelson-Aalen estimation allow some forms of data compression in the log-likelihood

$$\ell(\beta|X) = \beta^\top s - \mathscr{F}(X, \beta), \tag{2.32}$$

where $s = Z^\top y$ is vector-valued reduced data, Z is a $n \times (p-1)$ sub-set of covariates, y is an $n \times 1$ outcome, $X = (Z, y)$ is the full data, and $\mathscr{F}(X, \beta)$ is a normalizing term allowing none or little reduction. The general agreement is that most of statistical information lies in numerator s which we denote as a partially sufficient statistic (p.s.s.). We can exploit this reduction phenomenon to use alternative resampling schemes as shown below. A method that tries to incompletely reconstruct Nelson-Aalen as well as Cox risk-sets is described in Appendix B.5.1, page 95.

2.3.3 Inference reconstruction: simulation options

We distinguish between a log-likelihood evaluated on the original IPD x or on the IPD simulation X^*_b, $\forall b$. Denote the former and latter as $\ell(\beta|x)$ and $\ell^*_b = \ell(\beta|X^*_b)$. Accordingly denote outcome/covariate separation as $x = (Z, y)$ and $X^*_b = (Z^*_b, y^*_b)$. In our experiments the IPD log-likelihood $\ell(\beta|x)$ does not exploit form (2.32). On the contrary we can exploit property (2.32) to handle ℓ^*_b. Letting $\mathscr{M}(X^*_b) \equiv \max_\beta \ell(\beta|X^*_b)$ we may choose between the following options to execute step 1) of Algorithm 2.1.2.

1. Ordinary resampling: let $s^*_b = (Z^*_b)^\top y^*_b, \quad \forall b$.

2-3. Resampling with stochastic or deterministic adjustment: let $\bar{s}^* = \sum_b^B s^*_b$, or $\dot{s} = \dot{Z}^\top \dot{y}, \quad \forall b$.

By "ordinary" in point 1 we intend the p.s.s. varies from an IPD realization to another. Instead in points 2-3 we compute an expected p.s.s. estimate or use an original IPD p.s.s., if available, and fix them constant throughout all B repetitions. In point 2 and 3 the merit of introducing some sort of resampling adjustment is the following. For n small, or for a sub-optimal data simulation, the data sample X^*_b can be far from an IPD conditional distribution draw, introducing bias. Then we can try to reduce information loss by using a less noisy likelihood numerator. Under near-optimal resampling conditions (Section 2.2.1, page 35) we should expect $\bar{s}^* \to \dot{s}$ and point 2 may differ little from 3. To which extent such adjustment might improve on point 1 remains to be verified empirically.

2.3.4 Reconstruction: point estimates

Let $\dot{\beta} = \max_\beta \ell(\beta|x)$ be the original IPD MLE inference, and $\beta^*_b = \max_\beta \ell(\beta|X^*_b)$ be the simulated one, $b = 1, \dots B$ according to options of Section 2.3.3. We further consider functions of the MLE $\iota(\beta)$. The focus is in acquisition of the MLE simulated sample and on basic descriptions of it.

Here we consider the simulated sample average. If all the conditions satisfying Conjecture 2.1.3, page 23, are in place we should observe

$$\frac{1}{B}\sum_{b}^{B}\iota(\beta_b^*) \approx \iota(\dot{\beta}). \tag{2.33}$$

If X_b^* is obtained via Algorithm A.2.1 some columns may be missing and appropriate adjustments on the pair (s_b^*, Z_b^*) might be needed (see Appendix A.2.2, page 86). We shall generally accept occasional instability of $\max_\beta \ell_b^*$ due to random variability from one IPD simulation to another one. Especially if variability is non-white noise and X_b^* is a sub-optimal simulation, then $\max_\beta \ell_b^*$ could yield singular or unrepresentative values for β_b^*. In this case as part of a post-processing of the statistic simulation we could need to discard outlying values for $\iota(\beta_b^*)$ (see Appendix B.2, page 93). Similarly Fisher Information may not be available because the Hessian in not invertible, and these outputs are discarded. Because $\ell(\beta|x)$ is evaluated in a standard fashion we can use standard routines like `glm` or `coxph` when computing the MLE on original IPD x. Instead a specific implementation of (2.32) is needed (see Appendix B.1.1, and B.1.2, page 89).

2.3.5 Reconstruction: 95% empirical CIs quantiles

We extract the 2.5th and 97.5th quantile (95% empirical CIs) of the MLE simulated sample. This is similar to Efron percentile method where no statistic studentization and pivoting for quantile retrieval is used. This is motivated by MLE approximate Normality that yields similar results for both Efron and studentized quantiles. Next we compare our quantiles to a number of classic IPD bootstrap CIs (normal, basic, percent, or Bca). By Proposition 2.1.10 and Conjecture 2.1.2 we expect rough resemblance bethween the ordinary and MaxEnt bootstrap quantile, $\dot{\beta}_q$ and β_q^*, $q = 0.25, 0.975$. To compute ordinary quantiles we use module `bootstrap` and `bootci` from `R` package `boot`. We use 10000 and 100 (or 300, see Convention 2.4.1, page 45) resamples for the ordinary and MaxEnt bootstrap respectively.

2.3.6 Reconstruction: Nelson-Aalen / Breslow type estimates

In order estimate the cumulative hazard we generally employ the Breslow estimator (see for instance equation 4.17, page 141 of Aalen et al. (2008)). See Appendix B.3, page 93 for more computational details. The Nelson-Aalen estimator is obtained as special case.In the following we shall refer to both types of estimate as simply the cumulative hazard estimate (c.h.e.). The appropriate distinctions are made where needed. Denote with $\dot{A}(t)$ and $A_b^*(t)$ the original IPD and simulated c.h.e. respectively, for $b = 1, \ldots, B$. We see the generated c.h.e. simulations as approximating the ensamble of a cumulative hazard process with expectation estimate of the form $\bar{A}^*(\bar{t}^*)$, where $\bar{A}^*$ and $\bar{t}^*$ denote average cumulative event-counts and event-times respectively. For more computational details see Appendix B.4, page 94. By Conjecture 2.1.3 we expect to see $\bar{A}^*(\bar{t}^*) \approx \dot{A}(t)$.

2.3.7 Comparison with IPD estimates: bias definition

To evaluate reconstruction bias we take a simple difference between original and reconstructed IPD inference. We measure point estimate bias with the difference between the left and right term of (2.33). Similarly we assess empirical 95% CI bias with the difference beetween the IPD ordinary bootstrap quantile and our corresponding simulation. We assess c.h.e. reconstruciton bias by the difference between and IPD and simulated compound summary. In the latter case the summary is computed directly on the expected c.h.e.. The compound summary includes axis-specific first four quartiles, the range, and the mean. Here quartiles have a natural time-ordinal interpretation for each axis, given the cadlag property of the c.h.e..

SUMMARY. In this section we give practical methods to implement IPD inference reconstruction from reconstructed IPD. We focus on reproduction of MLE from a

linear, Poisson, and Logistic regression model.

proportional, or constant, hazards Cox model.

We focus on 95% empirical CIs of the simulated MLE and compare them to classic IPD bootstrap CIs. We further focus on reproduction of Nelson-Aalen or Breslow estimates. We propose three MLE simulation strategies based on how the compressed likelihood numerator is handled:

Random numerator.

Average numerator.

Original IPD numerator.

In order to compare the reconstructed point MLE to its IPD reference we use the average MLE simulation. Similarly we propose a procedure to recover an average cumulative hazard simulation. We define bias measures to assess the difference between the original and reconstructed IPD inference.

2.4 Methods: Data examples

In our experiments for IPD and IPD inference reproduction we use original IPD examples. We also review all main simulation options.

2.4.1 original IPD examples

We use IPD examples from a number of `R` packages or from Royston and Sauerbrei (2008) and from kind concession in one case. A list of 20 data-sets with their respective origin can be found in Table B.1 (Appendix B.6.1, page 96. The sample sizes from this data list ranges between 23 and 17260 with a median of 415 records.

Below we introduce some re-arrangements we make on the IPD in order to simplify our experiments.

2.4.2 IPD usage and re-arrangement

We define the original IPD format used in our experiments.

Stipulation 2.4.1. (Original IPD format) The IPD (2.1) is a numerical matrix derived from the transformation $\mathscr{D}(\dot{x}) = x$ where $\mathscr{D}$ maps to the standard design matrix. That is, all categorical variables with category l are converted into $(L-1)$ binary contrasts relative to reference level $l = l'$, for $l = 1, 2, \ldots, L$.

We adopt the above stipulation for convenience and without loss of generalization. To see this consider the Multinomial distribution is the analytic MaxEnt solution for a categorical variable with L categories, that would replace all $L - 1$ binary marginals above.

The number of data variables in our original IPDs bank ranges from few to many dozens, but we want to work with fewer variables for mere convenience. Then in our experiments we use IPD with mostly two, three, or four variables. For each IPD of Table B.1 we select only a reduced number of available variables by combinatorial means. This procedure yields several smaller IPDs from a wider single one.

Batches of different IPD typologies

For each of the 20 IPDs of Table B.1 (page 97) we apply a combinatorial procedure to generate many smaller sub IPDs with fewer variables. We design four main IPD typologies by number and type of covariates included. Figure 2.3, page 47 schematizes the characteristics of these data batches.

I. IPD with three variables (only one binary covariate):

   – All combinations of one binary treatment covariate by keeping one time and one time-event outcome fixed. Application to Cox regression and Nelson-Aalen estimation.

II. IPD with two variables (only one continuous covariate):

- All combinations of one continuous variable from the available ones. If application is Gaussian GLM one time covariate is kept fixed while all combinations on one continuous outcome are used. If application is Binomial or Poisson (without offset) GLM a binary outcome is kept fixed, and all combinations of a contiguous covariate are used.

III. IPD with three variables (one/two continuous covariates):

- All combinations of one continuous covariate from the available ones. The remnant two variables are kept fixed. If application is Gaussian GLM one time covariate and a real-valued outcome are kept fixed, while the second covariate varies. If application is Binomial, Poisson (with offset) GLM, Cox regression, or Breslow estimation, then one time and one time-event outcome are kept fixed while one covariate varies.

IV. IPD with three or four variables (two/three binary/continuous covariates).

- All combinations of one binary or continuous covariate from those available, by keeping one time, one time-event, and one additional binary/continuous variable fixed, with applications to all models of the previous point.

Thus IPD type I has only one binary/categorical covariate. IPD type II has only one continuous covariate, while type III has between one and two continuous covariates, depending if time is used as an offset or as a time outcome in a Poisson or Cox model respectively. Similarly IPD type IV has between two and three mixed binary/continuous covariates depending on how time is modeled. IPD with competing events outcomes are split by competing outcome. Further detail on how data is handled is found in Appendix B.6.2, page 96.

Analysis automatization on data

We produce a number of routines to automatize simulation of each IPD set. For a program extract see Appendix D, page 143. For each IPD type and for each simulation option (Section 2.2.3) we simulate B copies of the original IPD following Algorithm 2.1.1 or its variant Algorithm A.2.1. We use the following

Convention 2.4.1. (IPD simulation size) If the IPD sample size is less than 500 we set $B = 300$, and $B = 100$ otherwise.

For each simulated IPD the bias relative to the original IPD (Section 2.2.4) is computed. We derive B inference simulations (see Section 2.3.4, 2.3.5, and 2.3.6) from the respective simulated IPD following Algorithm 2.1.2, for each model type (Section 2.3.1) and for each simulation option of Section 2.3.3. For each simulated inference we compute the bias relative to the original IPD inference (Section 2.3.7).

2.4.3 Global simulation options

We combine the four data simulation options of Section 2.2.3, page 38, with the three statistic simulation options of Section 2.3.3, page 41. Figure 2.4, page 48, schematizes the data and inference simulation options. We obtain the global option combinations of Table 2.8, page 46. We roughly classify these global options by simulation performance:

Sub-optimal options: (1-1), (1-2), (2-1), (2-2).

Mildly optimal options: (3-1), (1-3), (2-3), (3-2),

Near-optimal options: (3-3), (2-4), (3-4), (1-4).

The number left to '-' refers to the IPD statistic simulation option while that right to '-' refers to the IPD simulation option.

Table 2.8: Global reconstruction options for the IPD inference and its underlying IPD. Methods of IPD reconstruction are broadly divided into Permutation-based that uses only incomplete correlation knowledge, and NORTAmax that uses complete correlation knowledge. Method 1-3 and 1-4 employ ordinary NORTAmax resampling (no likelihood numerator adjustment). See Figure 2.4, page 48 for further information. Log-lik.: Log-likelihood.

		Reconstruction option					
	Code	IPD inference	IPD	IPD marginal	Known moments	Correlation	Log-lik. nominator
Permutation	1-1	1	1	Gamma	2	incomplete	random
	2-1	2	1	Gamma	2	incomplete	average
	3-1	3	1	Gamma	2	incomplete	original IPD
	1-2	1	2	Johnson	4	incomplete	random
	2-2	2	2	Johnson	4	incomplete	average
	3-2	3	2	Johnson	4	incomplete	original IPD
NORTAmax	**1-3**	1	3	Gamma	2	complete	random
	2-3	2	3	Gamma	2	complete	average
	3-3	3	3	Gamma	2	complete	original IPD
	1-4	1	4	Johnson	4	complete	random
	2-4	2	4	Johnson	4	complete	average
	3-4	3	4	Johnson	4	complete	original IPD

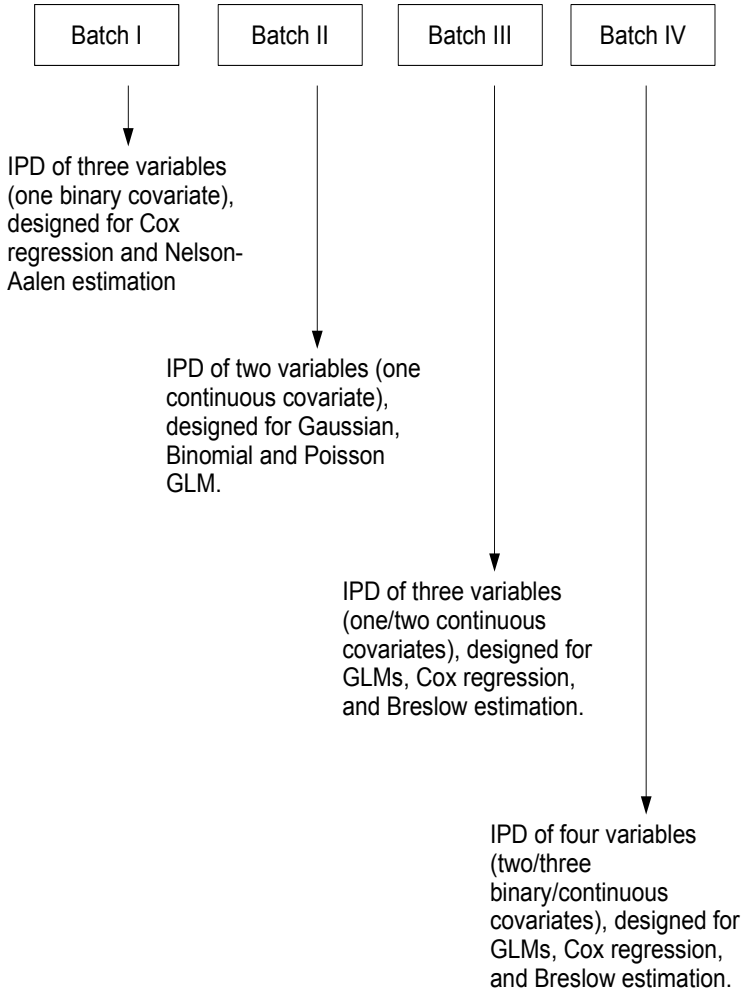

Figure 2.3: Experiments on IPD and IPD inference reconstruction are performed on IPD of different typologies, that is batches. The schema gives a brief description of each IPD batch. Each data batch has two goals. The first is to assess IPD simulation methods for increasing variables number and complexity. The second is to assess simulations of a specific IPD inference from the simulated IPD (see Section 2.3, page 40). The method of IPD and IPD inference simulation is be that of Figure 2.1 (page 28).

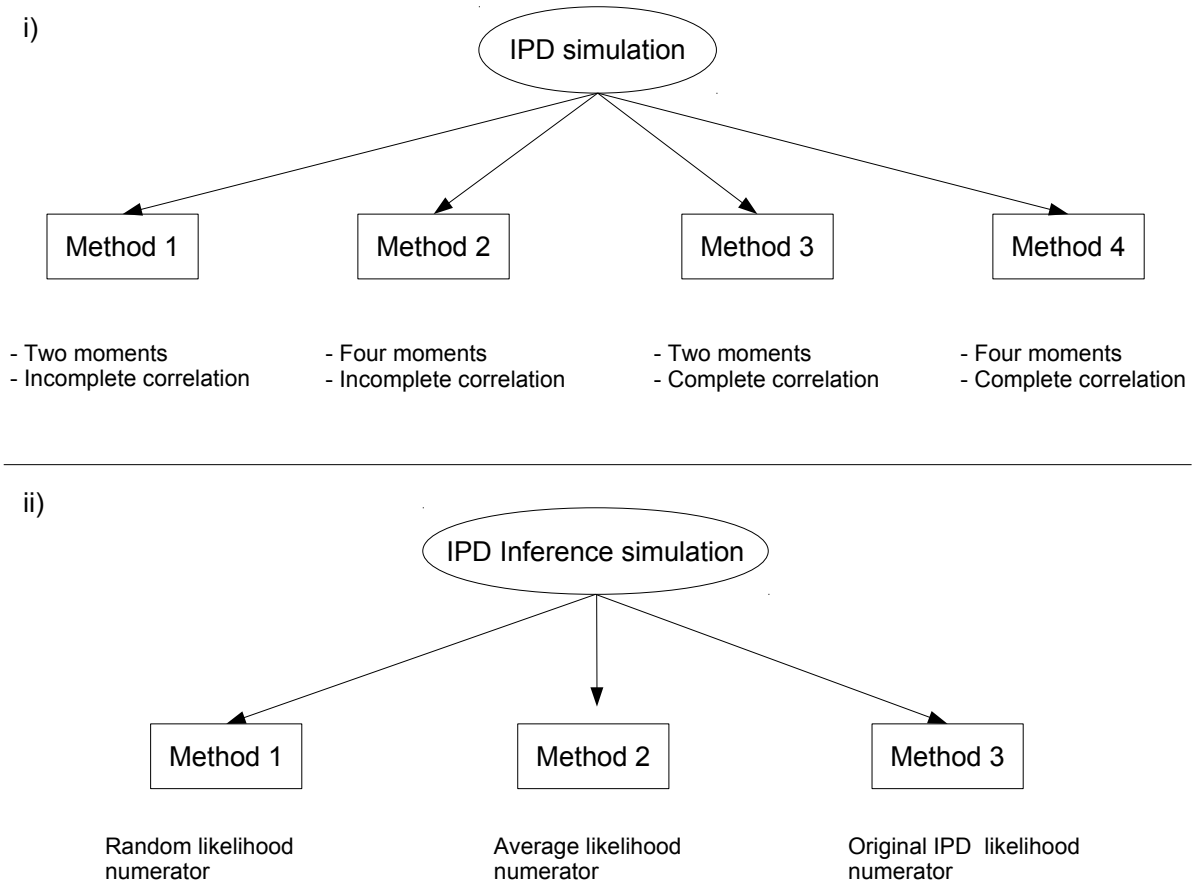

Figure 2.4: Schema for the global simulation options. i) The IPD is simulated under four different scenarios depending on the amount and completeness of empirical summary IPD information. All methods use Gamma and Johnson continuous marginals in the two or four moments case respectively, while a binary marginal is Bernoulli distributed with given mean. Method 1 and 2 use a permutation-based procedure (see Appendix A.2.2, page 86) to re-order the sampled marginals and induce incomplete correlation structure. Method 3 and 4 use NORTAmax resampling (see Algorithm 2.1.1, page 24) and recover complete correlation structure. ii) The simulated IPD is converted into an IPD inference. If the inference is likelihood-based (all models in Section 2.3.1, page 40 except Nelson-Aalen estimation), then the likelihood numerator can be handled according to three different strategies. Method 1 let the numerator vary with the corresponding IPD simulation. Method 2 set the numerator to the fixed average value across all IPD simulations. Method 3 set the numerator to a fixed deterministic value, ideally disclosed from the original IPD. For example, the likelihood numerator from a Cox model with a single binary covariate is the number of events in treatment group, which can either randomly vary or be fixed to an expectation estimate or to its true IPD value. The global simulation is performed according to all possible combinations between the data and inference simulation option (see Table 2.8, page 46).

Chapter 3

Experimental results

3.1 Similarity between reconstructed and original IPD

We give results on experimental evaluations on point A of Section 2.1.8. That is we empirically assess IPD reconstruction according to methodological implementations of Section 2.2. Of the original IPD sources (Table B.1, page 97) about 10% includes hundred records, 31% between hundred and four-hundred, 35% between four-hundred and thousand, and 24% over thousand records. We re-organize the original IPD sources into batches of different typology accordingly to Section 2.4.2. For each data batch the original IPD source is reproduced from its summaries only according to methods of Section 2.2.3 and 2.4.2. By Convention 2.4.1, page 45, and with an approximate median of 400 records per IPD we have roughly balanced proportions of IPD simulations with size $B = 300$ or $B = 100$. For each data batch and simulation option we count the number of reconstructed IPDs satisfying all or some similarity conditions (point 1 to 5, Section 2.2.4) relative to the total number of reconstructions. A comprehensive summary of this section and related Appendix material is given on page 52.

Table 3.1 reports the overall percentage of reconstructed IPDs similar to the original IPD reference. The percentage of similar marginal moments or correlations is also reported.

Generally we see the percentage of reconstructed IPDs that is overall similar to its original reference increases from method 1 to 4, that is with increasing knowledge and completeness of IPD moments and correlations information. Typically, the best IPD reconstruction is given when four marginal moments and a complete correlation matrix are both known. Recovery rates of first and second moments are always close to 100% regardless of data batch and simulation method. Recovery rates for higher moments varies with simulation method and data batch.

As expected Johnson-drawn continuous marginals (method 2 and 4) improve third and fourth moments recovery. This performance decreases with increasing batch order, and batch III and IV display (overall) lower recovery rates. Correlation recovery rates mostly varies with simulation method. Generally NORTAmax resampling (method 3 and 4) never displays correlation recovery rates lower than 79%. As expected incomplete correlations knowledge (method 1 and 2) yields

Table 3.1: Percentage of reconstructed IPDs satisfying overall similarity conditions of Section 2.2.4 (page 38) relative to original IPD reference. Marginal moments and correlations similarity is also assessed. Batch size is the number of distinct data-sets in each batch (Section 2.4.2, page 44). IPD simulation in each batch occurs according to method 1 to 4 of Section 2.2.3 (page 38).

			Similar (%)			
				Moments		
Data batch	Batch size	Method	*Overall*	*1st and 2nd*	*3rd and 4th*	*Correlation*
I	111	1	0.90	100.00	12.61	25.23
		2	26.13	100.00	92.79	27.93
		3	12.61	99.10	12.61	89.19
		4	88.29	100.00	91.89	94.59
II	241	1	16.60	99.17	16.60	100.00
		2	76.76	100.00	76.76	100.00
		3	15.77	98.76	16.60	95.02
		4	73.44	98.76	78.84	93.36
III	198	1	0.51	99.49	2.02	30.30
		2	17.17	99.49	63.64	28.79
		3	2.02	99.49	2.02	92.93
		4	56.57	97.98	65.66	86.87
IV	195	1	0.00	99.49	7.18	4.10
		2	1.54	99.49	48.72	4.10
		3	6.15	99.49	7.18	81.54
		4	41.54	98.97	51.79	79.49

lower recovery rates especially as the number of data variables increases (batch I, III, and IV). Below we give more details by data batch and we report results by IPD sample size in Appendix C.1.1, page 99.

Batch I includes 111 IPDs composed of three variables, one time and one time-event outcome, plus one binary/categorical covariate. Here we generally see that recovery performances are always very good when either higher moments (method 2 and 4) or a complete correlation matrix (method 3 and 4) or both are known.

Batch II includes 241 IPDs composed of two variables, one binary or continuous outcome, and one continuous covariate. In particular 106 data-sets have a continuous outcome and 135 a binary one. This latter sub-batch lends itself for an assessment of Johnson distribution third and fourth moments recovery ability. Table 3.2 shows the percentage of well recovered third and fourth

moments on average and for increasing IPD sample size. We see an overall good though not always perfect reproduction performance. The Johnson distribution incurs into a 15.56% reproduction error on average, which makes overall higher moments recovery in batch II a bit worst than batch I. Correlation recovery (here scalar) seems best under Algorithm A.2.1 (method 1 and 2).

Table 3.2: Johnson distribution recovery performance of higher moment features under simulation method 2 (Section 2.2.3, page 38), for increasing sample size. Johnson distribution performance: percentage of well recovered third and fourth moments (point 3 and 4 of Section 2.2.4, page 38) under simulation method 2 (Section 2.2.3, page 38) and for increasing sample size. The assessment is made on a sub-set of Batch II including IPD with only one continuous variable. Bottom row: unstratified overall averages.

		Similar (%)	
Subset batch (size)	Sample size	*3th Moment*	*4th Moment*
II.(135)	< 100	100.00	80.00
	(100,400]	91.11	88.89
	(400,1000]	95.00	80.00
	> 1000	100.00	86.67
		95.56	84.44

Batch III includes all 198 composed of three variables, one time and one binary (or continuous) outcome, and one continuous covariate. In particular 113 data-sets have a binary outcome and 85 a continuous one. As compared to batch I and II, here the overall recovery rate diminishes due to a lower third and fourth moment recovery performance. This is due to yet one more Johnson drawn marginal in the data.

Batch VI includes 195 IPDs composed of three to four variables. That is one time and one binary (or continuous) outcome, and between two to three continuous/binary covariates. Here overall performance is yet lower due to inclusion of more continuous Jonson-drawn marginals. Since NORTA-based correlation reproduction (method 3 and 4) depends on the continuous marginal (see Appendix A.1.2, page 80), a lower 3rd/4th moment recovery rate seems to negatively affect correlation recovery, ans seen under method 3 and 4.

Table C.2 and C.3 (pages 101 and 102) show further results on data batch III and IV. In batch IV a lower reproduction performance under method 4 is explained by worst fourth moment recovery. However the reproduction error is relatively tolerable in at least half of the reconstructed IPDs, indicating the overall IPD information recovery could still be relatively good here.

SUMMARY. We give results on IPD reconstruction from IPD summaries only, under four different simulation scenarios and IPD typologies (batches). We can always reconstruct IPD and often with high similarity to the original IPD, on average, based on our similarity criteria. Generally overall IPD reconstruction is best when higher moment features and a complete correlation matrix are both known (NORTAmax). Overall reconstruction accuracy diminishes when more continuous marginals are included (batch III and IV), because the Johnson distribution occasionally fails to well reproduce fourth moment features. This may also badly affect NORTA correlation recovery in batch IV. However in batch IV the magnitude of fourth moment simulation imprecision is often tolerable, indicating that the quality of the recovered IPD information may be practically better than what portraied by our relatively strict similarity criteria.

3.2 Similarity between reconstructed and original IPD inference

We give results on experimental evaluations on point B of Section 2.1.8. That is we perform all inferences of Section 2.3.1 on simulated and original IPDs by batch typology (Section 2.4.2) and simulation options (Section 2.4.3). The MC sample size reflects Convention 2.4.1. On page 58 we give a summary of this Section and of Appendix C.2 (page 103) where we give more detailed tabular results. Figure 3.1, page 53 is a graphical guide on how to read Figure 3.2 to 3.4. Results for each data batch are stratified by similarity of reconstructed IPD to the original reference, and by simulation method. In the boxplots below we mark purely NORTAmax approaches – no likelihood numerator adjustment (method 1-3 and 1-4) – with an asterisk.

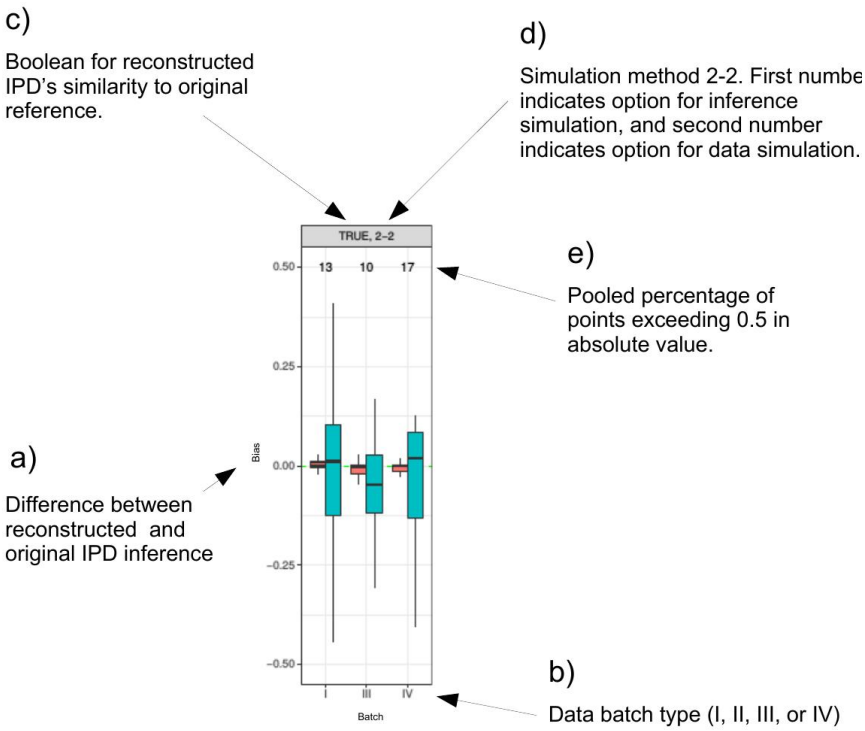

Figure 3.1: Guide to read Figure 3.2 to Figure 3.4 (pages 55 – 3.4). The figure shows a) bias distribution, the difference between a reconstructed and original IPD inference for each b) data batch (see Figure 2.3, page 47). Each panel is stratified by c) similarity (TRUE or FALSE) of the reconstructed IPD to its original reference and d) the global simulation method. This consist of a method to simulate the inference and a method to simulate the IPD on which the inference is computed (see Figure 2.4, page 48 and Section 2.2.4, page 38). For each batch on top of each boxplot is reported e) the percentage of points exceeding ±0.5. Outliers – points exceeding 1.5 times the inter-quartile-range on both directions – are not plotted for graphical clarity.

Figure 3.2, page 55, shows boxplots for the bias distribution of the average MLE and r.f.i.d. simulation, relative to original IPD value. Generally the bias distribution is well centered on zero regardless of method or data-similarity. We typically see that from panel 1-3 to 3-4 – NORTAmax methods – the MLE bias is tightly centered on zero. The r.f.i.d. bias is even more tightly centered on zero. Complete correlation information seems important to reduce bias (methods 1-3 to 3-4). Knowledge of four (methods 1-4 to 3-4, using Johnson continuous marginals) or two (methods 1-3 to 3-3, using Gamma continuous marginals) moments does not seem to yield very different bias distributions. We generally see that likelihood numerator adjustment might reduce bias only under incomplete correlation knowledge (panels FALSE-3-1 and FALSE-3-2). Outliers are not plotted for graphical clarity and we generally see the percentage of bias values exceeding ±0.5 is lower when reconstructed IPD is similar to its original reference. In particular methods 1-3 and 1-4 – ordinary NORTAmax resampling – typically yield lower outliers percentages, hence thinner bias distribution tails.

Figure 3.3, page 56, shows boxplots for the bias distribution of the 2.5th and 97.5th empirical quantile of the MLE simulation relative to any of the reference IPD basic, percent, Bca, or normal bootstrap quantiles. In general the bias distribution is wide and not always centered on zero, except under NORTAmax sampling (method 1-3 and 1-4) where bias is more tightly centered on zero. NORTAmax sampling with four (panel TRUE,1-4) or two (panel FALSE,1-3) moments seems to yield the least bias, if reconstructed IPD is similar and not similar to its reference respectively. In these two circumstances the bias distribution has also relatively thinner tails (see percentages of values exceeding ±0.5).

Figure 3.4, page 57, shows boxplots of the bias distribution of an aggregate index of the expected c.h.e., relative to the original IPD estimate. For each c.h.e. marginal axis, this index includes the first three quartiles, the mean, and the range values. Generally we see the bias distribution of the marginal cumulative events-count (y-axis) is very close to zero regardless of method and IPD similarity. Nevertheless ordinary NORTAmax resampling with four (panel TRUE,1-4) or two (panel FALSE,1-3) moments seems to yield the least bias in the case IPD is similar or dissimilar to reference respectively. The bias distribution of the marginal events-time line (x-axis) is drastically wider but more contained under the two conditions just mentioned. As a result of the nearly zero y-axis reconstruction bias, the expected c.h.e better approximates the original IPD graph the smaller the events-time reproduction error gets (also see Section 3.2). By the optimization scope, the events-time line reproduction error is bounded by a nearly one-to-one linear correlation between simulated and original variables (result not shown).

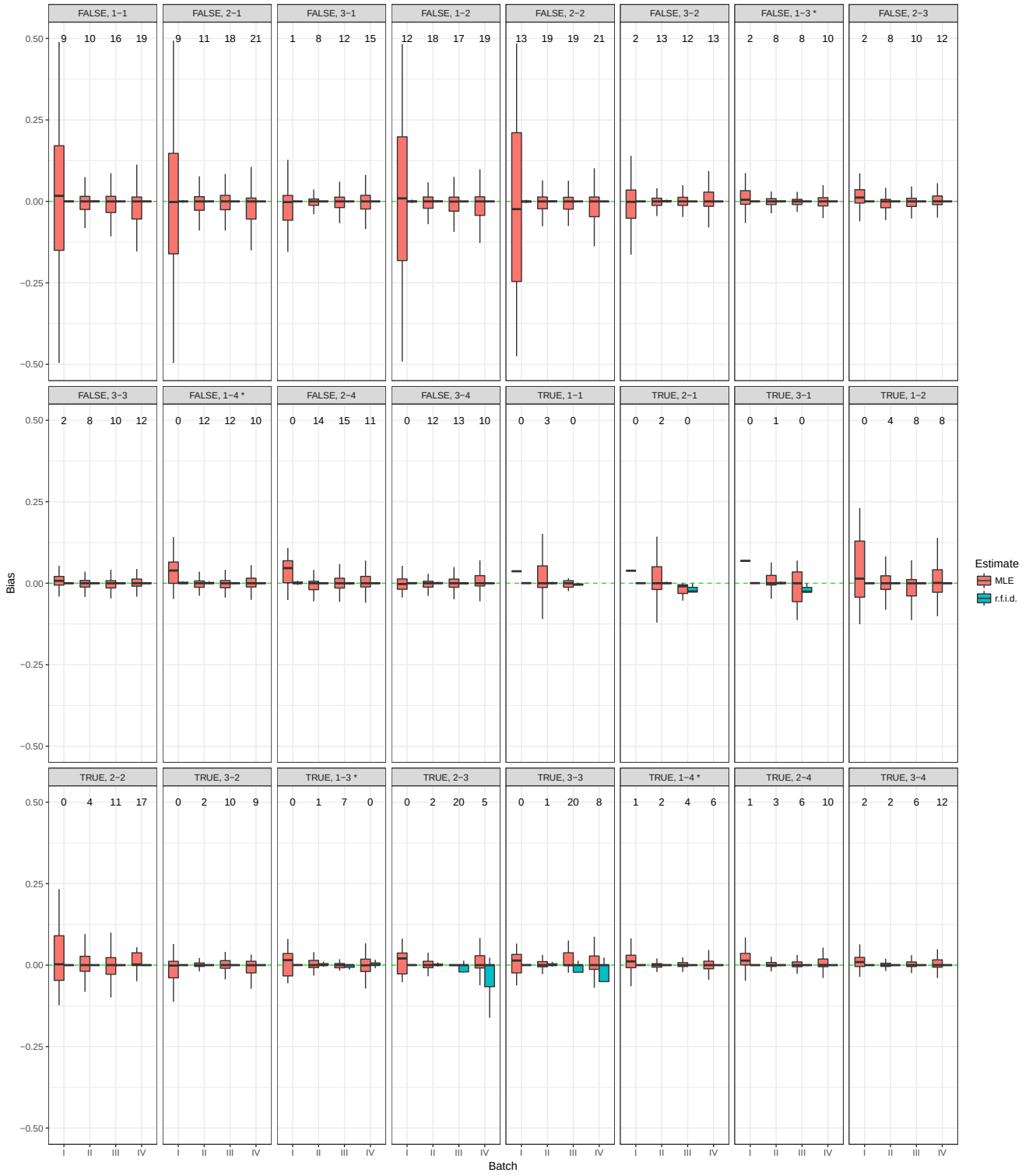

Figure 3.2: Bias distribution of average point estimate simulation versus IPD reference value, for different types of data batches (see Section 2.4.2, page 44). Outlying points are excluded for graphical clarity. The number on top of each boxplot pair is the overall percentage of bias instances exceeding ±0.5 across estimate types. Results are stratified by the Boolean for data similarity (see Table 3.1, page 50) and by simulation method (see Section 2.4.3, page 46). Find further information in Table C.4 (page 107), Table C.7 (page 110), Table C.9 (page 112), Table C.12 (page 115).

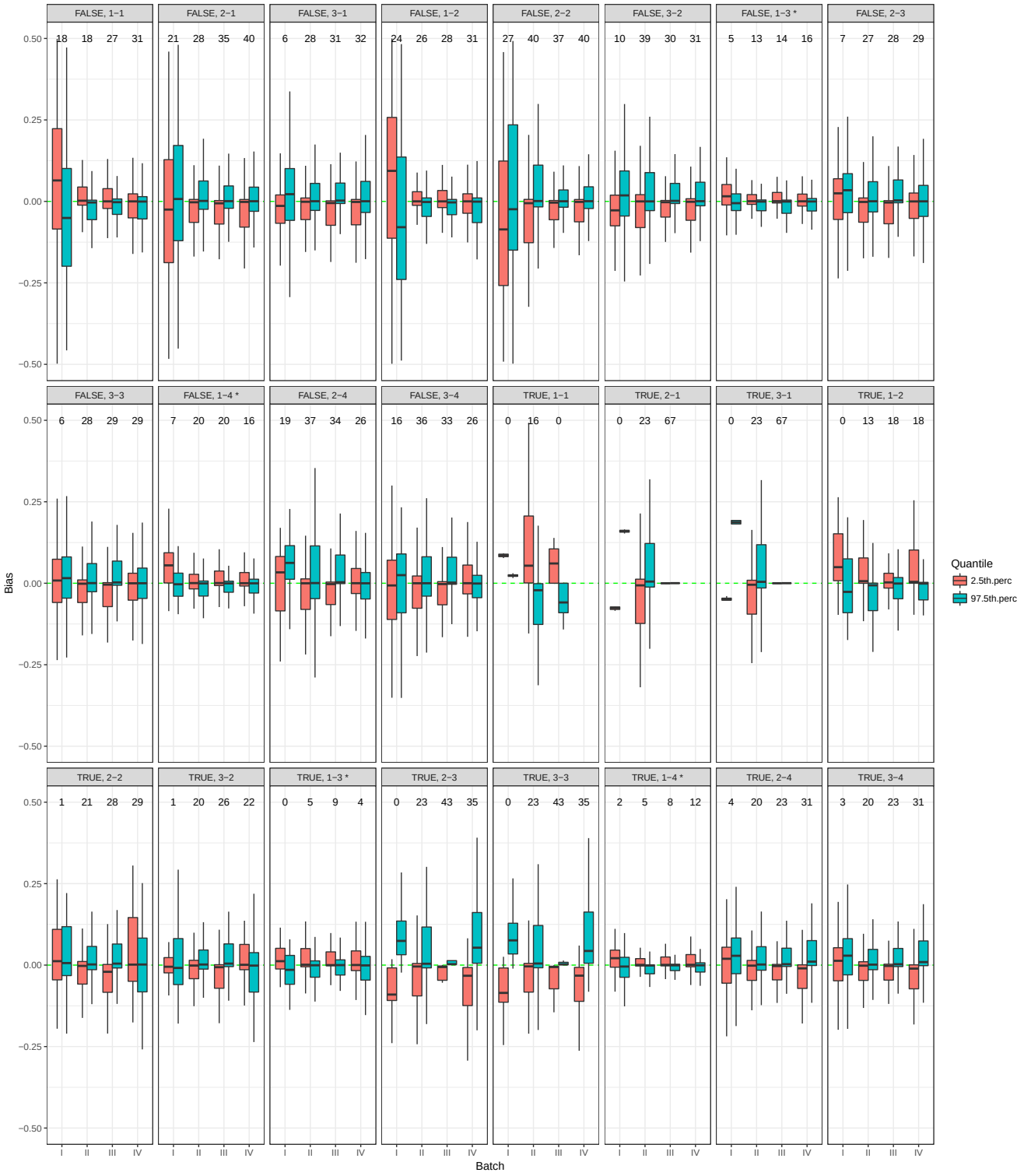

Figure 3.3: Bias distribution of 95% empirical CIs of the MLE simulation versus IPD (basic, normal, Bca, or percent) bootstrap reference quantiles, for different types of data batches (see Section 2.4.2, page 44). Outlying points are excluded for graphical clarity. The number on top of each boxplot pair is the overall percentage of bias instances exceeding ±0.5 across quantiles. Results are stratified by the boolean for data similarity (see Table 3.1, page 50) and by simulation method (see Section 2.4.3, page 46). Find further information in Table C.5 (page 108), Table C.8 (page 111), Table C.10 (page 113), Table C.13 (page 116).

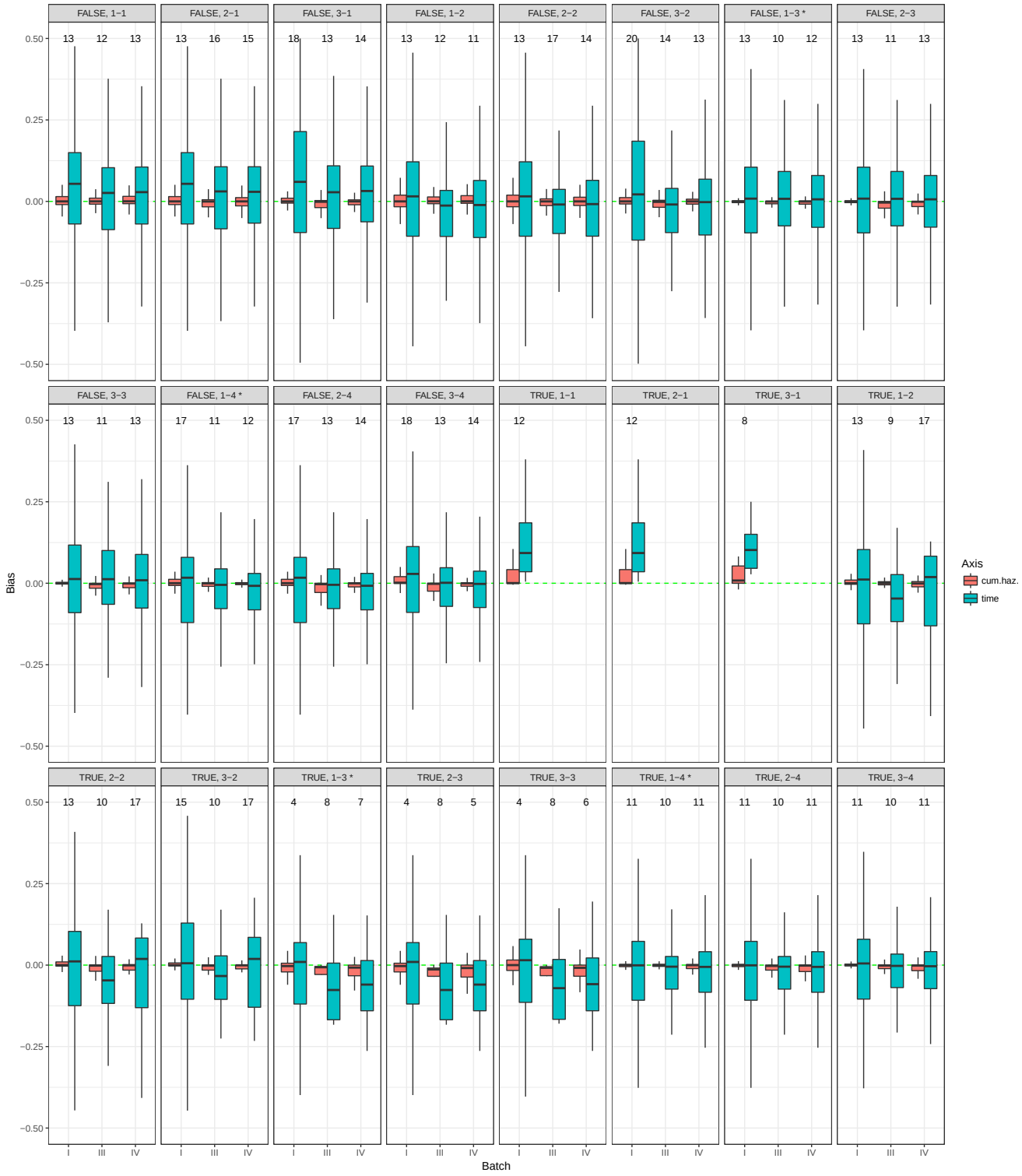

Figure 3.4: Bias distribution of an aggregate index of each marginal axis (cumulative events-count and log event-times) of the average cumulative hazard simulation versus IPD reference (Nelson-Aalen in batch I, or Breslow otherwise), in three data batches (see Section 2.4.2, page 44). The aggregate index is the axis first three quartiles, mean, and range. Outlying points are excluded for graphical clarity. On top of each boxplot pair: overall percentage of bias instances exceeding ±0.5 across axes. Results are stratified by the Boolean for data similarity (see Table 3.1, page 50) and by simulation method (see Section 2.4.3, page 46). Find further information in Table C.6 (page 109), Table C.11 (page 114), Table C.14 (page 117).

SUMMARY. We give results on IPD inference reconstruction from reconstructed IPD. We employ twelve simulation options on four different IPD typologies (batches). In particular we introduce the option to fix a constant (original) IPD log-likelihood numerator versus letting it vary with each simulation, that is the default. We can reconstruct MLEs from IPD GLM and Cox regressions under different GLM families and (baseline) hazards assumptions, as well as Nelson-Aalen or Breslow estimates. We also recover 95% empirical quantiles of the MLE simulation alongside original IPD bootstrap intervals. We show that under certain circumstances the IPD inference reconstruction is highly comparable to the original IPD inference. Generally MLE reconstruction is better when the original IPD complete correlation is known (NORTAmax). Usage of a constant likelihood numerator could be beneficial if no complete IPD correlation is available (see batch I). Recovery of bootstrap-like 95% intervals is generally best under ordinary NORTAmax resampling (method 1-3 and 1-4). Here any type of likelihood-numerator adjustment is typically detrimental. Reconstruction of marginal cumulative events-counts of the cumulative hazard graph is generally remarkably good irrespective of the amount of original IPD correlation or moments knowledge, although NORTAmax resampling typically yields slightly better results. On the other side reconstruction of the marginal events-time line is generally less accurate, but NORTAmax resampling typically yields the least reconstruction error. To sum up a good IPD inference reconstruction is generally possible even if the underlying IPD reconstruction is only approximate (w.r.t. higher degree marginal moments features), but so far as the original IPD correlation structure is well recovered.

3.3 Practical Examples: data and statistic reconstruction

We further elaborate on the example of Section 2.1.9, page 29. IPD `diab.2` is an example from data batch IV (see Section 2.4.2, page 44), and is a variables' sub-selection of data-set `diabetes` (see Appendix B.6.1, page 96). Table 2.1, page 29, shows an excerpt of `diab.2` along with its marginal moments (Table 2.2, page 30) and correlations (Table 2.3, page 30). We generate $B = 300$ realizations of `diab.2` under differently optimal resampling strategies. Then we check fidelity of the data simulations by graphical and descriptive means. Ultimately we compute inferences of interest on each data realization and obtain respective samples for that inference. Here we reconstruct a Cox model MLE and its corresponding Breslow or Nelson-Aalen estimate.

3.3.1 Example: sampling under sub-optimal settings

Suppose we can only recover an incomplete IPD correlation matrix. In our case we assume only first order empirical correlations are available. Also we can only recover mean and variance for each IPD real-valued marginal variable, but not higher moments information. The mean of each binary variable is always given. Since the correlation is incomplete NORTAmax resampling cannot be used. Instead we use Algorithm A.2.1, page 86, to recover the first degree correlation structure as following.

Table 3.3: Proportional Hazards Cox regression for the MLE of the log HR of 'Treatment' (β_1) and 'Age' (β_2) as estimated on the original `diab.2` IPD or as an average (MC mean) across 300 simulated MLEs. The method of MLE recover is based on an incomplete IPD correlation reconstruction (Algorithm A.2.1, page 86). HR: Hazard Ratio; Info. proxy: inverse of Fisher Information diagonal. Lower/Upper CI: $\hat{\beta} \pm 1.964 \sqrt{\text{Var}(\hat{\beta})}$; Log Lik.: $\ell(\hat{\beta})$. MC sd: standard deviation across all simulated MLEs

	Point Estimate									
	log HR		Info. proxy		Lower CI		Upper CI			
Estimate	$\hat{\beta}_1$	$\hat{\beta}_2$	$\text{Var}(\hat{\beta}_1)$	$\text{Var}(\hat{\beta}_2)$					Log Lik.	*AIC*
IPD	-0.782	0.004	0.029	0.00003	-1.113	-0.007	-0.451	0.015	-856.5	1717.0
MC mean	-0.403	-0.001	0.027	0.00003	-0.723	-0.012	-0.082	0.010	-855.5	1714.9
MC sd	0.119	0.004	0.002	0.00000	0.120	0.004	0.119	0.004	59.5	118.9

We first sample the real-valued marginals, `time` and `agedx`, from a Gamma distribution with corresponding mean and variance. Alongside we sample the binary marginals, `status` and `treat`, from a Bernoulli distribution with corresponding mean. Next we merge the simulated marginals by reconstructing their incomplete inter-dependence structure with a permutation-based heuristic. An

excerpt of this code is in Appendix D.2, page 144. Output from the comparison between the reconstructed and the original IPD is given in Appendix C.3.1, page 132. We see good agreement up to the second moment for all marginals. However, third an fourth moments of the Gamma-sampled continuous marginals (`time` and `agedx`) are not well recovered. By design the correlation matrix is not fully recovered. The `bool` message confirms that under our criteria the simulated data does not well reflect the observed distributional properties. In Figure 3.5 (page 60) we check agreement of the two continuous variables to the limiting Gamma density. Binary variables converge to a Bernoulli distribution by default, and we do not check that. The Gamma marginal seems to not

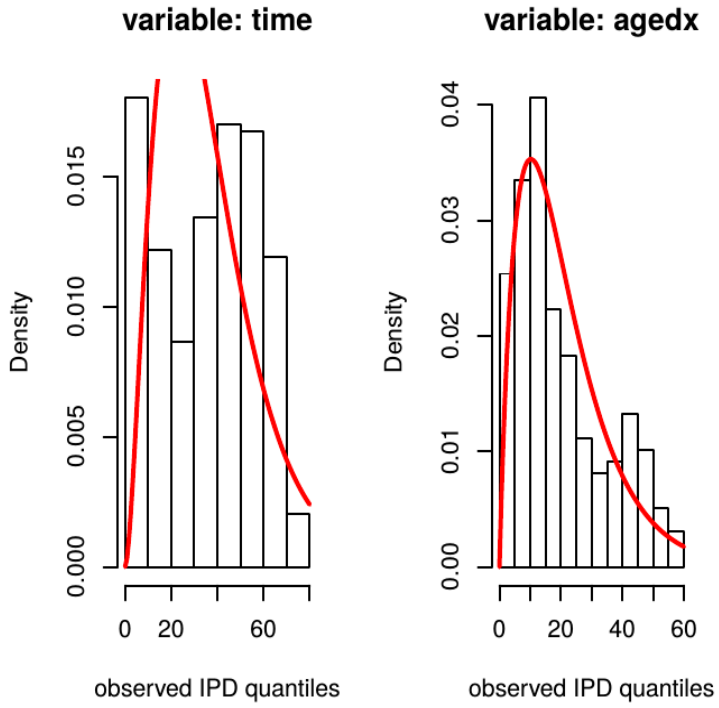

Figure 3.5: Agreement of IPD marginal histograms to its respective limiting Gamma density (red line), $k = 2$.

so adequately describe the limiting distribution for variable `time`. Differently given only the first two moment constraints, variable `agedx` is fairly well described by its limiting Gamma density at $n = 394$.

Next we compute the log Hazard Ratio (HR) vector and Breslow estimation on each data realization. Excerpt for this program is in Appendix D.2. A comparison between the reconstructed log HR and it original IPD value is given in Table 3.3, page 59 We see enough difference, except for the reciprocal Fisher Information diagonals that is well recovered. This indicates some key IPD

likelihood information is conserved that is reasonable with MaxEnt marginals, although under limited prior information. In Figure 3.6 (page 64) we display the generated log HR and r.f.i.d. samples with the original IPD and recovered point value. We also show an average Breslow estimate (see Section 2.3.6, page 42) over all generated simulations and stratified by treatment. This Breslow estimate recovery displays some deviation from its IPD counterpart especially for the treatment group.

In Table 3.4, page 61, we show a comparison between generated 95% quantiles for the log HR vector. MC indicates our generated sample versus ordinary bootstrap intervals computed on original IPD.

Table 3.4: Empirical quantiles for the log HR of Table 3.3. MC denotes the HRs generated from the 300 simulated IPDs. The method of HR simulation is based on an incomplete IPD correlation reconstruction (Algorithm A.2.1, page 86). Results compare the empirical quantiles of the MaxEnt bootstrap of size 300 against several types of non-parametric bootstrap CIs, as computed each on a 10000 sample of the original IPD. B: bootstrap sample size.

			Empirical Quantile			
			2.5th		97.5th	
Estimate	B	Bootstrap	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
MC	300		-0.627	-0.010	-0.154	0.007
IPD	10000	normal	-1.111	-0.007	-0.445	0.015
		basic	-1.098	-0.007	-0.438	0.015
		percent	-1.126	-0.007	-0.466	0.015
		Bca	-1.120	-0.007	-0.463	0.015

The result remains unsatisfactory at least for `beta.1` quantiles recovery.

3.3.2 Example: sampling under near-optimal settings

Here we suppose a complete empirical correlation matrix from IPD `diab.2` is available, that allows to perform NORTAmax resampling. We also suppose to have moments up to degree four for all continuous marginals, `time` and `agedx`. The latter are drawn from the respective Johnson distribution. We generate $B = 300$ bootstrap repetitions of data `diab.2` and a code excerpt is in Appendix D.2, page 144 Table 2.4 and 2.5, page 32, show all main IPD distributional summaries are well recovered on average. In Figure 2.2 (page 31) we see the Johnson marginal well approximates the original IPD marginal empirical frequencies for `time` and `agedx` at $n = 394$. Figure 3.7, page 65, shows approximately Normal shaped MaxEnt bootstrap distributions for the log HR of 'Treatment' and 'Age' – see Table 2.6, page 33, for more details. The expected Breslow curve is here remarkably close to its IPD counterpart. Table 2.7, page 34, shows 95% MaxEntBoot

quantiles for the log HR vector versus ordinary bootstrap intervals computed on IPD. We see here good agreement especially with type percent or Bca.

3.4 Practical Examples: long-run prediction

We show the MaxEnt bootstrap distribution can be a predictive alternative to a defect original IPD inference (see Remark 2.1.8, page 24). In Appendix C.3.3, page 136, we give a similar example and in Appendix C.3.4, page 137, we show careful constraints imposition can be important for accurate reconstruction. In Appendix C.3.5, page 138, we show MaxEnt bootstrap 95% quantiles can be more stable alternative t to the ordinary IPD counterparts.

3.4.1 Example: predictive use of MaxEntBoot sample

We consider IPD `wh.4` (batch I) that is a variables' sub-selection from the `whiteall1` dataset (Royston and Sauerbrei, 2008). Survival time, time-event status, and a binary treatment indicator are recorded for each patient. An IPD excerpt along with its empirical moment correlation matrix is given in Appendix C.3.2, page 133. There we see individuals in group `all10 = 1` strongly correlate with shorter survival times while modestly correlating with event occurrence.

The constant hazard in group `all10 = 0` as given by the classic MLE estimate on page 405 of Andersen et al. (1993), is 0.005, while that in group `all10 = 1` is 0.073, and their ratio is 14.280, showing an extremely greater hazard in treatment group relative to control. Exploding estimates may be a warning on a possible sparse data bias (Greenland et al., 2016) in the original IPD. Indeed we see here disproportionally less individual at risk in group `all10 = 0` than in group `all10 = 1`.

Table 3.5: Proportional Hazards Cox regression for the MLE of the log HR of 'Treatment' (β_1) as estimated on the true `wh.4` IPD or as an average (MC mean) across 300 simulated MLEs. The method of MLE simulation from the simulated IPDs is that of Algorithm 2.1.2, page 25, based on NORTAmax resampling. HR: Hazard Ratio. Lower/Upper CI: $\hat{\beta} \pm 1.964 \sqrt{\text{Var}(\hat{\beta})}$; Log Lik.: $\ell(\hat{\beta})$. MC sd: standard deviation across all simulated MLEs

	Point Estimate					
Estimate	$\hat{\beta}_1$	$\text{Var}(\hat{\beta}_1)$	Lower CI	Upper CI	Log Lik.	*AIC*
IPD	23.077	185007.70713	-819.968	866.122	-22122.1	44246.2
MC mean	4.328	0.00574	4.179	4.476	-22592.7	45187.3
MC sd	0.066	0.00025	0.064	0.069	393.9	787.8

An inspection of the Nelson-Aalen estimates in each group reveals a strong departure from the hazard ratio proportionality assumption. The original IPD Cox regression here yields a singular

MLE (see Appendix C.3.2). The IPD HR estimate and relative r.f.i.d. are nearly infinite and these original IPD inferences are little informative.

Next we generate $B = 100$ NORTAmax repetitions of IPD `wh.4` from the given empirical IPD summary constraints. The comparison between reconstructed and original IPD is in Appendix C.3.2 showing an overall good agreement. Table 3.5, page 62, compares the average MaxEnt bootstrap log HR or its r.f.i.d.m against their respective original IPD values. We see the original IPD estimates tend to explode, but the MaxEntBoot expectations retain enough numerical stability and interpretable. Here the expected HR estimate is exp(4.3) ≈ 73.7, 95% CI: (66-7 – 90.0).

In Figure 3.8 we plot the expected Nelson-Aalen estimate against its IPD counterpart. In both treatment groups the curves well agree. Table 3.6, page 63, compares MaxEntBoot and ordinary IPD bootstrap 95% confidence intervals. MaxEntBoot intervals well agree with the normal approx-

Table 3.6: Empirical quantiles for the log HR of Table 3.5. MC denotes the HRs generated from the 300 simulated IPDs. The method of HR simulation from the simulated IPDs is that of Algorithm 2.1.2, page 25, based on NORTAmax resampling. Results compare the empirical quantiles of the MaxEnt bootstrap of size 300 against several types of non-parametric bootstrap CIs, as computed each on a 10000 sample of the original IPD. B: bootstrap sample size.

			Empirical Quantile	
Estimate	B	Bootstrap	2.5th	97.5th
MC	300	MaxEnt	4.200	4.449
IPD	10000	normal	22.811	24.188
		basic	22.884	24.200
		percent	21.954	23.270
		Bca		

imation given in Table 3.5. Here ordinary IPD bootstrap intervals suffers the same instability of the original IPD point estimate. All Bca computations also fail in this IPD example.

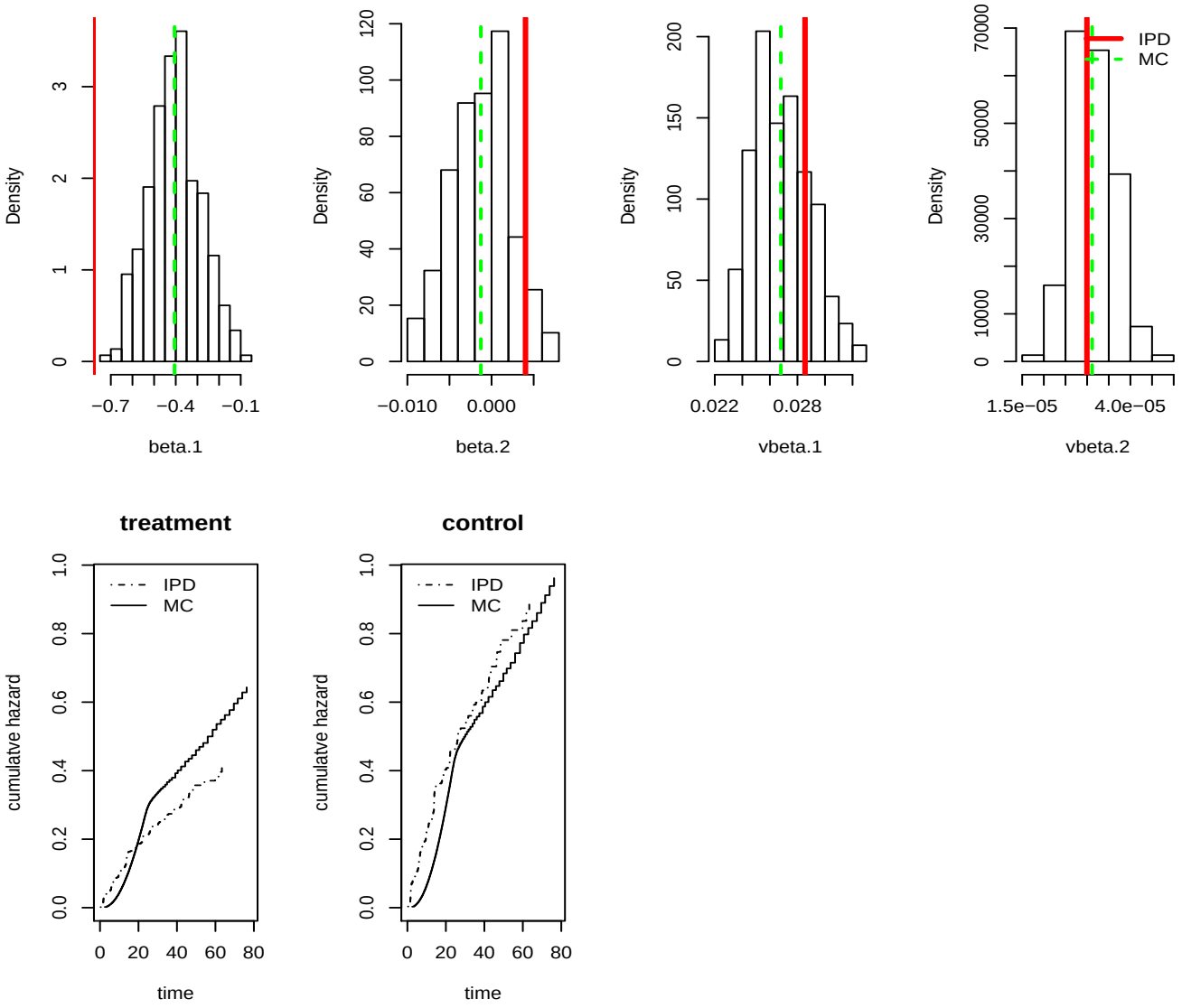

Figure 3.6: Generated samples for the log HR (`beta.`), and respective reciprocal Fisher Information diagonal (`vbeta.`), using non NORTAmax algorithms, alongside Breslow estimates in group treatment and control. MC = Monte Carlo average. IPD = reference estimate

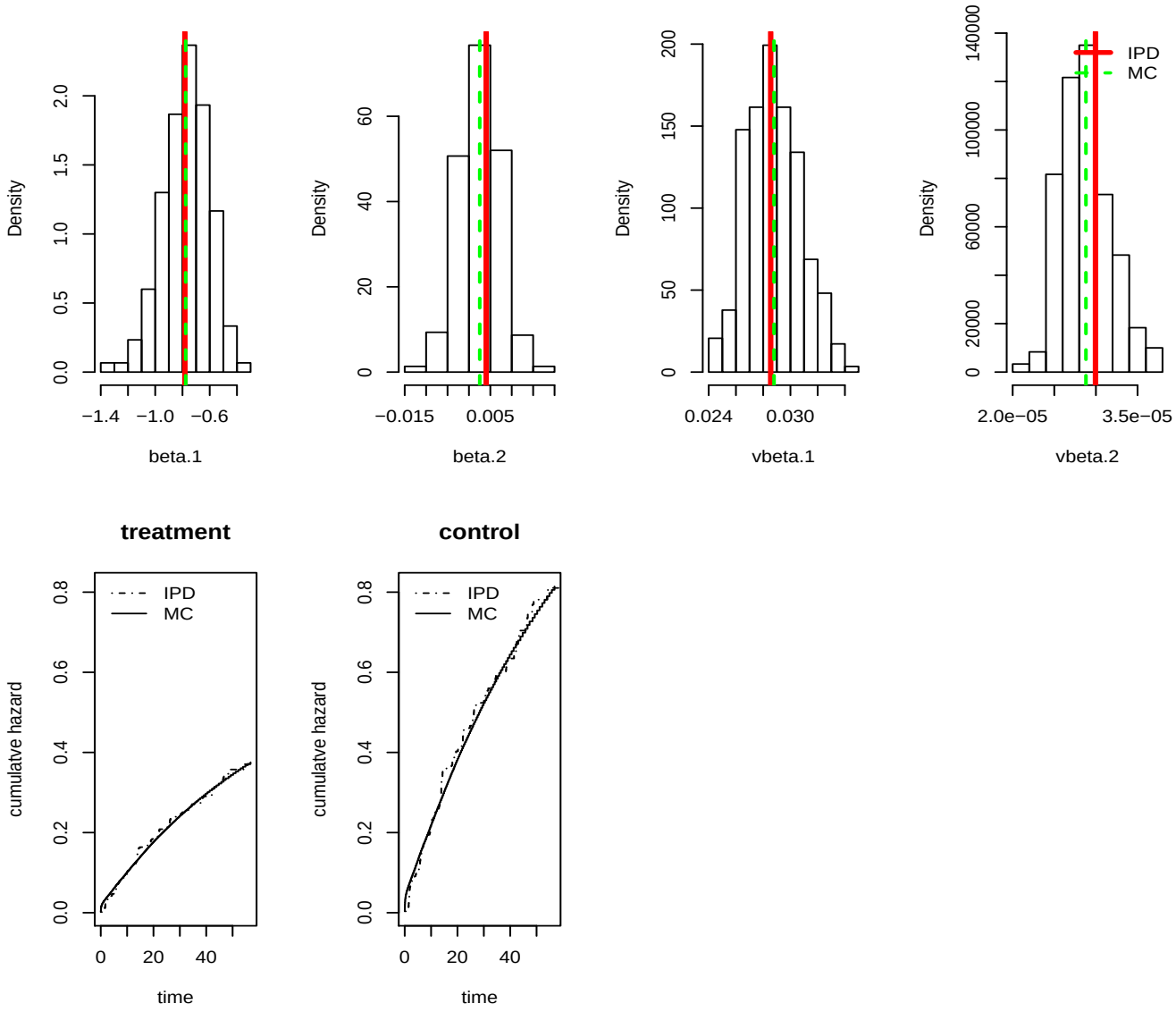

Figure 3.7: MaxEntBoot samples for the log HR (`beta.`), and respective reciprocal Fisher Information diagonal (`vbeta.`) alongside Breslow estimates in group treatment and control. MC = Monte Carlo average. IPD = reference estimate.

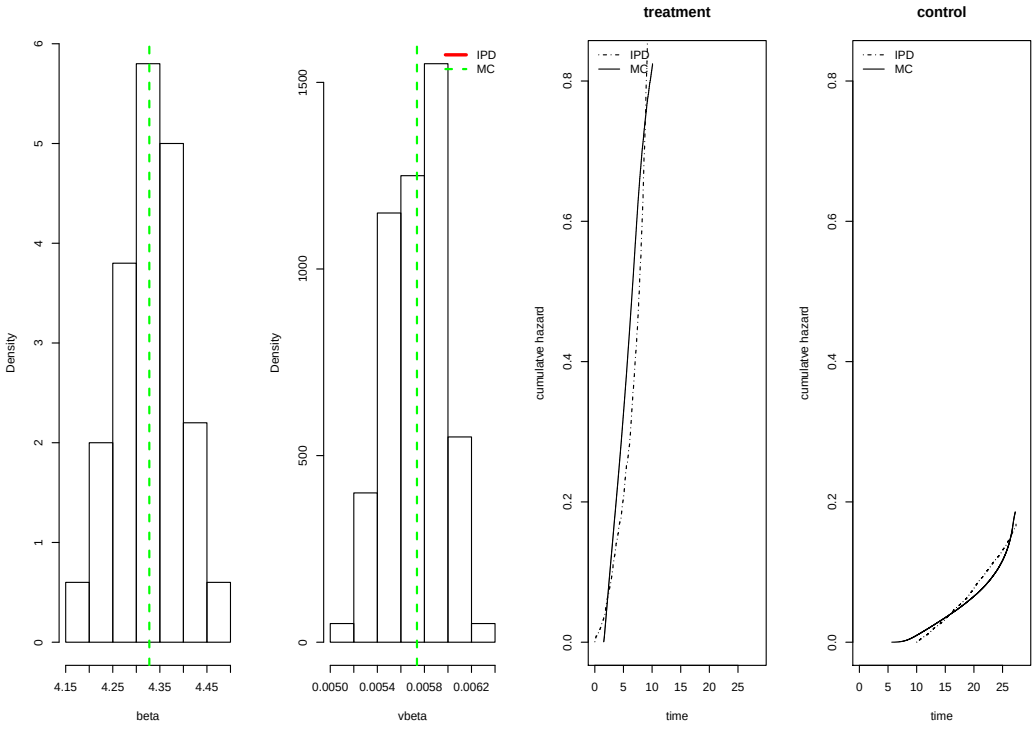

Figure 3.8: The `wh.4` data: MaxEntBoot samples for the log HR (`beta.`) and respective reciprocal Fisher Information diagonal (`vbeta.`) alongside Nelson-Aalen estimates in group treatment and control. MC = Monte Carlo average. IPD = reference estimate. Note the IPD point estimate is outside the plotting range because singular.

Chapter 4

Discussion

4.1 General

Today we see an increase in the digitization and storage of personal data, that is IPD. However due to privacy, legislative, or technical issues original IPD cannot always be publicly disclosed nor shared and must remain anonymous. Personal health data is an example. Researchers, companies, and institutions still typically need to perform IPD inferences even if original IPD is not available that is a typical challenge in fields like statistical disclosure control, meta-analysis, distributed computing, research synthesis and reproduction.

Here we offer a principled method to reconstruct original IPD and IPD inferences from knowledge of empirical IPD marginal moments and correlation matrix only. We argue such summary format entails limited IPD information loss while maintaining privacy. To unlock IPD information we apply a decompression procedure that reconstructs a stochastic, anonymous, representation of the IPD raw format. Such IPD reconstruction is performed by what we call NORTAmax resampling (Corollary 2.1.1.1, page 18) that draws from a multi-variate maximum entropy distribution based on a Gaussian copula given the above empirical IPD summaries. We argue NORTAmax resampling is asymptotically ($n \to \infty$) equal to draw from the IPD generating mechanism given empirical IPD marginal moments and correlation matrix.

Key to IPD inference reconstruction is to see NORTAmax resampling as a form of bootstrapping procedure – we call it MaxEnt bootstrap – taking the above empirical IPD summaries as only input data. The MaxEnt bootstrap can be relevant for the following.

Confirmation of reported IPD inferences.

Production of new IPD inferences.

IPD hypothesis testing (confirmatory and new).

IPD pooled estimation and hypothesis testing (confirmatory and new).

The premise "as if the original IPD was available" is understood implicit above. Such applications suggest MaxEnt bootstrap could be useful in SDC, meta-analysis, distributed computing, research synthesis and reproduction, when no original IPD is available.

Something similar to the term 'MaxEnt bootstrap' is already used in the time-series literature (Vinod, 2004, 2006). We think our method is general enough to include reconstruction of time-series inferences, although we do not provide empirical support for this claim. A next step should be to replicate Vinod's results with our method. Below we review and discuss formal and methodological aspects in more detail. A discussion of the generated empirical evidence follows. We proceed by discussing limitations, future directions, and connections to other work. We finish with some concluding remarks.

4.1.1 Formal issues

While this work is born applied we make an effort to give a formal argument behind the idea of IPD and IPD inference reconstruction, given that only certain empirical IPD summaries are available. The proposed explanation is obligingly limited and sometimes only includes conjecturing. Many intuitions are simply left to stand empirical validation later.

Our basic idea is to use information theory for the purpose of IPD reconstruction. Here known results such as the Conditional Limit Theorem provide a necessary basis, but are not alone sufficient to accomplish the purpose except when IPD marginals are independent. Our main contribution is to extend the basic formalism to cover the case of between IPD marginals dependence that is practically more relevant. We do this by seeing a connection between maximum entropy and copula theory.

We argue that a Gaussian copula with given MaxEnt marginals and fixed dependence structure identifies the joint MaxEnt distribution (Theorem 2.1.1). We show such Gaussian copula is the limit distribution of the generating IPD mechanism given empirical IPD marginal moments and correlation structure (Theorem 2.1.2). This argument uses a convenient copula factorization under continuous and differentiable marginals. This suggests that extending the Conditional Limit Theorem to the multi-dimensional dependent situation – a seemingly unavailable result – could be pursued that way, via copula theory, that is of independent interest.

If some marginals are discrete there could be several configurations of the Gaussian copula yielding the same joint distribution, intuitively of equal maximum entropy that shall not be an issue in practice. However under discrete marginals the used copula factorization is not longer possible and a generalization of Theorem 2.1.2 seems challenging. We suggest that continuation techniques proposed by Durante et al. (2012) or Faugeras (2013) could be here used to extend Theorem 2.1.2 to include discrete marginals (Conjecture 2.1.1). To avoid confusion we stress this continuation step must not find practical application. In our experiments we reconstruct discrete IPD marginals with no such continuation step and obtain reassuring results.

For what concerns IPD inference reconstruction we contribute with the main idea that MaxEnt Gaussian copula inversion defines a bootstrap estimator, where the MaxEnt distribution is the plug-in approximation for the data-generating mechanism given IPD empirical summary constraints.

We call this the MaxEnt bootstrap estimator, that as far as we know it is yet not proposed in the literature. Mimicking consistency proofs for the ordinary bootstrap we give intuitive consistency arguments for our MaxEnt bootstrap estimator in very simple cases (consistency of the mean, Proposition 2.1.10 and Conjecture 2.1.2). Here further work to strengthen this argument needs a more rigorous assessment of the distance between empirical MaxEnt frequencies and the classic e.d. estimate.

We conjecture the MaxEnt bootstrap average should be close to the original IPD inference that is practically relevant for IPD inference reconstruction (Conjecture 2.1.3). Intuitively this claim follows from the distributional constraints imposed on the plug-in distribution, by which resampling tends to more systematically occur around an IPD configuration similar to the original IPD on average. An extensive consistency assessment for the MaxEnt bootstrap estimator was out of scope here but empirical validation later support our intuitions, suggesting more theoretical work could be worth.

4.1.2 Methodological issues

IPD reconstruction is based on possession of empirical IPD marginal moments and Pearson correlation matrix as the only input data. IPD is reconstructed via a Gaussian copula (NORTAmax) identifying a joint MaxEnt distribution. NORTAmax optimization has typical cost $O(pk + p(p-1)/2)$ where k is the highest moment degree. If an analytic solution for the MaxEnt marginal is available we have $k = 0$.

There is only one Gaussian copula defined by the given IPD correlation matrix. However the Gaussian copula correlation structure is not assured to be semi-positive definite (see Appendix A.1.3, page 83). In our experiments Gaussian correlation optimization based on a product-moment estimate seems more stable, which is intuitive here. This is not a restriction on the format of the IPD empirical matrix and Clemen and Reilly (1999) give analytic transformations from common correlation indexes to the product-moment one. We confirm a simple Newton-Raphson routine suffices to find an IPD Pearson correlation projection into standard Normal space. Here optimization is conveniently broken down into separate two-dimensional tasks. However we need modifications to ensure semi-positive definiteness of the copula dependence structure.

We modify the NR-search with Algorithm A.1.1, page 82, that marks out-of-range s.n. correlation entries for later adjustment. Next Algorithm A.1.4 (page 84) finely tweak (marked) matrix entries deterministically or stochastically until the s.p.d. condition is fulfilled. This approach seems to work well without strongly altering the copula correlation structure, but this should be further tested under higher data dimension (p big).

We assess IPD reconstructions under different implementation strategies reflecting the amount and completeness of the input empirical IPD marginal moments and correlation matrix. By convention we replace the continuous MaxEnt marginal with the Johnson distribution. This choice is well grounded in Proposition 2.1.4, page 14, and inherently assess robustness of the method when some marginals only crudely approximate the MaxEnt distribution. We observe minor inconsistencies during Johnson resampling where negative quantiles are sometimes drawn from a positive

Johnson type (see Appendix A.1.1, page 79). This is most probably a software-related problem, but variable domain in the Johnson system is extrapolated by the optimization routine which could be prone to errors.

The next step should be to replace the Johnson approximation with the canonical MaxEnt solution – see Equation (2.10), page 12. At the time of writing there seems to be no readily avaialble software to optimize (2.10). A number of optimization approaches are available and that of Rockinger and Jondeau (2002) or Holly et al. (2011) seem promising. Optimization routines accomodating for more than four moments seem difficult and more work in this direction seems yet needed. It also remains to show how well a (approximate) MaxEnt marginal may generalize to irregular, non unimodal, IPD marginals such as, for instance, genetic sequencing data. In practice Pearsonian family distributions can flexibly reproduce non strictly unimodal features (Johnson, 1949). Also the general MaxEnt solution covers in principle a rahter large domain (Holly et al., 2011). Our empirical observation (not shown) suggests that as far as the IPD marginal variance (hence entropy) is contained, good information recovery is possible quite regardless of the exact density shape.

Our alternative to NORTAmax resampling is a routine that only processes incomplete correlation information while maintaining the same approximate MaxEnt marginals. In cases where only first-order IPD correlations are available, Algorithm A.2.1, page 86, reconstructs such incomplete dependence based on the mixture (A.17) and on a permutation procedure to find null correlated pair combinations. While this approach is never superior to NORTAmax, Algorithm A.2.1 could still be a simple alternative if the goal is a two-dimensional IPD reconstruction. A typical situation would be where marginal moments of only two variables, one outcome and a covariate along with their raw correlation index, are reported. Studies on such small IPD are infrequent but if the reported covariate is a compound index simultaneously accounting for several confounders, or risk-factors, then such small IPD is sometimes employed.

In a similarly incomplete summary reporting scenario of a two-arms survival IPD we might only recover arm-specific total number at-risk and events-count, but not the correlation between event-counts and treatment. Here an attempt to reconstruct the IPD Hazard Ratio for treatment could be via maximization of (1.2), page 5. Appendix B.5.1, page 95, proposes a simple approach to reconstruct the IPD at-risk set vector to insert in (1.2). We empirically see that such approximation can often well recover the HR (not shown). This may be particularly true if only administrative or low rate censoring occurs. In all other cases this simple at-risk sets reconstruction considers either a scenario where censored units remain infinitely at risk (sub-distribution hazard case) or one where they all leave the at-risk set before study end. In presence of right, non-administrative, censoring both these scenario may bias HR reproduction. If it is known that generic right censoring occurs one way to mitigate bias could be to average between the stay-in and stay-out cases, hoping this roughly approximates the original censoring process.

IPD reconstruction experiments were performed by re-arranging about 20 original IPDs into four data batches by different number and type of covariates. We kept the overall dimension quite low and never exceeding four variables per original IPD. Such choice was motivated by convenience and extension to higher dimension seems possible. The original IPD examples are mostly

chosen open source and reflect the complexity and variety or real data. For instance a number of examples from Royston and Sauerbrei (2008) purposely include original survival data where neither the baseline hazards are constant nor the PH assumption is met. To asses overall similarity between simulated and original IPD we define a number of marginal and dependence similarity criteria to be simultaneously fulfilled (Section 2.2.4, page 38). That is we simply check the distance between expected IPD distributional features against their original value. If only one marginal moment or correlation entry fails to match on average the original IPD value, the IPD simulation is deemed overall not similar to the IPD reference. As discussed later this rather conservative diagnostic approach may tend to inflate false negative outcomes relative to the actual quality of the recovered IPD information content.

To study NORTAmax ability to recover original IPD inferential content we consider a number of statistical procedures to be applied on reconstructed IPD.

Parametric: GLMs with Gaussian, Binomial, or Poisson family.

Semi-parametric: PH Cox modeling, and Breslow estimation.

Non-parametric: Nelson-Aalen estimation.

IPD inference recovery is assessed in a general multi-variate case but also in non trivial univariate ones. It is well known (Olkin and Sampson, 1998) that certain IPD GLM reconstructions from summary data allows only contained information loss, if only a single categorical covariate is used. In this respect it is interesting to evaluate parametric GLM reconstruction if the single covariate is continuous and we do so. Similarly we assess PH Cox regression reconstruction starting with inclusion of a single binary covariate. Next the number and types of covariates is increased. In the parametric and semi-parametric cases the mode of inference estimation slightly differs from the original IPD application because we also specially handle information entering the GLM or Cox likelihood numerator. We rewrite likelihood routines to allow explicit numerator control (Appendix B.1, page 89). Also, to speed up partial Cox likelihood and c.h.e. calculations we use a slightly different at-risk set computation (Equation (B.7) and Appendix B.3, page 91 and 93). Where possible, we use log-likelihood gradient and Hessian linearization to speed up optimization (Appendix B.1.2, page 91). We use a simple averaging procedure for the reconstructed c.h.e. (Appendix B.4, page 94) to make comparisons with the original IPD estimates possible.

We use a simple difference between reconstructed and original IPD inference to assess bias. While this choice is straightforward it could also be too conservative. In our experiments it is not unusual to see Gaussian intercept estimates with a value of two order of magnitude. Say the original IPD intercept estimate is 281 and the reconstructed one is 282. Here a distance of 1 is maybe big relative to zero but seems quite irrelevant in practical terms.

4.2 Empirical results

We could satisfactorily reconstruct IPD from its marginal moments and correlation matrix only, and, as a second step, reconstruct original IPD inferences from the reconstructed IPD. Generally the results seem to empirically support our formal arguments. First, NORTAmax resampling well reproduces key features of the empirical IPD distribution. This practically agrees with NORTAmax resampling being asymptotically equal to draw from the IPD joint distribution given IPD empirical distributional summaries (Theorem 2.1.2). In particular, we can empirically confirm this limiting result well extends to mixed binary-continuous data (Conjecture 2.1.1). Second, statistical transformation of a NORTAmax sample, the MaxEnt bootstrap, well recovers important features of the IPD inference distribution. This seems to confirm claims of MaxEnt bootstrap consistency (Proposition 2.1.10 and Conjecture 2.1.2) in even more general situations. Results also confirm approximate equality of the MaxEnt bootstrap average to the reference original IPD inference value (Conjecture 2.1.2). Below we discuss results in more detail.

4.2.1 IPD reconstruction

We could stochastically reconstruct (undisclosed) IPD from (ideally disclosed) IPD marginal moments and correlation matrix via NORTAmax resampling. The degree of reconstruction honesty seems to reasonably depend on the amount and completeness of the available IPD summaries. Reconstruction performance seems better when moments up to fourth degree and a complete (lower triangular) correlation matrix are both available. However the strength of this result varies depending on number and type of IPD marginal variables. We did not practically see copula indeterminacy issues related to usage of discrete variables. Occasional non semi-definite positive Gaussian copula correlation matrix could be satisfactorily repaired with our proposed heuristic. Ultimately most issues were posed by continuous marginals which we attribute to Johnson resampling and/or related software implementation.

We observe an excessive poor Johnson performance in about 10% of reconstructed IPD continuous marginals that mostly affects experiments in batch III and IV. Especially in batch IV the fourth moment is badly recovered, that suggests Johnson resampling could be currently problematic in higher dimensional cases. Good fourth moment recovery might not always be necessary but it could negatively impact IPD correlation recovery that is more important for IPD inference reconstruction. Nevertheless for growing n the fourth moment recovery is relatively tolerable in at least half the cases.

Even with Johnson resampling issues we think the obtained IPD emulation is overall good, suggesting method amelioration is possible and worth pursuing, for instance by usage of the canonical MaxEnt solution (Equation (2.10)). Such level of IPD emulation from simple IPD summaries is rather unexpected and sort of defies common sense. To the best of our knowledge NORTAmax resampling seems one of the first procedures to generally reconstruct dependent multi-variate original IPD from simple IPD summaries only.

4.2.2 IPD inference reconstruction

We could reconstruct a variety of commonly used IPD inferences from generated IPD emulations. The results bares practical benefits for disciplines like statistical results reproduction and distributed network computing, where no original IPD but only key IPD summaries can be disclosed. Meta-analysis and research synthesis could also greatly benefit from the method, provided the IPD summary format we suggest becomes more standardized. Hence via MaxEnt bootstrapping (MaxEntBoot) we could well recover information on:

a) (multi-variate) MLEs from Gaussian, Binomial, and Poisson GLMs, as well as PH Cox regression HRs.

b) Breslow or Nelson-Aalen cumulative hazard estimates.

c) Bootstrap-like 95% empirical CIs for estimates of point a).

The MaxEnt bootstrap average is generally well centred on the original IPD inference value and 95% interval quantiles are quite comparable to the IPD ordinary bootstrap. Similarly average cumulative hazard estimates well approximates the original IPD graph. All these observations seem to empirically support a more general notion of MaxEntBoot consistency that extends simpler claims made in Proposition 2.1.10 and Conjecture 2.1.2.

Standard MLE recovery methods mostly focus on a scalar value – typically a group effect – and are based on appraisal of (highly transformed) IPD summaries from different study sources. Our results imply that, for each study source, one can reconstruct study-specific IPDs and pool them to reproduce a fixed or random effect multi-variable IPD regression, that can be relevant for meta-analysis and research synthesis. For example one can Cox regress the pooled reconstructed IPD, introducing a study-specific baseline hazards or a frailty effect. Similarly one can apply a random-effects GLM regression. We yet need empirical confirmation for such random effect estimate recovery, but we think this task is feasible as far as the study-specific IPD is well reconstructed.

Standard cumulative hazard estimates recovery methods typically focus on a manual, time-consuming, extrapolation of published Kaplan-Meier or Nelson-Aalen graphs (Parmar et al., 1998). Our results imply a general (pooled) cumulative hazard estimation recovery is possible based on reported empirical IPD summary only. In such regard we are not aware of comparable methodologies in the literature. Our method can rather accurately recover IPD Breslow or Nelson-Aalen cumulative events-counts that we think is kind of surprising. Recovery of IPD events-time lines is more difficult because MaxEntBoot must well predict the time value at which an event-jump occurs. However, our experiments suggest time-line recovery is acceptable on average which results in an overall good c.h.e. reconstruction (see Figure 3.7 and 3.8, page 65 and 66, and Figure C.1 to C.3, page 139 to 141) that seems a rough interpolation. Our c.h.e. reconstructions lack appropriate confidence bounds estimates. Here we could plot all the c.h.e. bootstrap realizations under the expectation estimate to give an assessment of its range of variability.

Recovery of 95% bootstrap CIs requires original IPD and our method seems one of the first to reconstruct bootstrap-like 95% intervals from IPD summaries only. The MLE MaxEntBoot variance also typically agrees with the IPD point estimate based on Fisher Information. Here empirical results seem to suggest MaxEnt bootstrap consistency is holding quite generally, begging the question if a Central Limit Theorem rule of thumb similar to the ordinary bootstrap's holds here too. We confirm our MaxEnt intervals can be on average slightly tighter than their ordinary IPD counterparts (not shown). This seems to reflect the preference of the MaxEnt bootstrap to resample more closely around the expected inference, given its more constrained nature.

IPD inference reconstruction is generally better under NORTAmax approaches (Algorithm 2.1.1–2.1.2) with no likelihood numerator adjustment (methods 1-3 and 1-4). Here the difference between reconstructed and original IPD inference is generally well centered on zero. Difference dispersion is also generally acceptable relative to mean and s.d. of the original IPD inference. Likelihood-numerator adjustment could be beneficial, if no complete IPD correlation is available. Based on our similarity criteria third and fourth moment knowledge seems relevant for good IPD reconstruction, but results on IPD inference reconstruction show another picture.

Generally essential for good IPD inference reconstruction seems accurate and complete IPD correlation matrix recovery. Higher moments information often only refines IPD inference reconstruction performance. Our experiments show that knowledge of IPD marginal mean and variance only, along a complete correlation matrix, is often enough to roughly well recover an IPD inference. In this respect sub-optimal IPD reconstructions in batch IV could be an artifact of low specificity of point 1 to 5 of Section 2.2.4, suggesting the quality of recovered IPD information might be often better than what we diagnose.

We give direct examples of IPD inference reconstruction (see Section 3.3 and 3.4, page 59 and 62, and Appendix C.3.3 to C.3.5, page 136 to 138). We show the MaxEnt bootstrap is not only useful when original IPD is unavailable but also when the original IPD inference is of little use. For example a singular original IPD inference can be often caused by a sparse data bias (Greenland et al., 2016). Here the MaxEntBoot prediction is a long-run alternative for the original IPD inference that can dilute the original bias. We do not compare our method with the penalized estimation described in (Greenland et al., 2016) and simply acknowledge our results are also reasonable here.

In the example of Appendix C.3.3 we model the original discrete, tied, IPD time variable on a continuous scale. This produces a slightly positive significant MaxEntBoot group effect against a slightly negative non significant original IPD one. The difference is not big but appreciable. Also the reconstructed Nelson-Aalen curve under treatment grows somehow faster than the original estimate beginning from time point 20. The MaxEntBoot Fisher information is remarkably close to the IPD value on average. These discrepancies shall be understood as the attempt to smooth the original IPD time variable. Here reproduction of the original time tied structure needs more care and maybe introduction of non-standard constraints.

Overall our experiments show the MaxEnt bootstrap allows qualitatively good IPD inferences reconstructions from simple IPD summaries only. Such result would not be typically expected a priori and the MaxEnt the bootstrap seems one of the first procedures that accomplish doing so.

4.3 Limitations and further directions

NORTAmax resampling can reconstruct IPD and IPD inferences if only IPD marginal moments and correlation matrix are available, that could be useful in meta-analysis, research reproduction, and synthesis. In practice it is realistic to assume we might more easily recover mean and variance of some (not all) IPD marginals. However higher moments and, most crucially, pairwise correlations are today not often reported, that is a limitation of our proposed method. One way to mitigate the problem may be suggested by Yoneoka and Henmi (2016) who indirectly recover some variables correlations from easily recoverable summaries.

We do not show experiments on a practical meta-analytic example. Here one issue to address is missing IPD marginals for some study sources that would make IPD pooling problematic. Kohnen and Reiter (2009) could provide some ideas in the merit. We also do not show IPD regression reproduction with effects interactions. We believe such task should be feasible as far as overall IPD reconstruction is good.

It remains yet to assess NORTAmax robustness if the original IPD has a longitudinal structure. In theory NORTAmax does not apply here but we could try the following. Assume a categorical variable encodes the IPD clustering structure, with each category level having enough observations. Let the correlation between the categorical variable and any other IPD marginal be included in the given IPD correlation matrix. A MaxEnt analytic solution for the IPD categorical marginal is the Multinomial distribution and we apply NORTAmax resampling accordingly. This is similar to handle survival IPD with multiple events in one single block where the categorical variable is the event-specific indicator. In our experiments we recover IPD HRs and c.h.e.'s for each event or transition separately. This is sufficient for a complete competing risks or event-history analysis (Latouche et al., 2013)

Copula marginal smoothing discussed in Section 4.1.1 suggests NORTAmax could be based on a discrete MaxEnt marginal's continuous approximation. Beta and Gamma distributions respectively approximate Binomial or Poisson marginals after careful re-scaling of moment-based parameter estimates. These approximations are then rounded up and discretized which shall conserve marginal features and maybe most of inter-dependence structure. This would save computation time drastically since Gaussian copula correlation has analytic solution here. Preliminary observation (not shown) suggests such approach needs care and could be sub-optimal. For instance approximation of a Multinomial MaxEnt marginal with a Dirichlet distribution seems not straightforward here.

In Section 2.1.7 we propose to use NORTAmax for missing data imputation under MAR assumption. Here missing records are asymptotically drawn from the IPD generating mechanism under constraints on its empirical summaries, but the MAR condition seems difficult to assess from empirical IPD summaries only and more IPD information could be needed. If NORTAmax resampling poorly matches the given constraints it is a more or less biased limit description of the observed IPD and careful results interpretation is warranted. Also we implicitly assume empirical marginal moments and correlation matrix capture all relevant constraints on the current IPD. However this must not be the case and a more specific constrained optimization could be needed.

4.4 Connection to other topics

MaxEnt bootstrap is a relatively straightforward type of autoencoder where data is reduced on its first rather than second dimension. We use maybe one of the simplest encoding operation by compressing raw IPD into its marginal moments and correlation matrix $\bar{C}_x$. The joint MaxEnt distribution is the decoder. As any autoencoder the MaxEnt bootstrap is a generative model drawing samples from the IPD distribution satisfying constraint $\bar{C}_x$, that is the only input data needed for loss minimization (point 2 and 3 of Algorithm 2.1.1) based on a simple distance between $\bar{C}_x$ and its theoretical model values.

Salakhutdinov and Larochelle (2010); Fisher et al. (2018) propose Neural Networks (NNs) that among other useful properties could generally approximate the p-dimensional generating law of the input data. This is direct consequence of the Hammersley-Clifford theorem (Grimmett, 1973) that justifies representation of a generic probability distribution via a Markov network (field). This probability law has the form of a Gibbs or Boltzmann distribution that is the MaxEnt solution satisfying given network constraints (Robert, 1990).

Hence both these NNs and the MaxEnt bootstrap are energy-based generative models based on a MaxEnt principle, but the NNs need a relatively sophisticated implementation. Here MaxEnt bootstrap could be a simpler but more precise alternative to sample from the data generating distribution, although under the clause $p << n$. We could force a $p \geq n$ condition by artificially inflating n to promote simulation stability. Optimization costs should be lower than in NNs (see discussion in Section 4.1.2) where at least the number of layers and of layer-specific units must also be factored in the total expense. Also MaxEnt bootstrap is a generative model for the inference associated with the input data, that seems an advantage relative to the NNs especially in respect to hypothesis testing and predictive inference.

It is maybe interesting to read some of our results in connections to generic notions of evidential equivalence. The weak conditionality principle (C') (Birnbaum, 1962; Berger et al., 1988) is sometimes interpreted as the irrelevance of hypothetical evidence on the meaning of actually observed evidence (Cox, 1958). Here evidence means some type of inference or inferential content (Basu, 1975). A strengthened derivation of (C'), denoted (C), is sometimes seen as a directive for explicit conditional inference (Kalbfleisch, 1975; Fraser et al., 2004) and an argument against use of purely frequentist inference. In Conjecture 2.1.3, page 23, $\mathscr{M}(x)$ is actual evidence from observed data x, and the left term of (2.30) is expectation over hypothetical evidence. Then we can read (2.30) as an operative (C'), denote it (C*), where we introduce an averaging step over hypothetical evidence. Hence (C*) suggests rough equivalence between actual inference and expectation over hypothetical inferences that seems to generalize and help explain the non-directionality notion behind (C') and (C), as discussed in Kalbfleisch (1975), page 263, and Dawid (2014). Here (C*) seems to connect seemingly antithetic inferential notions similar to Efron (2012).

A related reading of (2.30) is that of random perturbation of the original IPD inference in the sense of (Yu et al., 2013). This seems also related to a generic idea of white noise irrelevance similar to Birnbaum (1964). That is, all essential inferential content, or signal, is locked in the constrain summary $\bar{C}_x$ and everything else being noise. The joint MaxEnt distribution is the tool to

re-propagate this noise or incertitude around the signal. A worrisome consequence of such view is that any data instance roughly described by $\bar{C}_x$ may yield approximately equal statistical evidence. Then one could deceitfully forge statistical evidence by fine tuning of $\bar{C}_x$.

4.5 Conclusions

We present a maximum entropy based resampling method (MaxEnt bootstrap) to reproduce reliable multi-variate IPD simulations from the following IPD summary data only:

marginal moments up to degree two (or at best four),

a complete correlation matrix lower (upper) triangular.

We show it is possible to roughly recover original IPD inferences from these IPD simulations by rather well reconstructing

multi-variable MLEs from GLM and Cox regression with bootstrap-like 95% CIs.

Nelson-Aalen or Breslow cumulative hazard estimates.

The method has potential application in a number of statistical disciplines where only summary IPD can be used but original IPD inferences are sought. This includes statistical disclosure control, meta-analysis, research reproduction and synthesis, as well as distributed network computing to name few. Our method is readily applicable in data anonymization and distributed network computing upon compliance to the above IPD summaries. Instead the remnant disciplines would greatly benefit from the method if the above IPD summaries would become more standardly reported. Many authoritative sources call for more reporting standardization (Liberati et al., 2009; Altman, 2015; Nature, 2017) and adoption of the above IPD summary format could comply to such purpose.

To the best of our knowledge our seems one of the first methods accomplishing generic IPD and IPD inference reconstruction from usage of the above IPD summaries only.

Appendix A

Method details: IPD reconstruction

A.1 NORTAmax scheme: details

We give further details about NORTAmax resampling (Section 2.2.1, page 35). Refer to Section D.1, page 143 for all mentioned `R` packages.

A.1.1 Johnson system distributions

The Johnson system is defined by the following general transformation of x,

$$z = \gamma + \delta f(y), \tag{A.1}$$

where $z \sim \mathcal{N}(0, 1)$, and

$$y = \frac{x - \xi}{\lambda}, \tag{A.2}$$

with $\delta > 0$, $\lambda > 0$, $\gamma \in \mathbb{R}$, and $f(\cdot)$ defining one of the following transforms: 1) log-normal (SL), 2) unbounded (SU), 3) bounded (SB), 4) normal (N). The parameters γ, δ, λ, ξ and the transform f must be determined from a set of first four moment constraints of x, $m(x)$, $m(x^2)$, $m(x^3)$, $m(x^4)$. The Hill algorithm (Hill et al., 1976; Hill, 1976) implemented in the `JohnsonDistribution` package provides the numerical tool to determine these unknowns. Once a solution for the unknown is found sampling from a Johnson distribution needs inversion of (A.1) as a function of a s.n. variate z. Below I describe in more detail each system.

SL system

Here $f(y) = \log(y)$, $y > 0$ where y is given by (A.2) and the density function is

$$p(y) = \delta \exp\left(-\frac{z^2}{2}\right) c, \tag{A.3}$$

with $c = 1/(y\sqrt{2\pi})$, and z as (A.1) after proper specification of f with $\xi < x$. To sample a quantile x from this distribution draw a s.n. variate and transform with

$$x = \xi + \exp\left(\frac{z-\gamma}{\delta}\right). \tag{A.4}$$

SU system

Here $f(y) = \log(x + \sqrt{1+y^2})$ with y as (A.2) and density (A.3) but with

$$c = \frac{1}{\sqrt{2\pi}\sqrt{1+y^2}}. \tag{A.5}$$

We can sample x using transform

$$x = \xi + \lambda \sinh\left(\frac{z-\gamma}{\delta}\right). \tag{A.6}$$

SB system

Here $f(y) = \log(y/1-y)$, $0 < y < 1$, with y as (A.2) and density (A.3) for $\xi < x < \xi + \lambda$ with

$$c = \frac{1}{\sqrt{2\pi}y(1-y)\lambda}. \tag{A.7}$$

We can sample x using transform

$$x = \xi + \lambda \sinh\left(\frac{z-\gamma}{\delta}\right). \tag{A.8}$$

All Johnson resampling is performed using package `moments` and `JohnsonDistribution`. The latter occasionally yields few negative values under bounded transformation (A.6). As a possible explanation for this the package author privately confirmed minor errors in the `FORTRAN` code. Nevertheless, these minor inconsistencies are never serious enough to compromise the module's functionality and can be easily handled by uniformly setting all negative values, if any, between zero and the minimum non-zero value of the resampled series.

A.1.2 Matrix correlation conversion: optimization

We give more details about optimization mapping IPD empirical correlation matrix $\bar{R}_x$ into a s.n. counterpart R_z with off-diagonal entry $\rho_{j\ell}$ satisfying Equation (2.23), page 17, $\forall\, j \neq \ell$.

Double integration techniques

We need to compute double integral (2.23). We substitute the Ψ-transform in (2.23) with

$$\Psi_j^* = \begin{cases} G_j^{-1}(z) & \text{if } G_j \equiv J_j \\ G_j^{-1}\left(\Phi(z)\right) & \text{otherwise} \end{cases} \tag{A.9}$$

where z is a s.n. variate, $\sim J_j$ is the Johnson distribution, and Φ is the cumulative s.n. distribution. We need Equation (A.9) to accommodate for the Johnson variate (see Appendix A.1.1). (2.23) may be solved for $\rho_{j\ell}$, $j \neq \ell$ using adaptive multivariate integration over hypercubes (package `cubature`). Here it is better to set a finite $z \times z$ range, say $(-5, 5) \times (-5, 5)$.

Occasionally adaptive integration may be unstable or fail. Monte Carlo integration it is an easy alternative here

$$E_G(X_j, X_\ell) \approx \frac{1}{K} \sum_i^K \Psi_j^*(Z_{i1}) \Psi_\ell^*(Z_{i2}), \tag{A.10}$$

for $j \neq \ell$, with Ψ_j^* as (A.9) and $(Z_{i1}, Z_{i2}) \sim \mathcal{N}_2(0, \rho_{j\ell})$ is a draw from the 0-mean bivariate Normal distribution with correlation $\rho_{j\ell}$, for $i = 1, 2, \ldots K$ and $\forall\, j \neq \ell$. Here $K = 10000$ or smaller can be enough.

Minimization: objective function

We wnat to minimize the distance between an empirical correlation $\bar{r}_{x_j x_\ell}$ and

$$\text{Cor}_\phi \left(\Psi_j^*(z_1), \Psi_\ell^*(z_2) \right), \tag{A.11}$$

relative to the bivariate s.n. density $\phi = \phi(z_1, z_2 | \rho_{j\ell})$ with parameter $\rho_{j\ell}$ with Ψ_j^* as (A.9) $\forall\, j \neq \ell$.

To obtain (A.11) subtract the right term of (2.23) by $E_{G_j}(X_j) E_{F_\ell}(X_\ell)$, the product of theoretical first moments, and divide all by $\sigma_{x_j} \sigma_{x_\ell}$, the product of theoretical standard deviations, where $\sigma_{x_j} = E_{G_j}(X_j^2) - E_{G_j}(X_j)^2$, $\forall\, j \neq \ell$. If these theoretical moments are unknown replace them with sample estimates. Our objective function is

$$\mathscr{O}(\rho_{j\ell}) = \text{Cor}_\phi \left(\Psi_j^*(z_1), \Psi_\ell^*(z_2) \right) - \bar{r}_{x_j x_\ell}, \tag{A.12}$$

where we must find the s.n. correlation $\rho_{j\ell}$ such that (A.12) is zero. The first derivative of $\mathscr{O}(\rho_{j\ell})$ relative to $\rho_{j\ell}$ is

$$\mathscr{O}'(\rho_{j\ell}) = \frac{1}{\sigma_{x_j} \sigma_{x_\ell}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Psi_j^*(z_1) \Psi_\ell^*(z_2) \phi' dz_1 dz_2, \tag{A.13}$$

where $\phi' = d\phi(z_1, z_2 | \rho_{j\ell}) / d\rho_{j\ell}$ equals

$$\phi' = \frac{\mathscr{Q}(\rho_{j\ell}) \phi(z_1, z_2 | \rho_{j\ell})}{2\pi \sqrt{1 - \rho_{j\ell}^2}}, \tag{A.14}$$

with

$$\mathscr{Q}(\rho_{j\ell}) = \left\{ \frac{c(\rho_{j\ell})\left(2\pi\sqrt{1-\rho_{j\ell}^2}\right)}{\left(2(1-\rho_{j\ell}^2)\right)^2} \right\} - \left\{ \frac{-2\pi\rho_{j\ell}}{\sqrt{1-\rho_{j\ell}^2}} \right\},$$

where

$$c(\rho_{j\ell}) = \left[4z_1z_2\left(1-\rho_{j\ell}^2\right)\right] - \left[4\rho_{j\ell}\left(z_1^2 + 2\rho_{j\ell}z_1z_2 - z_2^2\right)\right],$$

All integrals can be numerically solved using adaptive integration as mentioned above. Alternatively, to evaluate (A.11) by MC integration we similarly subtract then divide (A.10) by theoretical or sample first moments and standard deviations product respectively. An approximate objective function is obtained with appropriate substitution in (A.12). Then an MC approximation of (A.13) is

$$\sigma_{x_j}\sigma_{x_\ell}\mathscr{O}'(\rho_{j\ell}) \approx \frac{1}{K}\sum_{i}^{K} \Psi_j^*(Z_{i1})\Psi_\ell^*(Z_{i2})\frac{\mathscr{Q}(\rho_{j\ell})}{2\pi\sqrt{1-\rho_{j\ell}^2}}, \tag{A.15}$$

where $\mathscr{Q}(\rho_{j\ell})$ is as above after appropriate substitution with draw $(Z_{i2}, Z_{i2}) \sim \mathcal{N}_2(0, \rho_{j\ell})$ for $i = 1, 2, \ldots, K$. Again $K = 10000$ or smaller should suffice.

Newton-Raphson optimization

To find an approximate zero of (A.12) we implement a Newton-Raphson (NR) scheme. A good starting value for $\rho_{j\ell}$ is the empirical observation $\bar{r}_{x_jx_\ell}$. At iteration i NR optimization is defined by the updating step

$$\rho_{j\ell}^{(i+1)} = \rho_{j\ell}^{(i)} - \frac{\mathscr{O}(\rho_{j\ell}^{(i)})}{\mathscr{O}'(\rho_{j\ell}^{(i)})}, \tag{A.16}$$

where on the right term ratio we have (A.12) at the numerator and (A.13) at the denominator with the requirement that $0 \leq \|\rho_{j\ell}^{(i+1)}\| < 1$ where $\|\cdot\|$ denotes absolute value. In particular exclusion of the unity upper bound is required to avoid an infinite solution of (A.11).

Occasionally the NR search might get stuck around a flat slope that is not a minimum. For instance, this seems to occur more often when marginal models G_j and G_ℓ are strongly misspecified w.r.t. the empirically observed correlation. Then we might obtain illegal solution values. We take some simple precautions to mitigate this issue if occurs, by either forcing the search to re-start from legal values, or by stopping the search around a flat slope.

Algorithm A.1.1. (Safe-check snippet embedded in NR optimization).

If $\|\rho_{j\ell}^{(i+1)}\| \geq 1$

1) Repair value $\rho_{j\ell}^{(i+1)}$ and flag:

if $\|\rho_{j\ell}^{(i+1)}\| = 1$, set $\rho_{j\ell}^{(i+1)} = \text{sign}(\rho_{j\ell}^{(i+1)}) \cdot 0.99$,

else

1.1) produce coarse sequence r'_z with range $\text{sign}(\rho_{j\ell}^{(i+1)}) \cdot [.74, .99]$.

1.2) set $\rho_{j\ell}^{(i+1)} = \tilde{r}_z = \min_{r'_z} \mathscr{O}(r'_z)$.

2) Go to next NR iteration

if $\mathscr{O}'\left(\text{sign}(\rho_{j\ell}^{(i+1)}) \cdot 0.9\right) < 0.01$ stop NR search and return value $\rho_{j\ell}^{(i+1)}$.

$\mathscr{O}(\cdot)$ and $\mathscr{O}'(\cdot)$ are defined in Appendix 7.1 of the main document.

The middle step at point 2 kills the search if around a solution's extremal value the slope gets already flat. Such aborted solutions are flagged in the s.n. matrix and checked in a second phase (section below) especially w.r.t. to satisfaction of the semi-positive definite condition.

A.1.3 Matrix correlation conversion: insurance of semi positive definite condition

We observe that s.n. matrix non semi-positive definiteness is often associated with aborted solutions returned from Algorithm A.1.1. The s.p.d. condition can also be not simply met when all NR searches hit a true minimum. We adopt a simple approach to fix a non s.p.d. matrix, distinguishing between a non s.p.d. matrix having also aborted (flagged) solutions or not.

Flagged matrix

Since the s.p.d. condition might be not met due to matrix values resulting from aborted NR searches, we target only flagged matrix entries and tweak their values until the s.p.d. condition is met.

Algorithm A.1.2. (Targeted tweaking).

Take matrix lower, or upper, triangular:

1) Select flagged element $\grave{\rho}_j$, $j = 1, \ldots, p_0$ with $1 \leq p_0 \leq \left(\frac{p(p-1)}{2}\right)$.

2) Substitute $\grave{\rho}_j$ with $\rho_j = \grave{\rho}_j - \left[\text{sign}(\grave{\rho}_j) \cdot 0.005\right]$, for all $j = 1, \ldots, p_0$.

3) Repeat point 2) until resulting matrix is s.p.d., or exit Algorithm after 200 iterations.

We typically observe a result within 200 iterations. This means we down-scale initial value $\grave{\rho}_i$ by at most one unit. This agrees with aborted solutions from Algorithm A.1.1 having typically too high values.

Not flagged matrix

Non s.p.d. solutions can simply result from all NR searches correctly converging. Here the goal is also to down-scale extremal matrix values, selected according to a rank-based priority. The approach is more diffuse and includes randomness.

Algorithm A.1.3. (Sequential tweaking)
Take matrix lower, or upper, triangular:

Denote with $k = \frac{p(p-1)}{2}$ the rank of the most extremal element.

1) Sort elements from most to least extremal value, $\rho_{(k)}, \rho_{(k-1)}, \rho_{(k-2)}, \ldots, \rho_{(1)}$.

2) Set $i = 0$.

   2.1) Substitute $\rho_{(k-i)}$ with $\tilde{\rho}_{(k-i)} = \rho_{(k-i)} - [\text{sign}(\rho_{(k-i)}) \cdot u]$, where $u \sim U(0, 0.01)$.

   2.2) Check if resulting matrix is s.p.d.:

   if yes, stop Algorithm and return s.p.d. matrix,
   else set $i = i + 1$ and return to point 2.1). If $i = k$ return to point 2).

3) Repeat point 2) until resulting matrix is s.p.d. or exit Algorithm after 200 iterations.

$\sim U(0, 0.01)$ is the Uniform distribution between 0 and 0.01.

From our observations the wished result is typically obtained within 200 iterations and a warning is issued otherwise. An additional warning is issued if some sign in R_z changes during the process. As an alternative approach we can implement step 2.1) of Algorithm A.1.3 simultaneously for all not necessarily ranked elements.

General approach

As a general approach to insure the s.p.d. condition we have the following three options

Algorithm A.1.4. (Make matrix s.p.d.)
If s.n. matrix is not s.p.d.

if any matrix elements is flagged apply Algorithm A.1.2 (targeted),

else apply Algorithm A.1.3 (sequential).

- If both above fail apply step 2.1 of Algorithm A.1.3 simultaneously not sequentially.

Typically such fuzzy approach finds a s.p.d. s.n. solution that roughly agrees with the empirically observed correlation constraints.

A.2 Incomplete correlation imposition

We elaborate on material of Section 2.2.2, page 37.

A.2.1 First-order correlation imposition: mathematical derivation

We first derive[1] relation (2.31), page 37. Consider a pair of unrelated variables X_1 and X_2 arranged in the matrix $X = (X_2, X_2)$. We want to induce $\text{Cor}(X_1, X_2) = \rho_{12}$. To this end arrange X such to have minimal correlation. Denote such arrangement with $Z = (Z_1, Z_2)$ where $\text{Cor}(Z) \approx 0$. Similarly arrange X such to have extremal correlation. Denote such arrangement with $U = (U_1, U_2)$ where $\text{Cor}(U) = \Re_{12}$ is the extremal correlation between X_1 and X_2. We require $\text{sign}(\rho_{12}) = \text{sign}(\Re_{12})$ and $\|\rho_{12}\| \leq \|\Re_{12}\|$. Next define the mixture

$$\alpha Z + (1 - \alpha)U, \tag{A.17}$$

for $0 \leq \alpha \leq 1$. Let $\sigma_1 = \sqrt{\text{Var}}(X_1)$ accordingly $\sigma_2 = \sqrt{\text{Var}}(X_2)$. For simplicity set $E(X_1) = E(X_2) = 0$. Set the equality

$$\begin{aligned}\sigma_1\sigma_2\rho_{12} &= \text{Var}\left[\alpha Z + (1 - \alpha)U\right] \\ &= \alpha E(Z_1 Z_2) + (1 - \alpha)E(U_1 U_2) \\ &= \alpha\text{Cov}(Z) + (1 - \alpha)\text{Cov}(U),\end{aligned} \tag{A.18}$$

that by dividing for $(\sigma_1\sigma_2)$ on both sides becomes

$$\begin{aligned}\rho_{12} &= \alpha\text{Cor}(Z) + (1 - \alpha)\text{Cor}(U) \\ &= 0 + (1 - \alpha)\Re_{12} \\ &= (1 - \alpha)\Re_{12}.\end{aligned} \tag{A.19}$$

Thus

$$\alpha = 1 - \frac{\rho_{12}}{\Re_{12}}, \tag{A.20}$$

and mixture (A.17) has correlation ρ_{12}.

A sample version of (A.17) for a $n \times 2$ matrix $X^*_{\ell j} = \left(X^*_\ell, X^*_j\right)$ is given by (2.31), where on the right of (A.20) we set $\rho_{12} = \bar{r}_{x_j x_\ell}$ and repeat Bernoulli experiment $\sim B(\alpha)$ for n independent trials, $\forall\, j \neq \ell$. Consider a $n \times p$ data matrix X^*, $p \geq 2$. For a fixed reference variable X^*_ℓ (say the first data column), $j \neq \ell$, we can reiterate (2.31) for all remnant $p - 1$ columns. This ensures we can easily reproduce all $(p - 1)$ first-order correlations of X^* relative to X^*_ℓ. Reproduction of higher order correlations that simultaneously satisfies all previous correlations needs a computationally unfeasible multi-nested implementation of (2.31). Hence we implement (2.31) for first-order correlations only. We describe such procedure below with details on how to construct Z and U.

[1] I must thank Prof. Ludger Rüschendorf, Department of Stochastic, Institute of Mathematics, Albert-Ludwigs-University of Freiburg, for suggesting these computations.

A.2.2 Permutation-based algorithm for incomplete correlation imposition

We modify Algorithm 2.1.1, page 24, to allow for incomplete correlation imposition in the following way. Keep step 1) and 2) of Algorithm 2.1.1 while adding new steps:

For $b = 1, \ldots, B$

3) Draw $(X_j^*)_b \sim F_j^*$, for all $j = 1, \ldots, p$.

Obtain the uncorrelated $n \times p$ realization X_b^*.

Next we impose incomplete correlation on X_b^*. To this end fix the reference empirical correlation matrix $\bar{R}_x$ with entry $\bar{r}_{x_j x_\ell}$ and fix a reference column X_ℓ^*, $\ell \neq j$, say $\ell = 1$, for all $j = 2, \ldots, p$. Fix a reference correlation function $\mathrm{Cor}(\cdot)$ consistent with $\bar{R}_x$ (see Appendix A.2.3). Use notation $(X_{\ell j})_b = (X_\ell, X_j)_b$ (here sorting/permuting is dimension invariant). Introduce new step

Algorithm A.2.1. (Step 4: incomplete correlation imposition)

For $b = 1, \ldots, B$

4.1) Apply random permutation $(X_{\ell j}^*)_b \mapsto Z_b$ such that $\mathrm{Cor}(Z_b) \approx 0$,

Iteratively apply 4.1) until $\mathrm{Cor}(Z_b) \approx 0$. Save each intermediate result Z_b'.
If after n cycles $\mathrm{Cor}(Z_b) \neq 0$ return $Z_b'' = \min_{\mathrm{Cor}(Z_b')} Z_b'$.

4.2) Sort $(X_{\ell l}^*)_b \mapsto U_b$ according to $\mathrm{sign}(\bar{r}_{x_j x_\ell}) = \mathrm{sign}(\Re_b)$, under constraint $\|\bar{r}_{x_j x_\ell}\| \leq \|\Re_b\|$

4.3) Compute extremal correlation $\Re_b = \mathrm{Cor}(U_b)$.

4.4) Compute $n \times 1$ vector $I(\alpha_b)$ of n i.i.d. items drawn $\sim B(\alpha_b)$, $\alpha_b = 1 - (\bar{r}_{x_j x_\ell} / \Re_b)$.

4.5) Compute $(X_{-,j\ell}^*)_b = [I(\alpha_b) \circ Z_b] + [1 - I(\alpha_b) \circ U_b]$.

Sort $n \times 2$ matrix $(X_{-,j\ell}^*)_b$ by reference $(X_{-,\ell}^*)_b$. Return $(X_{-,j}^*)_b$ only.

Repeat step 4.1 to 4.5 for all $j = 2, \ldots, p$ columns. Then merge to obtain $n \times p$ realization

$$(X_-^*)_b = (U_\ell, X_{-,2}^*, \ldots, X_{-,j}^*, \ldots, X_{-,p}^*)_b.$$

Due to random fluctuations constraints in point 4.2 need not being always satisfied. Then $\alpha_b > 1$, $I(\alpha_b)$ cannot be computed, and vector $(X_{-,j}^*)_b$ is void. Here letting $\bar{R}_x$ be rank-based aims at mitigating this aspect. Additionally we could replace the faulty $\Re_b$, $\forall h$, with an expectation estimate $\bar{\Re} = \frac{1}{B} \sum_b^B \Re_b$ (median or maximum could do it as well). While such replacement seems reasonable we practically observe that entirely discarding realizations with a missing column, or similarly setting $(X_{-,j}^*)_b = 0$ if $\alpha_b > 1$, is the least biasing approach and we employ the latter. In order to reduce computation run-time step 4.1 of Algorithm A.2.1 is programmed in C++ and then imported into an R function that implements the rest of the Algorithm.

A.2.3 Rank correlation index

In order to insure the most stability in step 4.3) of Algorithm A.2.1 we employ a rank-correlation index instead of a moment-based one. Our rank index is similar to Spearman's. However to stress the influence of potential ties our ranking function uses the minimum ranking value between ties contrary to classic Spearman that uses the average.

Appendix B

Method details: IPD inference reconstruction

We give further details about Section 2.3, page 40.

B.1 Maximization objective functions

We consider log-likelihoods of Section 2.3.1, page 40. We give expressions for the log-likelihood (2.32) that uses a p.s.s. term (see Section 2.3.2) under different plug-in options (Section 2.3.3, page 41). These expressions are written in `C++` to avoid excessive computation time during optimization. We perform all likelihood maximization using package `maxLik` (see Appendix D.1, page 143), and obtain MLE $\hat{\beta}$. We tested the expressions output for consistency with standard modules. In particular MLE output was compared with that of `glm` or `coxph`. Optimization of our expressions did not yield materially different outputs from that of the standard modules (results not shown).

B.1.1 Log-likelihood expressions

Let data $X = (t, y, Z)$ include a $n \times 1$ time variable, $t \in \mathbb{R}^+$, a binary or continuous outcome, $y_{n\times 1}$, and a $n \times (p-2)$ set of covariates Z. When clear from context Z is understood as $(1, Z)$ after inclusion of an intercept unity vector-column. Approximately Normal 95% CIs are computed as $\hat{\beta} \pm 1.96\hat{\sigma}$ where $\hat{\sigma}$ is the variance estimate for $\hat{\beta}$. If not otherwise specified we compute AIC as $2k - 2\ell(\hat{\beta})$ where k is the length of the MLE vector and $\ell(\hat{\beta})$ is the value of the log-likelihood at its maximum.

Linear model

Here we simply write $X = (y, Z)$ where y is a $n \times 1$ continuous outcome and the time variable is incorporated in the $n \times k$ covariate-set Z, $k = (p-1)$. Reduced data is $s = Z^\top y$ and $s' = y^\top y$. In the

linear case we directly use the OLS close form solution,

$$\hat{\beta} = \left(Z^\top Z\right)^{-1} s. \tag{B.1}$$

The error term is

$$\hat{e} = (s' - Z\hat{\beta})^\top (s' - Z\hat{\beta}) \tag{B.2}$$

An estimate for the covariance matrix is

$$\Sigma = \hat{e}\left(Z^\top Z\right)^{-1}. \tag{B.3}$$

The variance is $\hat{\sigma}^2 = \text{diag}(\Sigma)$. An estimate for the log-likelihood maximum is

$$\ell_{\max} = \frac{\hat{e}}{2\sqrt{\hat{e}/(n-k)}}. \tag{B.4}$$

An AIC estimate is $2k - 2\ell_{\max}$.

Poisson model

Here outcome y is an event indicator taking value 0 (no event or censored) or 1 (event). Covariate set Z is modeled alongside time offset-variable t. The used log-likelihood in compact form is

$$\ell(\beta) = \beta^\top s - 1^\top p(\beta, Z), \tag{B.5}$$

where $s = Z^\top y$ and $p(\beta, Z) = \exp(Z\beta + \log(t))$ where $1^\top$ denotes the $1 \times n$ unity vector and t the $n \times 1$ time vector. The $\exp(\cdot)$ operator is applied element-wise. Gradient and Hessian of (B.5) are given in Section B.1.2.

Logistic model

Here y is a binary outcome and the time variable is incorporated in the covariate-set Z. The used log-likelihood has the same form of (B.5) where $p(\beta, Z) = \log(1 + \exp(Z\beta))$, and $\exp(\cdot)$ is applied element-wise. Gradient and Hessian are given in Section B.1.2.

Cox model

Here, outcome y is an event indicator taking value 0 (no event, or censored), or 1 (event). Denote with $X(t) = \text{sort}_t(X)$ the data X sorted by ascending order of time variable t, where $\text{sort}_a(v)$ is the function sorting v by ascending order of a. If $a = v$ we simply write $\text{sort}(v)$ to denote sorting by ascending order of v. Accordingly denote with $y(t)$ and $Z(t)$ the event-indicator and covariate-set sorted by time. Denote the sub-set of times where an event occurs with $t'_1, \ldots, t'_j, \ldots, t'_T$ where

$T \leq n$. Let β be a $k \times 1$ parameter vector. To reduce computation time we use an approximate definition for the risk set. Let the extended, or sub-distribution, risk set be $n \times 1$ vector

$$\mathscr{R}_\beta(t) = \text{sort}\left(\sum_{t_i<t} \text{rev}\left(\exp(Z(t_i)\beta)\right)\right), \tag{B.6}$$

where $\sum_{t_i<t}$ denotes cumulative summation over censored and uncensored event-times, $Z(t_i)\beta$ is a dot product, and rev$(\cdot)$ denotes the inverse of sort$(\cdot)$ function. Function exp$(\cdot)$ is applied element-wise. Next, compute the canonical risk set by discarding the subset of elements with no event,

$$\mathscr{R}_\beta(t') \approx \left\{\mathscr{R}_\beta(t_i) : y(t_i) = 1, \quad i = 1, \ldots, n\right\} \tag{B.7}$$

where $\mathscr{R}_\beta(t')$ is a $T \times 1$ vector and $\mathscr{R}_\beta(t_i)$ is the i-th element of (B.6). The approximation in (B.7) is typically very good and computationally faster. The Cox log-likelihood denominator is

$$\mathscr{F}(Z, \beta) = \sum_{j}^{T} \mathscr{R}_\beta(t'_j), \tag{B.8}$$

where $\mathscr{R}_\beta(t'_j)$ is the j-th element of (B.7).

If a stratification variable is modeled (B.8) is computed for each stratum. Then a stratified denominator is obtained after summation over the stratum-specific denominators. This may be typically useful in meta-analysis when each stratum denotes an independent data source and a stratum models one source-specific baseline cumulative hazard. The Cox partial log-likelihood is

$$\ell(\beta) = \beta^\top s - \mathscr{F}(Z, \beta), \tag{B.9}$$

where $s = Z^\top y$, and $\mathscr{F}(Z, \beta)$ is as in (B.8). Gradient and Hessian of (B.9) are given in Section B.1.2, below.

B.1.2 Linearized gradients and Hessians

To speed up computations we employ a linearized version of the gradients and Hessians that is approximately equal to the original respective expressions.

Poisson model

The gradient of (B.5) is computed as the $1 \times k$ vector

$$\nabla\ell(\beta) = s^\top - g, \tag{B.10}$$

where $g = 1^\top A$ with $A = Z \circ p(\beta, Z)$ and $\circ$ denotes column-wise Schur (element-wise) multiplication, with $p(\beta, Z) = \exp(Z\beta + \log(t))$ as above. An approximate Hessian is given by the linearized operation

$$\nabla^2\ell(\beta) \approx (-1)^\top B, \tag{B.11}$$

where $B = A^\top Z$ and A is defined as above. As usual (Efron and Hinkley, 1978) minus (B.11) is used as working Fisher Information[1] for normal approximations of the MLE.

Logistic model

The gradient here is computed as in (B.10), with

$$p(\beta, Z) = \frac{\exp(Z\beta)}{1 + \exp(Z\beta)}.$$

An approximate minus log-likelihood Hessian is computed as in (B.11) where

$$A = Z \circ p(\beta, Z) \circ 1 - p(\beta, Z),$$

and $p(\beta, Z)$ is defined exactly as above.

Cox model

Consider the $n \times k$ matrix

$$A = \text{sort}\left(\sum_{t_i<t} \text{rev}\left(\exp(Z(t_i) \circ Z(t_i)\beta)\right)\right), \tag{B.12}$$

where $\circ$ here denotes column-wise Schur (element-wise) multiplication. Let $A' = A \circ (1/B)$ where multiplication is column-wise and $B = \mathscr{R}_\beta(t)$ denotes the $n \times 1$ vector of extended at-risk counts as in (B.6). Let the $T \times k$ weight matrix of event-labeled rows be

$$C = \{A'(t_i) : y(t_i) = 1, \quad i = 1, \ldots, n\}, \tag{B.13}$$

where $A'(t_i)$ is the i-th row of (B.12) labeled by the event indicator $y(t_i) = \{0, 1\}$. Let $C' = 1^\top C$ be a $1 \times k$ weights vector where $1^\top$ is a $1 \times T$ unity vector. The gradient of (B.9) is computed as the $1 \times k$ vector

$$\nabla \ell(\beta) \approx s^\top - C'. \tag{B.14}$$

With a stratification variable all computations up to (B.13), included marginal summation afterward, are repeated for each stratum. An overall gradient is given by element-wise summation of stratum-specific gradients.

An approximate Hessian is computed as follows. Denote with A the $n \times k$ matrix with row

$$A(t_i) = \text{sort}\left(\sum_{t_i<t} \text{rev}\left(w(t_i)\right)\right), \tag{B.15}$$

[1] With an abuse of notation we often refer to the diagonal of the inverse of (B.11) as simply Fisher Information. For $k > 1$ this approach clearly simplifies the appraisal of the whole information by only restricting to the diagonal (see Section 2.3.1, page 40).

and

$$w(t_i) = cz(t_i)^\top z(t_i),$$

where $c = \exp\left[z(t_i)\beta\right]$ and $z(t_i)$ is the $1 \times k$ row-vector of time-ordered covariate matrix $Z(t)$. Let $A' = A \circ (1/B)$ as above. From A' compute the matrix C of only event-labeled rows, as in (B.13), and marginally sum up to get the $1 \times k$ vector C'. Next row-stack C' onto itself k times to get the $k \times k$ matrix C''. Define the $n \times k$ matrix

$$B' = \text{sort}\left(\sum_{t_i<t} \text{rev}\left(Z(t) \circ \exp(Z(t)\beta)\right)\right), \tag{B.16}$$

Similar to above let B'' be the shrunk $T \times k$ matrix after keeping only the event-labeled rows of B' Define the $n \times k$ matrix $D = B' \circ (1/B^2)$ where B is defined as above. Define the shrunk $T \times k$ matrix D' after keeping only the event-labeled rows of D. Define the $k \times k$ matrix $D'' = (D')^\top B''$. An approximate Hessian is

$$\nabla^2 \ell(\beta) \approx (-1)\left[C'' - D''\right]. \tag{B.17}$$

With a stratification variable computations up to (B.17) are repeated for each stratum. An overall Hessian is given by element-wise summation of stratum-specific Hessians.

B.2 Post-simulation trimming of outliers

Likelihood maximization on a simulated IPD may be occasionally unstable, due to random variation or bias in the IPD realization. The resulting MLE might be singular or an outlier and needs to be discarded. To do so we assess the range of the generated MLE distribution in the following. Compute the median of the generated MLE distribution and the median absolute deviation (m.a.d.). The latter is defined as the median of the difference between each sample point and the median, in absolute value. If the maximum (minimum) distribution value exceeds the median ± m.a.d. times 3.5, then keep only the distribution elements ranging between the median ± m.a.d. times 2.5. The latter is a conservative choice to avoid unrepresentative heavy tails values. We do not discard outliers during cumulative hazard estimation.

B.3 Expressions for the cumulative hazard estimate

We use a general Breslow expression to compute a cumulative hazard estimate, and derive a Nelson-Aalen estimate as special case. The cumulative risk set is

$$\sum_{t'_j<t} \frac{1}{\mathscr{R}_{\hat{\beta}}(t'_j)}, \tag{B.18}$$

where $\mathscr{R}_{\hat{\beta}}(t'_j)$ is as (B.7) except replacement of β with its MLE $\hat{\beta}$. Let the Breslow nominator be

$$\exp(\hat{\beta}^\top \tilde{z}), \tag{B.19}$$

where $\tilde{z}$ is a $(p-2)\times 1$ vector of predictive covariate values. The Breslow estimate is given by

$$A(t'|\hat{\beta},\tilde{z}) = \sum_{t'_j<t} \frac{\exp(\hat{\beta}^\top \tilde{z})}{\mathscr{R}_{\hat{\beta}}(t'_j)}. \tag{B.20}$$

If $\tilde{z}$ contains a binary treatment/control covariate we stratify computations accordingly. A Nelson-Aalen estimator is derived if we have a single and binary covariate g as

$$A(t'_g) = \sum_{t'_{gj}<t} \frac{1}{\mathscr{R}(t'_{gj})}, \quad g = 0, 1.$$

Because these c.h.e. expressions are based on a slightly different risk-set computation than usual, we checked their output for consistency with those of standard reference tools like `survfit`. Our expression did not lead to major differences in output compared to `survfit` (results not shown).

B.4 Computation of an "expected" cumulative hazard curve

In Section 2.3.6, page 42 we consider an expected cumulative hazard estimate, as the average across all c.h.e. simulations. To this end denote with A^*_b and t^*_b the marginal y and x axis of $A^*_b(t)$, that is the cumulative event-count and event-time $T\times 1$ vectors $\forall b = 1,\dots B$. Denote with A^*_{ib} and t^*_{ib} the element of A^*_b and t^*_b respectively for $i = 1,\dots T$ event-time points. The time-events count T should vary from sample to sample but for simplicity let keep it fixed for the moment.

To obtain a single overall graph we separately average each axis of the c.h.e. simulation to obtain $\bar{A}^*(\bar{t}^*)$, where

$$\bar{A}^* = \left(\bar{A}^*_1,\dots,\bar{A}^*_i,\dots,\bar{A}^*_T\right), \tag{B.21}$$

is the overall y-axis with

$$\bar{A}^*_i = \frac{1}{B}\sum_b^B A^*_{ib}, \quad \forall i = 1,\dots,T, \tag{B.22}$$

being the average cumulative events-count value at event-time point i, and where

$$\bar{t}^* = \left(\bar{t}^*_1,\dots,\bar{t}^*_i,\dots,\bar{t}^*_T\right), \tag{B.23}$$

is the overall x-axis with

$$\bar{t}^*_i = \frac{1}{B}\sum_b^B t^*_{ib}, \quad \forall i = 1,\dots,T, \tag{B.24}$$

being the average event-time value at point i.

We now account for different time-event lengths between simulations. We consider $T_{\max} = \max_T(T_1,\dots,T_b,\dots,T_B)$ where T_b is the number of time-events in realization $A^*_b(t)$. If the c.h.e. is Nelson-Aalen we intend $T_b = T_{0b}\wedge T_{1b}$ and $T_{\max} = T_{0,\max}\wedge T_{1,\max}$. Let T^*, accordingly $T^* = T^*_0\wedge T^*_1$ if c.h.e. is Nelson-Aalen, be the reference IPD count of time-events.

Algorithm B.4.1. (Average c.h.e. $\bar{A}^*(\bar{t}^*)$)

1) For all $b = 1, \dots, B$, increment A_b^* by $T_{\max} - T_b$ points.

   Accordingly, increment t_b^* by $T_{\max} - T_b$ points.

2) Compute average marginal dimension according to (B.21)-(B.22) and (B.23)-(B.24), page 94 (where $T = T_{\max}$).

3) For each marginal dimension, discard all points with index greater greater than T^*, $T^* \leq T_{\max}$.

   If T^* is not available, use $\bar{T} = \frac{1}{B}\sum_b T_b$.

Denote with

$$\Delta t_b^* = (t_2^* - t_1^*, \dots, t_i^* - t_{i-1}^*, \dots, t_{T_b}^* - t_{T_b-1}^*)_b \tag{B.25}$$

the vector of time jumps for simulation $b = 1, \dots, B$. Denote with $\bar{dt}_b^*$ the average of (B.25). Point 1) of Algorithm B.4.1 is implemented as follows.

Algorithm B.4.2. (Points augmentation)

1) Extend y-axis by repeating its last time-point value, $A_{T_b b}^*$, $T_{\max} - T_b$ times.

2) Extend x-axis by evenly sequencing $T_{\max} - T_b$ points between $t_{T_b}^* + \bar{dt}_b^*$ and $t_{T_b}^* + \left[(T_{\max} - T_b)\,\bar{dt}_b^*\right]$.

Repeat for all $b = 1, \dots, B$ under the approximation of constant average time-jump increment.

B.5 One covariate: sub-optimal reconstructions

We introduce reconstruction alternatives especially useful when exact information on data correlation is missing.

B.5.1 Survival models: incomplete at-risk denominators

Imagine we want to implement (1.2) given T large and baseline counts (n_{11}, n_{01}), that is the total number at risk in each arm at study start $t = 1$. We assume numerator s is given and time events can be right-censored. Because the exact count of at-risk units for $t > 1$ is missing we need a procedure to recover them. The probability that an event occurs in arm $x = 1$, at any time, is $p_1 = s/T$. We run $T - 1$ independent Bernoulli experiment with parameter p_1 that is $y_t \sim Bern(p_1)$, for $t = 2, \dots, T$. Denote with $Y_t = \sum_{t<T} y_t$ the simulated cumulative sum of at-risk units at time index $t = 2, \dots, T$. The simulated counts of at-risk units in arm 1 is $n_1^* = (n_{11}, n_{11} - Y_2, \dots, n_{11} - Y_t, \dots, n_{11} - Y_T)$. To simulate risk-sets in $x = 0$ we take $1 - y_t$, $t = 2, \dots, T$. Denote with $Y_t^- = \sum_{t<T}(1 - y_t)$ the cumulative sum of at-risk units in arm 0. The simulated counts of at-risk units in arm 0 is

$n_0^* = (n_{01}, n_{01} - Y_2^-, \ldots, n_{01} - Y_t^-, \ldots, n_{01} - Y_T^-)$. For both n_1^* and n_0^* we set to zero possible negative values at the end of the respective series. This method would only work when the covariate x is categorical. For more than two categories the above procedure can be generalized through a Multinomial sampling scheme.

The described simulation method does not account for right censoring. That is all censored units never leave the risk-set (1.2) that defines to a Fine-Gray sub-distribution log HR with all censoring events coded as a competing event. Hence we need to incorporate censoring information in the above procedure. A different approach is to now assume all censored units leave the risk set in each arm. Then we may replace p_1 with $\bar{p}_1 = (s/T) + [(n_{11} - s)/(n - T)]$, that is p_1 plus the probability that one censoring occurs in arm 1 at any time; $n = n_{11} + n_{01}$. Here we sample $y_t \sim Binom(\nu, \bar{p}_1)$ where $\nu = 1 + [(n_{11} - s)/T]$ is an adjusted size parameter that is an unit added to an average count of censoring events, which we always round to be integer-valued. To define the risk-sets in group $x = 0$ we use the term $(\nu - y_t)$, $t = 2, \ldots, T$. Cumulative at-risk sums are computed as above. This describes a model in which the censoring process is exhaustive at a sort of constant rate. We could find a compromising approach by averaging between leave-all-in and leave-all-out outputs. Alternatively we could average directly on the two types of risk-sets and proceed with a single estimation step. The single advantage of reproducing risk-sets in this way is that all input summaries are, typically, more easily recovered since no correlation is needed. The proposed risk sets can also be used to compute an approximation of the Nelson-Aalen estimate.

B.5.2 Knowledge of likelihood numerator

Imagine IPD consists of one outcome and only one continuous covariate. The type of application here may be GLM. If a p.s.s. from reference IPD is available (see point 3. from Section 2.3.3, page 41) to be plugged in the numerator of (2.32) then the dependence structure between outcome and covariate does not need to be known. Here IPD reconstruction needs only execution up to step 2) of Algorithm 2.1.2, page 25.

B.6 Data examples

B.6.1 Data list

We give the list IPD examples we use in our experiments. Data from IST Genova is kindly allowed for methodological purposes only and it is otherwise protected. A number of IPD examples are taken as real data excerpts from well known `R` packages.

B.6.2 Data re-organization

IPD of Table B.1 is re-organized in different batches as explained in Section 2.4.3, page 46. To automatize this re-organization we adopt a convention on the order of IPD marginal variables. For each data-set we put the time variable as first column and the event outcome as second column. If

Data name	Source
`aml`	package `survival`
`mgus`	package `survival`
`mgus1`	package `survival`
`lung`	package `survival`
`rats`	package `survival`
`aidssi`	package `mstate`
`abortion`	package `etm`
`sir.cont`	package `etm`
`bc`	package `flexsurv`
`diabetes`	package `timereg`
`cirrhosis`	University of Oslo
`oncocard`	IST Genova
`prost`	Prostate cancer (Royston and Sauerbrei, 2008)
`gbsg`	GBSG breast cancer (Royston and Sauerbrei, 2008)
`glio`	Glioma (Royston and Sauerbrei, 2008)
`kidney`	Kidney cancer (Royston and Sauerbrei, 2008)
`myel`	Myeloma (Royston and Sauerbrei, 2008)
`pbc`	PBC (Royston and Sauerbrei, 2008)
`rott`	Rotterdam breast cancer (Royston and Sauerbrei, 2008)
`wh`	Whiteall1 (Royston and Sauerbrei, 2008)

Table B.1: List of employed data-sets with source origin. All examples from (Royston and Sauerbrei, 2008) can be found at "portal.uni-freiburg.de/imbi/Royston-Sauerbrei-book/index.htmldatasets". Data from University of Oslo can be found at "http://www.uio.no/studier/emner/matnat/math/STK4080/h12/r-trial-project.html". Data from IST Genova is kindly allowed by Dr P Bruzzi, but masked due to privacy issues.

any, we put a continuous variable as third column. Similarly, we put a treatment outcome as fourth column. All other variables, if any, have no specific order. Appropriate instructions are given to mark the third column as the outcome variable to use, if the model is a linear regression.

Appendix C

Further results

C.1 IPD reconstruction

We report further results on IPD reconstruction from Section 3.1, page 49.

C.1.1 Similarity by increasing data sample size

In Table C.1 we give results similar to Table 3.1 but stratified by IPD sample size. In batch III and IV similarity percentages under method 4 increase until stratum $(400, 1000]$ and decrease afterward. In batch IV the decrease is even stronger. Performance decline in batch III and IV is due to Johnson distribution failure to well recover third and fourth moments features of some continuous variables as shown in Table 3.1. In Table C.2 we see that the fraction of wrongly reproduced continuous variables seems to occur more frequently in data-sets with more than thousand records in batch III and IV. The issue is more evident in data batch IV where the similarity percentage for the third and fourth moments is only 31.43 and 28.57 for data-sets with more than thousand records, under method 2 and 4 respectively. In Table C.3 we count the total number of continuous variables in each data-set from batch III and IV and for each sample size stratum. Similarly we count all correct solutions for the unknown Johnson parameter vector. Code 0 means no error while a code greater than zero means issues during parameters optimization, that indicates the Johnson solution might not comply to all moments constraints. Hence we count the number of continuous variables having a faulty Johnson solution, stratified by sample size. For method 4 we also compute the percentage of IPD reconstructions with dissimilar third and fourth moments relative to original IPD value by sample size. Accordingly we compute the median difference between simulated and original third or fourth degree moment. The number of included continuous variables and faulty Johnson solutions increase from batch III to IV. In particular from batch III to IV the overall count of continuous variables increases from 49 to 156 in IPDs with more than thousand records. Similarly from batch III to IV the overall count of faulty Johnson solutions increases from 6 to 58 in IPDs with more than thousand records. The percentage of IPD reconstructions with dissimilar fourth moment relative to IPD reference jumps from 15.38 %, in batch III, to 71.43 %, in batch

Table C.1: Percentage of IPD reconstructions satisfying overall similarity conditions defined in Section 2.2.4 (page 38) relative to original IPD reference, for increasing IPD sample size. Batch: see Section 2.4.2 (page 44). For batch see Table 3.1.

		Overall similar (%)			
Sample size	Method	*Batch I.*	*Batch II.*	*Batch III.*	*Batch IV.*
< 100	1	0.00	18.92	3.03	0.00
	2	0.00	64.86	18.18	0.00
	3	0.00	21.62	9.09	0.00
	4	42.86	59.46	51.52	31.82
(100,400]	1	0.00	19.75	0.00	0.00
	2	9.52	77.78	6.85	0.00
	3	0.00	17.28	0.00	0.00
	4	100.00	76.54	49.32	36.36
(400,1000]	1	3.12	17.14	0.00	0.00
	2	25.00	75.71	19.15	2.08
	3	43.75	15.71	2.13	25.00
	4	87.50	75.71	70.21	77.08
> 1000	1	0.00	9.43	0.00	0.00
	2	37.25	84.91	31.11	2.86
	3	0.00	9.43	0.00	0.00
	4	90.20	75.47	57.78	24.29

IV, in the last sample size stratum. Interestingly half of the dissimilar simulated fourth moments in batch IV shows only a mild discrepancy value of 1.96 from its IPD reference in sample sizes greater than thousand.

Table C.2: IPD batch III and IV: percentage of IPD reconstructions satisfying similarity conditions defined in point 3 and 4 of Section 2.2.4 (page 38) – difference between simulated and IPD original third (fourth) moment – by sample size. Correlation similarity is also reported (see point 5 of Section 2.2.4).

		Similar (%)			
		Batch III.		*Batch IV.*	
Sample size	Method	*Moment 3rd and 4th*	*Correlation*	*Moment 3rd and 4th*	*Correlation*
< 100	1	9.09	27.27	0.00	4.55
	2	63.64	21.21	50.00	0.00
	3	9.09	78.79	0.00	59.09
	4	63.64	75.76	59.09	59.09
(100,400]	1	0.00	21.92	0.00	0.00
	2	52.05	20.55	40.00	0.00
	3	0.00	98.63	0.00	80.00
	4	56.16	91.78	49.09	74.55
(400,1000]	1	2.13	25.53	29.17	2.08
	2	74.47	27.66	83.33	2.08
	3	2.13	97.87	29.17	91.67
	4	76.60	91.49	85.42	87.50
> 1000	1	0.00	51.11	0.00	8.57
	2	71.11	48.89	31.43	10.00
	3	0.00	88.89	0.00	82.86
	4	71.11	82.22	28.57	84.29

Table C.3: Simulation method 4, IPD batch III and IV: total count of included IPD continuous variables and of faulty solutions for the Johnson parameter vector by IPD sample size. In column 'Not similar' we report the percentage of differences between the simulated and IPD original third and fourth moment exceeding the threshold according to point 3 and 4 of Section 2.2.4 (page 2.2.4) respectively. Median bias: median of the mean absolute value differences.

		Total count of		Moment 3rd		Moment 4th	
Data batch	Sample size	*Variables*	*Fault codes*	*Not similar (%)*	*Median bias*	*Not similar (%)*	*Median bias*
III	< 100	35	2	5.56	1.14	50.00	4.07
	(100,400]	62	8	12.50	-2.12	18.75	-16.47
	(400,1000]	73	9	5.41	-5.08	24.32	-3.11
	> 1000	49	6	0.00		15.38	1.55
IV	< 100	58	2	0.00		40.91	3.16
	(100,400]	98	20	48.72	-2.14	48.72	-22.65
	(400,1000]	154	21	3.12	-5.27	25.00	-1.46
	> 1000	156	58	0.00		71.43	1.96

C.2 IPD inference reconstruction

We report results on IPD inference simulation from Section 3.2, page 53.

C.2.1 Inferences reconstruction in data batch I

This data batch is explicitly designed to perform Cox regression with one binary/categorical covariate, and respective stratified Nelson-Aalen estimation. In all but one case the covariate is a binary treatment/control indicator. Results are presented by simulation method and IPD simulation similarity to its reference. For additional results stratified by estimate's rank, or IPD bootstrap type see Appendix C.2.5, page 118 (Table C.15, page 118, and Table C.22, page 125). Additional results for the x-axis, the event-time line, of the Nelson Aalen graph can be found in Appendix C.2.6.

Table C.4, page 107, shows results on the average HR and r.f.i.d. simulation. In all but one case the log-likelihood Hessian is scalar and the r.f.i.d. is one-to-one with the Fisher Information. Results from data `wh.4` are discarded since bias there is clearly outlying (see Section 3.4.1, page 62). Mean and dispersion of the difference between simulated and original HR are always acceptable relative to mean and s.d. of original IPD HR, quite regardless of method and IPD similarity to reference. The difference dispersion seems however lower if IPD is similar to original reference or under NORTAmax resampling (methods 1-3 to 3-4). Difference dispersion is also low if the original IPD Cox likelihood numerator is known (methods 3-1, 3-2, 3-3, 3-4). Here the benefit seems stronger if IPD is dissimilar to reference. The mean difference between simulated and original r.f.i.d. is generally tightly close to zero quite regardless of simulation method and IPD similarity to reference. Overall performance under ordinary NORTAmax resampling (method 1-3 against 1-4) is rather comparable.

Table C.5, page 108, shows results on empirical 95% quantiles of the HR simulation versus any of the normal, basic, Bca, and percent original IPD bootstrap CIs. Additional to results from data `wh.4` also results from data `rats.2` are discarded (see Section 3.4.1, page 62, and Appendix C.3.5, page 138). Overall difference between reconstructed and IPD quantiles is comparable with that observed for the HR in Table C.4. The difference is tighter around zero under ordinary NORTAmax resampling (method 1-3 and 1-4) quite irrespective of moment degree and of IPD similarity to reference. Under non NORTAmax resampling, knowledge of the IPD Cox likelihood nominator or its average estimate (method 3-1, 3-2, 2-3 and 3-3) helps mitigating bias.

Table C.6, page 109, shows results on the cumulative events-count (y-axis) of the average Nelson-Aalen simulation stratified by one treatment covariate. We consider y-axis quartile values, or an aggregate index, also including y-axis range and mean value. Focus is on the difference between the simulated and IPD original y-axis quartile or its aggregate index. The average quartile difference is remarkably close to zero almost everywhere. Column 'Biased' reports two types of percentages. On the left it reports the percentage of the aggregate index differences exceeding ± 0.1. On the right it reports the same percentage after excluding the difference at the y-axis maximum, that by the cadlag property of the Nelson-Aalen estimate corresponds to the difference at the last event-time point. These percentages roughly decrease from method 1 to 4 and by excluding values

at the last event-time point it reduces the percentage of differences greater than ±0.1 even further, indicating that most bias accumulates toward the end of the event-time line. Table C.26, page C.26, reports the difference between the simulated and IPD original event-time line (log scale). In general mean and s.d. difference are lower under NORTAmax resampling (methods 3 and 4) in both treatment groups and quite regardless of IPD similarity to reference.

C.2.2 Inferences reconstruction in data batch II

This data batch is explicitly designed to perform GLM regression of Gaussian, Poisson and Binomial family with one continuous covariate. The batch is divided into IPDs with continuous or binary outcome, depending if regression is Gaussian or not respectively. We consider the average MLE, and r.f.i.d. vector simulation relative to original IPD value. Here intercept and slope estimates are pooled together. We consider 95% empirical quantiles of the MLE simulation, relative to classic IPD bootstrap intervals. Results are presented by simulation method and IPD similarity to its reference. For additional results stratified by estimate rank, model, or IPD bootstrap type see Appendix C.2.5, page 118 (Table C.16, page 119, Table C.19, 122, and Table C.23, page 126).

Table C.7, page 110, shows results on the average MLE and r.f.i.d. vector simulation. Few points of the difference between simulated and original r.f.i.d. (1% of the total) exceeding ±100 are discarded. The simulated r.f.i.d. may explode in an isolated run due to random log-likelihood fluctuations especially if the original IPD r.f.i.d. estimate is also unstable (for instance see occasional explosion of original IPD r.f.i.d. dispersion in the top right half of the table). Typically such outliers should be automatically trimmed away but few unrepresentative points may remain in the sample. Most of the discarded points are linked to intercept estimation in the Binomial or Gaussian regression (result not shown). The remaining reported values exceeding ±1000 are set to Infinity (Inf) for conciseness. Overall difference between simulated and original MLE is typically lower if IPD is similar to reference. Here the difference mean and dispersion are always acceptable relative to mean and s.d. of original IPD MLE under NORTAmax resampling (methods 1-3 to 3-4). Ordinary NORTAmax resampling (method 1-3, or 1-4) performs quite comparably irrespective of marginal moment maximal degree and only roughly irrespective of IPD similarity to reference. Although non NORTAmax IPD re-ordering performs very well – see method 1 and 2 in Table 3.1, page 50 – we cannot observe a corresponding good IPD inferential reconstruction as shown by typically high difference dispersion's under method 1-1 and 2-2 quite irrespective of IPD similarity to reference. This may actually indicate that Algorithm A.2.1 introduces unnecessary distorting noise even if correlation is scalar. Here knowledge of the original IPD likelihood numerator (method 3-2) seems to mitigate bias if IPD seems at least similar to reference. Similar patterns are observed for the r.f.i.d..

Table C.8, page 111, shows results on empirical 95% quantiles of the HR simulation versus any of the normal, basic, Bca, and percent original IPD bootstrap CIs. Analogously to above we discard few instances of bias exceeding ±100 (1% of the total sample). Generally the overall difference between reconstructed and IPD quantiles seem quite large except under NORTAmax resampling (method 1-3 and 1-4) where the quantiles difference mean and dispersion is acceptable, relative to

mean and s.d. of the original IPD value. This outcome roughly holds irrespective of IPD similarity to reference.

C.2.3 Inferences reconstruction in data batch III

This data batch is explicitly designed to perform GLM and Cox regression with two continuous covariates. Used GLM families are the Gaussian, Poisson and Binomial. Breslow estimation is also performed. This batch is divided into IPDs with continuous or binary outcome, depending if the performed regression is or it is not Gaussian. We model a log time offset in the GLM Poisson regression which amounts to perform a Cox regression under a parametric assumption of constant baseline hazards. For additional results stratified by estimate rank, model, or reference bootstrap type see Appendix C.2.5, page 118 (Table C.17, page 120, Table C.20, page 123, and Table C.24, page 127). Additional results for the x-axis, the time line, of the average Breslow simulation can be found in the Appendix C.2.6.

Table C.9, page 112, shows results on the average MLE and r.f.i.d. vector simulation. We discard difference points exceeding ±100 for a total of 1% of the entire sample. The overall difference between simulated and original MLE is comparably lower under ordinary NORTAmax resampling (method 1-3 and 1-4) quite regardless of moment degree and of IPD similarity to reference. Here the difference mean and dispersion are always acceptable relative to mean and s.d. of original IPD MLE. The reconstruction performance is also generally better if IPD is similar to reference. A similar pattern is observed for the r.f.i.d..

Table C.10, page 113, shows results on empirical 95% quantiles of the HR simulation versus any of the normal, basic, Bca, and percent original IPD bootstrap CIs. We discard bias exceeding ±100 (2% of the total sample). Generally the overall difference between reconstructed and IPD quantiles seem quite large except under NORTAmax resampling (method 1-3 and 1-4) where the quantiles difference mean and dispersion is acceptable, relative to mean and s.d. of the original IPD value. This outcome roughly holds irrespective of IPD similarity to reference. If IPD is not similar to reference non ordinary NORTAmax resampling (method 1-1 and 1-2) perform also relatively well.

Table C.11, page 114, shows results on the cumulative events-count (y-axis) of the average Breslow simulation for a single continuous covariate. Here we have fewer IPD examples available for Cox and Breslow modelling. We consider y-axis quartile values or an aggregate index also including y-axis range and mean value. Focus is on the difference between the simulated and IPD original y-axis quartile or its aggregate index. The average 1st or 2nd quartile difference is remarkably close to zero almost everywhere. Several approaches yield unacceptable mean bias at the third quartile if IPD is not similar to reference. As seen in Table C.6, page 109, bias generally accumulates toward the end of the time-line (first column of 'Biased'). Excluding bias at the last time point, bias remains always below ±0.1 (second column of 'Biased') under most methods especially if IPD is similar to its reference. Overall difference seems slightly lower under ordinary NORTAmax sampling (method 1-3 and 1-4), irrespective of IPD similarity reference. Table C.27, page 130, reports the difference between the simulated and IPD original event-time line (log scale).

Mean and s.d. difference are generally lower under NORTAmax sampling (methods 1-3 to 3-4) regardless of IPD similarity to reference.

C.2.4 Inferences reconstruction in data batch IV

This data batch is designed to perform the same regression models of batch III but using mixed binary and continuous covariates instead. Overall we model at most three covariates. For additional results stratified by estimate rank, model, or reference bootstrap type see Appendix C.2.5, page 118 (Table C.18, page 121, Table C.21, page 124, Table C.25, page 128, and Table C.28, page 131). Additional results for the x-axis, the event-time line, of the Breslow graph can be found in Appendix C.2.6. Table C.12, page 115, shows results on the average MLE and r.f.i.d. vector simulation. Difference points exceeding ±100 are discarded for a total of 2% of the entire sample. Similar to batch III the overall difference between simulated and original MLE is here comparably lower under ordinary NORTAmax resampling (method 1-3 and 1-4), quite regardless of moment degree and of IPD similarity to reference. Difference mean and dispersion are always acceptable relative to mean and s.d. of original IPD MLE. NORTAmax performance is particularly remarkable if IPD is dissimilar to reference, which comprises the majority of IPD examples here. NORTAmax IPD reconstruction was relatively worst mainly due to Johnson marginal inaccuracies at fourth moment reconstruction (see Table 3.1, page 50, or Table C.2 and C.3, pages 101 and 102). However those fourth moment inaccuracies seem to not excessively affect IPD inference reconstruction which remains good here. Similar bias patterns are observed for the r.f.i.d..

Table C.13, page 116, shows results on simulated empirical 95% interval percentiles for the MLE vector elements. Reference percentiles are any of the type normal, basic, percent, and Bca bootstrap CIs, computed on original IPD. We discard bias exceeding 100 in absolute value, which amounts to 2% of the total sample. Overall ordinary sampling (methods 1-1, 1-2, 1-3, and 1-4) display lower bias patterns for both percentiles quite regardless of IPD similarity to reference. Difference dispersion is yet lower under ordinary NORTAmax resampling (method 1-3 and 1-4), if IPD is similar to reference.

Table C.14, page 117, shows results on the cumulative events-count (y-axis) of the average Breslow simulation stratified by one treatment covariate. Similar to previous results we often see a remarkable performance with widespread near zero average bias at the first two quartiles of the y-axis. Method 1-1 to 3-1 yield unacceptable mean bias (over 1000 indicated with Inf) if IPD is not similar to reference, especially in treatment group. This is likely due to few exploding MLE cases under treatment. Exploding bias often occurs at the third quartile for some methods confirming the notion that errors accumulates toward the end of the event-time line. After excluding points at the last event-time point only few bias instances exceed ±0.1 under most methods and especially if IPD is similar to reference. In general, however, ordinary resampling approaches (methods 1-2, 1-3, 1-4, but 1-1 excluded) interestingly yield comparable and overall low bias, quite regardless of IPD similarity to reference. Table C.28, page 131, reports the difference between the simulated and IPD original event-time line (log scale). As seen in batch III mean and s.d. difference are generally lower under NORTAmax sampling (methods 1-3 to 3-4) regardless of IPD similarity to reference.

Table C.4: IPD batch I (110 data-sets). Results for the average HR (from Cox regression with one binary covariate) and reciprocal Fisher Information diagonal simulation. Mean and standard deviation (± s.d.) of original IPD estimate and of difference between simulated and original IPD inference (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

			Maximum Likelihood Estimate							1 / Fisher Information diagonal						
			Biased (%)			*IPD Value*		*Bias*		*Biased (%)*			*IPD Value*		*Bias*	
Similar data	*Method*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1		73	39	17	0.14	0.65	-0.00	0.45	1	1	0	0.03	0.06	-0.00	0.04
	2-1	99	74	39	17	0.14	0.65	-0.02	0.47	1	1	0	0.03	0.06	-0.00	0.04
	3-1		17	2	1	0.14	0.65	-0.03	0.12	0	0	0	0.03	0.06	-0.00	0.00
	1-2		88	52	22	0.12	0.69	-0.01	0.52	1	1	0	0.04	0.07	-0.01	0.05
	2-2	74	89	52	23	0.12	0.69	-0.03	0.54	1	1	0	0.04	0.07	-0.01	0.05
	3-2		21	2	2	0.12	0.69	-0.03	0.14	0	0	0	0.04	0.07	-0.00	0.00
	1-3		9	2	2	0.16	0.67	0.02	0.09	0	0	0	0.03	0.07	0.00	0.01
	2-3	87	10	2	2	0.16	0.67	0.02	0.09	0	0	0	0.03	0.07	0.00	0.00
	3-3		5	2	2	0.16	0.67	0.02	0.09	0	0	0	0.03	0.07	0.00	0.00
	1-4		23	0	0	-0.01	0.47	0.03	0.07	0	0	0	0.08	0.08	0.00	0.01
	2-4	12	23	0	0	-0.01	0.47	0.04	0.07	0	0	0	0.08	0.08	0.00	0.01
	3-4		15	0	0	-0.01	0.47	0.00	0.07	0	0	0	0.08	0.08	0.00	0.00
TRUE	1-1		0	0	0	0.47		0.04		0	0	0	0.02		0.00	
	2-1	1	0	0	0	0.47		0.04		0	0	0	0.02		0.00	
	3-1		0	0	0	0.47		0.07		0	0	0	0.02		0.00	
	1-2		31	0	0	0.23	0.49	0.03	0.10	0	0	0	0.01	0.01	-0.00	0.00
	2-2	26	31	0	0	0.23	0.49	0.03	0.10	0	0	0	0.01	0.01	-0.00	0.00
	3-2		10	0	0	0.23	0.49	-0.01	0.07	0	0	0	0.01	0.01	0.00	0.00
	1-3		0	0	0	0.02	0.49	0.01	0.04	0	0	0	0.02	0.01	0.00	0.00
	2-3	13	0	0	0	0.02	0.49	0.01	0.04	0	0	0	0.02	0.01	0.00	0.00
	3-3		0	0	0	0.02	0.49	0.01	0.04	0	0	0	0.02	0.01	0.00	0.00
	1-4		7	1	0	0.17	0.67	0.02	0.07	0	0	0	0.03	0.06	0.00	0.01
	2-4	88	8	1	0	0.17	0.67	0.02	0.07	0	0	0	0.03	0.06	0.00	0.00
	3-4		3	1	1	0.17	0.67	0.02	0.06	0	0	0	0.03	0.06	0.00	0.00

Table C.5: IPD batch I (109 data-sets). Results for 95% empirical CIs of the HR simulation. Mean and standard deviation (± s.d.) of original IPD CIs (any of the type percent, normal Bca and basic IPD bootstrap quantiles) and of difference between reconstructed and IPD quantile (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

			2.5th Percentile							97.5th Percentile						
			Biased (%)			*IPD Value*		*Bias*		*Biased (%)*			*IPD Value*		*Bias*	
Similar data	*Method*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1		73	38	17	-0.13	0.61	0.04	0.36	68	32	17	0.49	0.65	-0.10	0.36
	2-1	99	76	38	20	-0.13	0.62	-0.06	0.39	72	36	19	0.49	0.65	-0.01	0.42
	3-1		20	8	4	-0.13	0.61	-0.04	0.17	43	13	5	0.49	0.65	0.02	0.26
	1-2		83	50	23	-0.18	0.64	0.03	0.42	84	42	22	0.51	0.70	-0.14	0.41
	2-2	73	85	49	28	-0.18	0.64	-0.09	0.43	83	44	23	0.51	0.70	-0.02	0.48
	3-2		25	10	7	-0.18	0.64	-0.05	0.20	44	16	8	0.51	0.70	0.04	0.31
	1-3		15	5	4	-0.10	0.63	0.05	0.14	17	4	2	0.53	0.67	-0.00	0.13
	2-3	87	33	10	3	-0.10	0.63	0.02	0.20	40	10	6	0.53	0.67	0.06	0.24
	3-3		34	8	3	-0.10	0.62	0.02	0.19	36	11	6	0.53	0.67	0.05	0.24
	1-4		35	20	10	-0.55	0.57	0.12	0.28	22	10	4	0.55	0.69	-0.06	0.24
	2-4	12	55	31	22	-0.55	0.57	0.08	0.40	61	22	16	0.55	0.69	0.03	0.39
	3-4		63	29	20	-0.56	0.58	0.05	0.41	47	22	12	0.55	0.70	-0.02	0.38
TRUE	1-1		0	0	0	0.23	0.01	0.09	0.01	0	0	0	0.70	0.01	0.02	0.01
	2-1	1	0	0	0	0.22	0.01	-0.08	0.01	100	0	0	0.71	0.01	0.16	0.01
	3-1		0	0	0	0.22	0.01	-0.05	0.01	100	0	0	0.71	0.01	0.19	0.01
	1-2		32	0	0	0.04	0.49	0.06	0.09	35	0	0	0.45	0.49	-0.01	0.10
	2-2	27	36	3	2	0.04	0.49	0.04	0.14	41	0	0	0.45	0.49	0.02	0.10
	3-2		11	0	0	0.04	0.49	0.01	0.06	33	3	2	0.46	0.48	-0.02	0.14
	1-3		11	0	0	-0.27	0.49	0.02	0.06	5	0	0	0.31	0.47	-0.02	0.05
	2-3	13	32	0	0	-0.27	0.49	-0.08	0.07	48	5	0	0.30	0.47	0.11	0.10
	3-3		41	0	0	-0.27	0.49	-0.08	0.07	45	4	0	0.30	0.47	0.10	0.09
	1-4		10	2	1	-0.06	0.59	0.03	0.09	10	1	0	0.49	0.64	-0.00	0.07
	2-4	88	28	5	1	-0.06	0.59	0.00	0.13	32	8	3	0.49	0.64	0.07	0.17
	3-4		26	3	0	-0.06	0.59	0.00	0.12	35	8	3	0.49	0.64	0.07	0.20

Table C.6: IPD batch I (111 data-sets). Results on the marginal y-axis (cumulative events-count) of the average Nelson-Aalen simulation in control and treatment group. Average bias (difference) between 1st, 2nd, 3rd y-axis quartile and original IPD value. Biased: percentage of quartile differences, including that at y-axis range and average, exceeding ±0.1 (first column). Same computation after excluding difference at y-axis maximum (second column 2). Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Section 2.2.3, page 38. Size: number of IPD examples relative to total batch size (%).

				Control					Treatment				
						Mean bias quartile					*Mean bias quartile*		
Similar data	*Type*	*Method*	*Size (%)*	*Biased (%)*		*1st*	*2nd*	*3rd*	*Biased (%)*		*1st*	*2nd*	*3rd*
FALSE	Nelson-Aalen	1	99	17	6	-0.00	0.00	0.01	19	8	-0.00	-0.01	-0.01
		2	74	19	9	-0.00	0.00	0.02	22	11	-0.01	-0.01	-0.01
		3	87	12	1	0.00	0.00	0.01	13	2	0.00	0.00	0.01
		4	12	15	3	0.01	0.01	0.01	15	4	0.01	0.02	0.03
TRUE		1	1	39	27	-0.01	-0.01	0.09	17	0	0.00	0.00	0.01
		2	26	11	1	-0.00	-0.00	0.00	11	1	-0.00	-0.00	0.01
		3	13	15	2	-0.00	-0.01	-0.00	13	0	-0.01	-0.03	-0.04
		4	88	12	1	-0.00	-0.00	0.01	12	2	-0.00	-0.00	0.01

Table C.7: IPD batch II (241 data-sets). Results for the average MLE (from GLM and Cox regression with one continuous covariate) and reciprocal Fisher Information diagonal simulation. Mean and standard deviation (± s.d.) of original IPD estimate and of difference between simulated and original IPD inference (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

			Maximum Likelihood Estimate							1 / Fisher Information diagonal						
			Biased (%)			*IPD Value*		*Bias*		*Biased (%)*			*IPD Value*		*Bias*	
Similar data	*Method*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1		35	19	14	12.8	76.1	0.70	28.64	11	7	5	14.8	167.8	-0.05	4.96
	2-1	83	36	20	15	12.8	76.1	2.61	48.07	10	6	5	25.7	314.0	0.21	6.39
	3-1		26	13	10	12.8	76.1	Inf	Inf	10	5	4	134.6	Inf	-0.07	6.53
	1-2		39	27	22	9.2	41.0	-0.33	6.27	20	13	10	9.3	56.0	-0.13	2.68
	2-2	23	42	30	25	9.2	41.0	9.52	87.47	17	9	8	9.0	56.7	0.24	7.82
	3-2		28	16	13	9.2	41.0	10.30	98.95	15	9	7	26.9	236.8	0.45	5.66
	1-3		26	13	9	12.4	75.4	0.06	1.60	11	7	6	14.5	165.9	-0.09	4.49
	2-3	84	29	13	10	12.4	75.4	2.53	43.12	11	5	4	30.2	333.8	0.03	5.47
	3-3		26	13	9	12.4	75.4	5.04	67.81	9	5	4	132.1	Inf	-0.19	4.80
	1-4		29	19	12	7.7	37.6	0.33	5.18	18	12	12	7.1	51.0	0.13	2.72
	2-4	27	31	23	15	7.7	37.6	9.50	94.32	12	8	7	6.6	52.3	-0.04	1.37
	3-4		24	15	13	7.7	37.6	9.32	97.53	12	6	5	22.2	221.3	0.05	0.83
TRUE	1-1		34	12	4	0.8	7.7	-0.00	0.27	5	1	1	0.3	0.7	0.01	0.18
	2-1	17	32	12	3	0.8	7.7	0.02	0.26	6	2	1	0.3	0.7	-0.00	0.08
	3-1		14	4	1	0.8	7.7	0.02	0.20	10	3	1	0.3	0.7	0.01	0.10
	1-2		30	12	6	10.7	74.7	0.87	27.13	4	2	2	23.9	317.0	-0.13	1.87
	2-2	77	29	12	6	10.7	74.7	0.89	27.29	4	2	2	23.9	317.0	0.02	1.39
	3-2		12	5	3	10.7	74.7	0.00	0.35	4	2	1	23.9	317.0	-0.09	2.37
	1-3		17	5	1	1.5	9.8	-0.01	0.13	8	1	1	0.3	0.7	-0.00	0.08
	2-3	16	20	8	4	1.5	9.8	-0.02	0.18	9	1	1	0.3	0.7	-0.00	0.08
	3-3		20	6	2	1.5	9.8	-0.01	0.17	8	1	1	0.3	0.7	-0.00	0.08
	1-4		10	5	3	11.3	76.5	0.00	0.60	3	2	2	25.3	323.9	-0.32	5.66
	2-4	73	13	5	4	11.3	76.5	-0.03	0.81	5	2	2	25.3	323.9	-0.11	3.48
	3-4		11	5	3	11.3	76.5	-0.01	0.80	5	2	1	25.3	323.9	-0.20	3.11

Table C.8: IPD batch II (241 data-sets). Results for 95% empirical CIs of the HR simulation. Mean and standard deviation (± s.d.) of original IPD CIs (any of the type percent, normal Bca and basic IPD bootstrap quantiles) and of difference between reconstructed and IPD quantile (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

			2.5th Percentile							97.5th Percentile						
			Biased (%)			*IPD Value*		*Bias*		*Biased (%)*			*IPD Value*		*Bias*	
Similar data	*Method*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1		38	23	18	8.5	44.3	-0.32	6.32	38	23	17	12.2	58.2	-0.23	5.24
	2-1	83	47	32	27	7.3	30.4	-1.04	7.96	51	33	25	11.4	51.2	0.46	7.56
	3-1		49	33	26	9.1	49.6	-0.58	6.92	50	32	26	14.6	92.1	0.14	8.16
	1-2		46	33	28	6.2	36.0	0.34	7.70	44	30	22	12.0	47.7	-0.64	7.89
	2-2	23	60	42	36	6.7	36.3	-1.45	8.66	60	44	36	12.3	50.3	0.38	8.87
	3-2		57	39	35	7.1	36.7	-0.65	9.60	57	42	35	12.0	49.2	0.72	9.68
	1-3		32	16	12	9.7	58.1	0.08	3.50	31	17	13	15.4	96.8	-0.11	3.54
	2-3	84	49	32	26	10.0	58.5	-0.60	6.70	51	32	26	15.3	97.4	0.64	7.50
	3-3		49	32	26	8.7	48.7	-0.53	6.96	50	32	26	15.4	97.6	0.25	8.05
	1-4		38	24	19	5.6	33.0	0.06	3.11	40	25	20	10.6	45.6	-0.26	4.82
	2-4	27	57	40	35	6.0	33.7	-0.90	9.00	60	41	33	9.8	44.6	0.66	8.28
	3-4		56	37	32	6.0	33.7	-0.77	9.08	55	38	32	9.8	44.6	0.73	8.08
TRUE	1-1		51	28	16	0.2	7.7	0.25	1.03	48	27	16	1.6	8.1	-0.26	1.25
	2-1	17	58	33	23	0.2	7.6	-0.62	2.13	57	39	23	1.5	8.1	0.59	1.93
	3-1		54	35	25	0.2	7.7	-0.59	2.19	54	33	21	1.6	8.2	0.54	2.08
	1-2		36	19	12	7.0	41.1	-0.13	4.89	37	19	13	9.3	53.5	-0.09	3.46
	2-2	77	44	29	21	7.0	41.1	-0.82	6.32	45	28	20	9.3	53.4	0.58	5.54
	3-2		43	25	19	6.8	41.8	-0.30	4.99	43	27	20	13.1	96.4	0.38	6.37
	1-3		29	8	4	1.0	9.6	0.03	0.23	27	11	5	2.3	10.2	-0.04	0.25
	2-3	16	55	34	25	1.0	9.6	-0.78	2.69	56	31	22	2.3	10.2	0.63	1.98
	3-3		55	33	25	1.0	9.6	-0.76	2.62	54	31	21	2.3	10.2	0.65	2.10
	1-4		18	8	5	9.1	58.4	0.05	1.57	18	7	4	13.9	98.6	-0.04	2.10
	2-4	73	44	26	20	9.1	58.2	-0.46	6.42	45	26	20	11.1	79.2	0.44	5.83
	3-4		43	26	20	8.7	55.8	-0.40	6.61	44	26	20	12.0	89.2	0.41	5.69

Table C.9: IPD batch III (191 data-sets). Results for the average MLE (from GLM and Cox regression with two covariates) and reciprocal Fisher Information diagonal simulation. Mean and standard deviation (± s.d.) of original IPD estimate and of difference between simulated and original IPD inference (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

			Maximum Likelihood Estimate							1 / Fisher Information diagonal						
			Biased (%)			*IPD Value*		*Bias*		*Biased (%)*			*IPD Value*		*Bias*	
Similar data	*Method*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1		47	33	24	4.5	32.7	0.14	5.33	13	8	6	11.3	124.3	-0.24	4.02
	2-1	99	48	34	26	4.6	32.8	0.00	6.02	16	10	8	8.1	79.2	-0.52	4.88
	3-1		36	20	14	6.4	48.0	0.33	4.62	13	9	6	8.0	79.6	-0.15	4.28
	1-2		48	34	25	4.1	20.2	0.07	4.98	14	9	6	227.4	Inf	-0.19	5.27
	2-2	83	47	36	29	4.1	20.3	-0.05	5.90	16	11	8	71.4	Inf	-0.43	6.79
	3-2		30	18	13	7.5	52.8	0.80	7.45	12	8	6	13.6	139.0	0.03	3.98
	1-3		28	13	10	6.4	48.1	-0.05	1.84	11	7	5	12.6	128.0	-0.11	3.97
	2-3	98	33	16	11	6.4	48.3	-0.06	2.85	13	8	6	7.9	80.1	0.00	3.40
	3-3		32	17	12	6.4	48.3	0.10	3.35	13	8	6	11.2	125.6	0.02	3.71
	1-4		29	18	15	8.5	57.6	-0.08	3.37	17	11	8	12.4	103.2	0.01	4.94
	2-4	45	34	22	19	8.6	58.0	0.31	8.26	15	8	6	331.3	Inf	0.19	5.46
	3-4		28	16	14	8.6	58.0	0.35	9.23	13	7	5	132.0	Inf	-0.15	3.43
TRUE	1-1		0	0	0	0.5	0.9	-0.00	0.02	0	0	0	0.0	0.0	-0.00	0.00
	2-1	1	0	0	0	0.5	0.9	-0.02	0.03	0	0	0	0.0	0.0	-0.02	0.02
	3-1		33	0	0	0.5	0.9	-0.01	0.09	0	0	0	0.0	0.0	-0.02	0.02
	1-2		36	21	15	1.7	10.6	0.00	0.54	12	5	2	0.9	5.7	-0.06	0.47
	2-2	17	37	21	15	1.7	10.6	-0.08	0.63	14	8	6	0.9	5.7	-0.30	2.37
	3-2		24	16	14	1.7	10.6	-0.04	0.45	15	8	6	0.9	5.7	-0.28	2.47
	1-3		13	7	7	5.2	15.5	0.03	0.18	13	7	7	7.0	18.2	-0.09	0.28
	2-3	2	27	27	27	5.2	15.5	-0.13	0.40	27	13	13	7.0	18.2	-2.55	6.62
	3-3		47	27	27	5.2	15.5	-0.16	1.65	27	13	13	7.0	18.2	-2.67	6.92
	1-4		17	6	4	4.9	39.5	0.00	1.04	7	3	2	10.6	136.5	-0.08	1.36
	2-4	55	21	11	7	4.9	39.5	0.01	1.32	10	7	5	304.5	Inf	-0.01	2.96
	3-4		21	10	7	4.9	39.5	-0.02	1.04	11	7	5	10.6	136.9	-0.16	1.50

Table C.10: IPD batch III (191 data-sets). Results for 95% empirical CIs of the MLE simulation. Mean and standard deviation (± s.d.) of original IPD CIs (any of the type percent, normal Bca and basic IPD bootstrap quantiles) and of difference between reconstructed and IPD quantile (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

			2.5th Percentile							97.5th Percentile						
			Biased (%)			*IPD Value*		*Bias*		*Biased (%)*			*IPD Value*		*Bias*	
Similar data	*Method*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1		48	34	26	1.8	17.1	0.23	7.18	47	33	26	6.7	47.3	-0.07	6.50
	2-1	99	54	41	35	1.7	17.1	-1.33	9.89	53	40	33	5.3	22.8	1.52	11.38
	3-1		52	36	29	2.0	20.1	-1.99	10.01	52	37	30	7.3	51.6	2.20	10.30
	1-2		48	35	28	1.9	18.2	0.35	7.66	48	35	26	7.6	52.0	-0.01	7.59
	2-2	83	54	43	36	1.8	18.1	-0.93	9.22	53	42	34	5.9	24.5	1.16	10.93
	3-2		47	33	26	2.0	18.3	-1.67	8.63	48	35	28	7.6	51.7	2.02	9.78
	1-3		36	19	13	2.7	24.7	0.37	5.30	35	19	14	10.6	80.0	-0.41	5.27
	2-3	98	50	33	27	2.0	17.1	-1.49	8.84	50	34	26	6.9	51.5	1.29	7.08
	3-3		50	35	28	1.9	16.8	-1.31	8.00	50	33	27	6.9	51.4	1.33	7.22
	1-4		39	24	18	3.0	26.7	0.73	7.66	38	23	18	10.4	66.3	-0.77	6.52
	2-4	45	49	36	30	2.6	18.6	-1.94	9.86	52	37	31	10.4	75.6	1.89	8.58
	3-4		48	34	30	2.6	18.8	-1.76	9.06	51	36	30	10.4	75.2	1.73	8.23
TRUE	1-1		33	0	0	0.2	0.6	0.06	0.05	25	0	0	0.9	0.9	-0.06	0.05
	2-1	1	67	67	67	0.2	0.6	-1.40	1.04	67	67	67	0.9	0.9	1.29	0.96
	3-1		67	67	67	0.2	0.6	-1.38	1.03	67	67	67	0.9	0.9	1.29	0.96
	1-2		38	24	18	1.0	10.2	0.00	0.60	38	25	18	2.5	11.5	0.07	0.73
	2-2	17	53	32	28	1.0	10.2	-1.89	8.66	44	34	29	2.5	11.5	1.94	8.17
	3-2		48	29	27	1.0	10.2	-1.99	8.69	43	30	25	2.5	11.5	2.25	8.71
	1-3		18	12	10	2.9	12.6	0.18	0.57	20	10	8	7.5	18.5	-0.08	0.38
	2-3	2	53	47	42	2.9	12.6	-12.57	25.13	53	47	45	7.5	18.5	12.07	23.36
	3-3		53	45	40	2.9	12.6	-12.61	25.40	53	47	47	7.5	18.5	12.14	23.60
	1-4		25	12	8	1.5	32.2	0.01	3.76	26	12	8	7.9	63.3	-0.07	3.95
	2-4	55	43	29	23	1.6	15.6	-1.74	9.73	43	29	22	4.4	22.2	1.97	9.95
	3-4		42	29	23	1.6	15.6	-1.77	9.90	43	30	22	4.4	22.1	1.95	9.84

Table C.11: IPD batch III (106 data-sets). Results on the marginal *y*-axis (cumulative events-count) of the average Breslow simulation in control and treatment group. Average bias (difference) between 1st, 2nd, 3rd *y*-axis quartile and original IPD value. Biased: percentage of quartile differences, including that at *y*-axis range and average, exceeding ±0.1 (first column). Same computation after excluding difference at *y*-axis maximum (second column 2). Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

						Mean bias quartile		
Similar data	*Type*	*Method*	*Size (%)*	*Biased (%)*		*1st*	*2nd*	*3rd*
FALSE	Breslow	1-1	100	19	7	-0.01	-0.02	-0.04
		2-1		24	12	-0.03	-0.07	Inf
		3-1		22	10	-0.01	-0.03	Inf
		1-2	79	21	9	-0.01	-0.02	-0.05
		2-2		27	16	-0.03	-0.08	Inf
		3-2		26	15	-0.02	-0.04	Inf
		1-3	99	13	0	-0.00	-0.00	-0.01
		2-3		16	4	-0.01	-0.02	Inf
		3-3		15	3	-0.01	-0.01	Inf
		1-4	35	13	0	-0.00	-0.00	-0.00
		2-4		18	5	-0.01	-0.02	Inf
		3-4		17	5	-0.01	-0.02	Inf
TRUE		1-2	21	13	0	-0.00	-0.00	-0.01
		2-2		18	4	-0.01	-0.02	-0.03
		3-2		17	2	-0.00	-0.01	-0.03
		1-3	1	17	0	-0.01	-0.03	-0.07
		2-3		17	0	-0.01	-0.03	-0.08
		3-3		17	0	-0.01	-0.03	-0.08
		1-4	65	13	0	-0.00	0.00	-0.00
		2-4		15	3	-0.01	-0.01	-0.03
		3-4		14	2	-0.00	-0.01	-0.02

Table C.12: IPD batch IV (194 data-sets). Results for the average MLE (from GLM and Cox regression with mixed continuous/binary covariates) and reciprocal Fisher Information diagonal simulation. Mean and standard deviation (± s.d.) of original IPD estimate and of difference between simulated and original IPD inference (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

			Maximum Likelihood Estimate							1 / Fisher Information diagonal						
			Biased (%)			*IPD Value*		*Bias*		*Biased (%)*			*IPD Value*		*Bias*	
Similar data	*Method*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1		54	39	29	4.2	30.6	-0.33	7.40	14	9	6	25.2	457.5	-0.31	5.65
	2-1	100	54	40	32	4.3	30.6	-0.33	7.53	18	13	9	17.0	133.0	-0.81	6.90
	3-1		39	23	18	5.1	37.0	0.05	5.76	17	12	9	17.4	140.8	-0.98	8.58
	1-2		52	39	30	4.4	31.2	-0.33	7.73	15	8	6	15.3	118.4	0.15	5.08
	2-2	98	53	40	32	4.4	31.3	-0.33	8.00	19	12	9	24.2	456.0	-0.34	5.93
	3-2		36	21	16	4.9	35.9	0.14	5.32	16	10	8	16.6	124.5	-0.47	7.48
	1-3		31	17	12	5.2	38.7	0.00	1.67	13	8	6	26.6	202.4	-0.17	4.81
	2-3	94	33	18	14	5.1	38.8	0.12	3.63	15	10	8	26.5	203.4	-0.49	4.70
	3-3		30	17	13	5.2	38.8	-0.05	3.73	15	10	8	26.2	202.9	-0.50	5.04
	1-4		28	16	13	4.2	37.7	0.00	2.88	13	7	7	24.4	169.2	-0.12	6.25
	2-4	59	29	18	13	4.2	37.8	0.25	6.03	14	8	6	39.6	596.9	-0.28	5.00
	3-4		25	14	11	4.5	39.7	0.31	6.63	13	8	6	24.4	170.3	-0.20	5.40
TRUE	1-2		44	22	16	4.3	22.6	0.21	1.14	6	3	0	1.0	3.3	-0.02	0.07
	2-2	2	44	25	22	4.3	22.6	0.21	1.25	12	12	12	1.0	3.3	-0.34	1.22
	3-2		19	12	9	4.3	22.6	0.03	1.03	12	9	9	1.0	3.3	-0.34	1.24
	1-3		22	1	0	5.9	20.2	-0.01	0.09	0	0	0	0.8	1.3	0.00	0.01
	2-3	6	26	12	6	5.9	20.2	0.03	0.24	28	21	5	0.8	1.3	-0.12	0.23
	3-3		30	18	10	5.9	20.2	0.03	0.33	28	21	6	0.8	1.3	-0.12	0.23
	1-4		22	12	8	6.6	37.8	-0.03	1.26	11	5	4	34.6	465.3	-0.02	4.74
	2-4	41	28	17	11	6.6	37.8	0.02	1.36	19	13	9	21.0	181.0	-0.30	4.15
	3-4		29	18	14	6.6	37.9	-0.04	2.51	19	14	9	21.1	181.3	-0.60	4.09

Table C.13: IPD batch IV (194 data-sets). Results for 95% empirical CIs of the MLE simulation. Mean and standard deviation (± s.d.) of original IPD CIs (any of the type percent, normal Bca and basic IPD bootstrap quantiles) and of difference between reconstructed and IPD quantile (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

			2.5th Percentile							97.5th Percentile						
			Biased (%)			*IPD Value*		*Bias*		*Biased (%)*			*IPD Value*		*Bias*	
Similar data	*Method*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1		54	40	30	2.2	26.6	-0.10	8.27	53	39	30	7.0	39.6	-0.61	8.58
	2-1	100	60	46	40	2.0	26.7	-2.15	12.50	60	45	38	6.2	39.0	1.98	11.79
	3-1		55	37	31	1.7	27.3	-2.94	12.29	55	36	30	7.3	45.5	3.02	12.53
	1-2		53	40	30	2.3	27.0	-0.09	8.29	53	38	30	7.0	39.8	-0.60	8.87
	2-2	98	59	47	40	1.9	26.2	-2.10	12.28	58	46	37	5.2	30.5	2.04	11.63
	3-2		52	34	29	1.4	25.3	-2.82	11.53	51	36	30	6.6	42.6	3.10	12.49
	1-3		37	21	15	2.5	32.4	0.30	4.04	36	21	15	8.6	51.3	-0.33	4.42
	2-3	94	52	34	28	2.2	31.9	-2.18	10.02	52	33	27	7.7	45.3	2.25	10.04
	3-3		53	34	28	2.2	31.6	-2.23	10.18	52	33	27	7.6	44.8	2.22	10.16
	1-4		35	20	15	1.5	31.7	0.39	5.35	34	20	15	7.8	51.2	-0.46	5.67
	2-4	59	50	32	25	1.5	30.9	-1.06	8.21	50	30	23	6.9	46.1	1.29	8.62
	3-4		51	31	24	1.3	32.1	-1.12	8.03	49	31	24	6.6	42.7	1.17	7.92
TRUE	1-2		47	21	17	3.8	22.1	0.19	1.08	40	27	18	5.7	24.9	0.20	1.12
	2-2	2	54	30	29	3.8	22.1	-6.49	20.52	60	32	29	5.7	24.9	7.11	20.71
	3-2		48	23	23	3.8	22.1	-6.66	20.27	43	23	21	5.7	24.9	6.96	21.19
	1-3		31	13	3	4.8	19.6	0.01	0.20	31	10	4	7.0	20.8	0.01	0.21
	2-3	6	52	37	33	4.8	19.5	-3.66	5.62	59	41	36	7.0	20.8	3.83	5.87
	3-3		52	37	34	4.8	19.6	-3.66	5.64	61	40	37	7.0	20.8	3.84	5.85
	1-4		29	15	12	4.4	31.7	0.06	2.44	29	16	12	9.4	48.1	-0.10	2.34
	2-4	41	49	35	31	4.2	31.3	-3.78	11.34	50	36	31	8.4	42.3	3.82	10.85
	3-4		50	35	30	4.2	31.4	-3.84	11.61	51	36	31	8.5	43.1	4.02	11.38

Table C.14: IPD batch IV (195 data-sets). Results on the marginal *y*-axis (cumulative events-count) of the average Breslow simulation in control and treatment group. Average bias (difference) between 1st, 2nd, 3rd *y*-axis quartile and original IPD value. Biased: percentage of quartile differences, including that at *y*-axis range and average, exceeding ±0.1 (first column). Same computation after excluding difference at *y*-axis maximum (second column 2). Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

				Control					Treatment				
						Mean bias quartile					*Mean bias quartile*		
Similar data	*Type*	*Method*	*Size (%)*	*Biased (%)*		*1st*	*2nd*	*3rd*	*Biased (%)*		*1st*	*2nd*	*3rd*
FALSE	Breslow	1-1		18	7	-0.00	-0.01	-0.02	20	8	Inf	Inf	Inf
		2-1	100	19	8	-0.00	-0.00	-0.01	25	14	Inf	Inf	Inf
		3-1		17	6	Inf	Inf	Inf	23	13	Inf	Inf	Inf
		1-2		18	8	-0.00	-0.01	-0.02	21	8	-0.01	-0.01	-0.03
		2-2	98	18	7	-0.00	-0.01	-0.03	26	15	Inf	Inf	Inf
		3-2		18	6	-0.00	-0.00	Inf	24	13	Inf	Inf	Inf
		1-3		14	2	0.00	-0.00	-0.00	13	2	-0.00	0.00	0.01
		2-3	94	15	3	0.00	-0.00	Inf	18	7	-0.01	Inf	Inf
		3-3		15	3	0.00	0.00	Inf	18	7	-0.01	Inf	Inf
		1-4		12	2	-0.00	-0.00	-0.00	12	2	0.00	0.00	0.00
		2-4	58	13	3	-0.00	-0.00	Inf	19	9	-0.01	-0.02	Inf
		3-4		12	3	0.00	0.00	0.45	19	8	-0.01	-0.02	Inf
TRUE		1-2		11	0	-0.00	-0.01	-0.00	6	0	0.00	0.01	0.01
		2-2	2	11	0	-0.00	-0.00	-0.00	11	0	0.00	0.00	0.01
		3-2		11	0	-0.00	-0.00	-0.01	11	0	-0.00	-0.01	0.00
		1-3		24	8	-0.01	-0.04	-0.09	18	2	-0.00	-0.03	-0.06
		2-3	6	25	10	-0.01	-0.04	-0.10	18	2	-0.00	-0.03	-0.07
		3-3		25	10	-0.01	-0.04	-0.10	18	2	-0.00	-0.02	-0.06
		1-4		16	4	-0.00	-0.01	-0.01	15	2	-0.01	-0.01	-0.01
		2-4	42	18	5	-0.00	-0.00	0.00	17	5	-0.01	Inf	Inf
		3-4		16	4	0.00	0.00	0.02	16	4	-0.01	Inf	Inf

C.2.5 More on MLE and 95% empirical CIs reconstruction

We give further results from Appendix C.2, page 103 on point and interval estimates. For sake of clearness we report results for a reduced selection of used sampling methods. Such selection may vary from batch to batch mainly depending on the experiment size there and on method relevance. Also, we only report results on the MLE (Table C.15 to C.21, page 118 to 124), because in most istances the bias dimension of the simulated r.f.i.d. is small.If the MLE is vector-valued we rank each vector element in ascending order. For instance a MLE intercept (if any) has rank 1 while the first covariate effect has rank 2, and so on. If no intercept is modeled rank 1 is assigned to the first covariate effect. Here we stratify MLE results by parameter rank and then by employed statistical model. Tables C.22 – C.25 (pages 125 – 128) show further results on 95% empirical intervals stratified by type of IPD reference bootstrap CI.

Table C.15: IPD batch I (109 data-sets). Results for the average HR (from Cox regression with one binary covariate) and reciprocal Fisher Information diagonal simulation. Mean and standard deviation (± s.d.) of original IPD estimate and of difference between simulated and original IPD inference (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Rank: MLE vector element (1: intercept, 2: first covariate, and so on). Size: number of IPD examples relative to total batch size (%).

				Maximum Likelihood Estimate						
				Biased (%)			*IPD Value*		*Bias*	
Similar data	*Method*	*Rank*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1	1	99	72	39	17	0.13	0.64	0.00	0.45
	3-1		99	17	1	1	0.13	0.64	-0.02	0.11
	1-3		87	8	2	2	0.15	0.65	0.02	0.09
TRUE	1-2		26	31	0	0	0.23	0.49	0.03	0.10
	1-3		13	0	0	0	0.02	0.49	0.01	0.04
	1-4		88	7	1	0	0.15	0.65	0.02	0.06

C.2.6 Expected marginal event-time of the cumulative hazard estimate

Here we report tabular results on the marginal x-axis (event-time line) of the expected c.h.e. (Table C.26 to C.28, pages 129–131). The format is slightly different from tabular results on the y-axis

Table C.16: IPD batch II (241 data-sets). Results for the average MLE (from GLM and Cox regression with one continuous covariate) and reciprocal Fisher Information diagonal simulation. Mean and standard deviation (± s.d.) of original IPD estimate and of difference between simulated and original IPD inference (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Rank: MLE vector element (1: intercept, 2: first covariate, and so on). Size: number of IPD examples relative to total batch size (%).

				Maximum Likelihood Estimate						
				Biased (%)			*IPD Value*		*Bias*	
Similar data	*Method*	*Rank*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1	1	83	55	30	21	24.70	105.88	1.63	39.85
	1-1	2	83	15	8	7	0.87	11.01	-0.22	7.30
	1-3	1	84	38	19	15	23.90	104.89	0.14	2.24
	1-3	2	84	14	6	4	0.86	10.90	-0.01	0.33
TRUE	1-2	1	77	52	20	11	20.59	104.27	2.07	37.71
	1-2	2	77	8	4	2	0.90	11.18	-0.32	7.05
	1-4	1	73	17	8	5	21.73	106.68	0.01	0.73
	1-4	2	73	4	1	1	0.93	11.44	-0.01	0.42

(see Section 3.2, page 53). For sake of conciseness we just focus on the average difference between the simulated and IPD original mean time-line.

Table C.17: IPD batch III (191 data-sets). Results for the average MLE (from GLM and Cox regression with two continuous covariates) and reciprocal Fisher Information diagonal simulation. Mean and standard deviation (± s.d.) of original IPD estimate and of difference between simulated and original IPD inference (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Rank: MLE vector element (1: intercept, 2: first covariate, and so on). Size: number of IPD examples relative to total batch size (%).

				Maximum Likelihood Estimate						
				Biased (%)			*IPD Value*		*Bias*	
Similar data	*Method*	*Rank*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1	1	98	70	52	40	9.55	47.65	0.41	6.93
	1-1	2	99	20	11	6	0.44	8.55	0.17	2.21
	1-1	3	99	42	27	19	0.42	4.55	-0.49	4.86
	1-3	1	98	42	23	17	13.39	70.29	-0.09	2.68
	1-3	2	98	14	2	2	0.86	11.11	-0.04	0.16
	1-3	3	98	17	8	7	0.41	4.59	0.01	0.75
TRUE	1-2	1	17	60	35	26	3.61	15.38	0.03	0.77
	1-2	2	17	13	7	2	-0.06	0.43	-0.04	0.12
	1-2	3	17	18	12	9	0.02	0.28	-0.00	0.29
	1-4	1	55	28	11	7	9.80	56.83	0.04	1.10
	1-4	2	55	3	1	1	0.70	11.05	0.01	0.25
	1-4	3	55	16	5	3	0.50	5.08	-0.12	1.57

Table C.18: IPD batch IV (194 data-sets). Results for the average MLE (from GLM and Cox regression with mixed continuous/binary covariates) and reciprocal Fisher Information diagonal simulation. Mean and standard deviation (± s.d.) of original IPD estimate and of difference between simulated and original IPD inference (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Rank: MLE vector element (1: intercept, 2: first covariate, and so on). Size: number of IPD examples relative to total batch size (%).

				Maximum Likelihood Estimate						
				Biased (%)			*IPD Value*		*Bias*	
Similar data	*Method*	*Rank*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1	1	100	70	54	44	12.59	52.88	-0.33	7.71
	1-1	2	100	32	20	11	-0.06	1.53	0.09	0.88
	1-1	3	100	53	37	28	1.44	10.55	-0.98	10.11
	1-1	4	97	65	50	40	0.27	8.93	-0.21	9.16
	1-3	1	94	46	29	23	15.44	65.47	-0.14	2.33
	1-3	2	94	19	6	3	-0.09	1.56	-0.02	0.19
	1-3	3	94	28	17	11	1.44	11.95	0.13	1.68
	1-3	4	91	32	16	10	0.36	23.53	0.14	1.73
TRUE	**1-3**	1	6	25	0	0	19.68	32.29	0.04	0.09
	1-3	2	6	27	0	0	0.39	0.86	-0.04	0.09
	1-3	3	6	11	0	0	-1.71	2.53	-0.01	0.06
	1-3	4	6	25	8	0	0.78	5.46	-0.02	0.12
	1-4	1	41	37	21	17	19.56	61.98	-0.13	1.34
	1-4	2	41	10	4	1	-0.12	1.73	0.00	0.17
	1-4	3	41	20	11	7	0.48	11.95	0.16	1.69
	1-4	4	39	18	9	6	2.67	28.00	-0.15	1.51

Table C.19: IPD batch II (241 data-sets). Results for the average MLE (from GLM and Cox regression with one continuous covariate) and reciprocal Fisher Information diagonal simulation. Mean and standard deviation (± s.d.) of original IPD estimate and of difference between simulated and original IPD inference (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

				Maximum Likelihood Estimate						
				Biased (%)			*IPD Value*		*Bias*	
Similar data	*Method*	*Model*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1	gaussian	42	39	27	21	37.66	127.10	2.27	49.11
	1-1	poisson	41	21	8	5	-0.44	1.03	0.04	0.34
	1-1	binomial	41	45	21	16	0.38	3.45	-0.25	2.00
	1-3	gaussian	41	16	8	6	38.02	128.28	0.03	1.07
	1-3	poisson	43	20	7	4	-0.44	1.02	0.05	0.41
	1-3	binomial	43	40	23	18	0.32	3.39	0.10	2.52
TRUE	1-2	gaussian	34	34	19	12	37.74	135.51	3.04	50.49
	1-2	poisson	42	18	2	0	-0.46	1.11	-0.00	0.11
	1-2	binomial	42	38	15	7	-0.02	1.96	-0.02	0.32
	1-4	gaussian	33	15	9	8	39.72	138.99	0.02	1.10
	1-4	poisson	41	4	1	1	-0.45	1.13	-0.00	0.06
	1-4	binomial	41	13	5	2	0.23	2.82	-0.00	0.14

Table C.20: IPD batch III (191 data-sets). Results for the average MLE (from GLM and Cox regression with two continuous covariates) and reciprocal Fisher Information diagonal simulation. Mean and standard deviation (± s.d.) of original IPD estimate and of difference between simulated and original IPD inference (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

				Maximum Likelihood Estimate						
				Biased (%)			*IPD Value*		*Bias*	
Similar data	*Method*	*Model*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1	gaussian	44	51	38	32	17.39	59.70	0.78	9.39
	1-1	poisson	55	53	31	18	-2.53	3.43	0.03	0.62
	1-1	binomial	55	49	38	29	0.73	3.44	-0.23	3.12
	1-1	cox	55	23	12	3	0.03	0.49	-0.02	0.22
	1-3	gaussian	43	18	10	9	24.13	88.33	-0.14	2.81
	1-3	poisson	55	25	7	4	-2.56	3.43	0.01	0.29
	1-3	binomial	55	43	22	17	0.73	3.46	-0.04	1.79
	1-3	cox	55	9	2	1	0.03	0.49	-0.00	0.08
TRUE	1-2	gaussian	6	30	21	9	8.84	21.64	0.01	0.28
	1-2	poisson	12	45	18	14	-2.13	3.42	-0.01	0.50
	1-2	binomial	12	39	29	21	1.11	3.17	0.02	0.73
	1-2	cox	12	14	5	5	0.17	0.42	-0.03	0.16
	1-4	gaussian	19	27	10	7	25.18	82.83	-0.03	2.18
	1-4	poisson	36	14	4	2	-2.55	3.49	0.00	0.14
	1-4	binomial	36	19	8	5	0.55	2.56	0.02	0.43
	1-4	cox	36	3	1	0	0.01	0.44	0.00	0.06

Table C.21: Data batch IV (194 data-sets). Results for the average MLE (from GLM and Cox regression with mixed continuous/binary covariates) and reciprocal Fisher Information diagonal simulation. Mean and standard deviation (± s.d.) of original IPD estimate and of difference between simulated and original IPD inference (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

				Maximum Likelihood Estimate						
				Biased (%)			*IPD Value*		*Bias*	
Similar data	*Method*	*Model*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1	gaussian	93	64	52	46	15.22	54.98	-0.85	13.50
	1-1	poisson	100	52	34	20	-1.49	3.08	0.01	0.61
	1-1	binomial	100	51	37	29	0.49	2.58	-0.25	2.24
	1-1	cox	100	40	25	12	0.09	0.65	-0.03	0.35
	1-3	gaussian	87	32	21	15	18.31	69.57	0.07	2.43
	1-3	poisson	94	26	11	7	-1.61	3.14	0.00	0.37
	1-3	binomial	94	41	23	17	0.47	2.64	-0.04	1.82
	1-3	cox	94	18	7	4	0.09	0.66	-0.02	0.15
TRUE	**1-3**	gaussian	6	67	4	0	18.27	33.43	-0.02	0.16
	1-3	poisson	6	0	0	0	0.28	0.47	0.01	0.03
	1-3	binomial	6	4	0	0	0.68	1.39	-0.01	0.03
	1-3	cox	6	4	0	0	0.03	0.33	0.01	0.04
	1-4	gaussian	38	30	17	13	22.49	67.27	-0.02	2.28
	1-4	poisson	41	19	9	6	-0.92	2.43	-0.02	0.17
	1-4	binomial	41	21	11	9	0.72	2.46	-0.05	0.39
	1-4	cox	41	13	6	1	0.08	0.70	-0.01	0.11

Table C.22: IPD batch I (109 data-sets). Results for 95% empirical CIs of the MLE simulation. Mean and standard deviation (± s.d.) of original IPD CIs (any of the type percent, normal Bca and basic IPD bootstrap quantiles) and of difference between reconstructed and IPD quantile (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Type: IPD reference bootstrap CI type. Size: number of IPD examples relative to total batch size (%).

				2.5th Percentile							97.5th Percentile						
				Biased (%)			*IPD Value*		*Bias*		*Biased (%)*			*IPD Value*		*Bias*	
Similar data	*Method*	*Type*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1	basic	99	72	39	17	-0.12	0.58	0.04	0.36	68	33	17	0.47	0.61	-0.09	0.35
	1-1	BCa	99	69	31	12	-0.12	0.67	0.04	0.34	65	22	13	0.60	0.71	-0.14	0.33
	1-1	normal	99	72	39	17	-0.14	0.61	0.05	0.37	68	33	17	0.48	0.65	-0.11	0.39
	1-1	percent	99	75	40	20	-0.12	0.61	0.04	0.37	72	37	18	0.47	0.65	-0.09	0.37
	1-3	basic	87	14	4	3	-0.10	0.59	0.04	0.14	14	4	3	0.49	0.62	-0.00	0.12
	1-3	BCa	87	15	5	5	-0.08	0.71	0.06	0.15	22	4	2	0.67	0.74	0.02	0.12
	1-3	normal	87	14	6	5	-0.11	0.63	0.06	0.17	16	5	3	0.51	0.67	-0.02	0.18
	1-3	percent	87	17	3	2	-0.10	0.62	0.04	0.10	18	4	1	0.49	0.67	0.00	0.10
TRUE	1-2	basic	27	31	0	0	0.03	0.50	0.06	0.10	38	0	0	0.43	0.49	-0.01	0.10
	1-2	BCa	27	27	0	0	0.16	0.42	0.06	0.09	33	0	0	0.61	0.45	-0.04	0.09
	1-2	normal	27	34	0	0	0.02	0.51	0.06	0.10	34	0	0	0.43	0.49	-0.00	0.10
	1-2	percent	27	34	0	0	0.02	0.52	0.07	0.10	34	0	0	0.42	0.50	0.00	0.10
	1-4	basic	88	11	2	1	-0.06	0.56	0.03	0.09	10	1	0	0.46	0.61	-0.00	0.07
	1-4	BCa	88	8	2	2	-0.06	0.67	0.04	0.09	8	2	0	0.61	0.72	-0.01	0.08
	1-4	normal	88	9	3	1	-0.07	0.57	0.03	0.09	9	1	0	0.46	0.63	-0.00	0.07
	1-4	percent	88	9	1	1	-0.06	0.59	0.03	0.08	10	1	0	0.47	0.64	-0.01	0.07

Table C.23: IPD batch II (241 data-sets). Results for 95% empirical CIs of the MLE simulation. Mean and standard deviation (± s.d.) of original IPD CIs (any of the type percent, normal Bca and basic IPD bootstrap quantiles) and of difference between reconstructed and IPD quantile (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Type: IPD reference bootstrap CI type. Size: number of IPD examples relative to total batch size (%).

				2.5th Percentile							97.5th Percentile						
				Biased (%)			*IPD Value*		*Bias*		*Biased (%)*			*IPD Value*		*Bias*	
Similar data	*Method*	*Type*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1	basic	83	39	23	17	8.24	43.77	-0.07	5.88	38	22	17	11.50	56.34	-0.03	3.57
	1-1	BCa	83	37	23	18	8.70	45.80	-0.67	6.85	40	23	18	12.83	61.42	-0.38	5.32
	1-1	normal	83	38	23	18	8.30	44.09	-0.05	6.66	39	22	17	12.21	57.59	-0.22	6.31
	1-1	percent	83	37	23	17	8.73	43.88	-0.53	5.91	38	23	16	12.37	57.79	-0.31	5.40
	1-3	basic	84	32	16	11	9.40	56.80	0.34	2.93	31	17	12	14.90	94.45	0.25	2.94
	1-3	BCa	84	33	18	14	10.08	60.92	-0.23	3.33	32	18	15	16.33	102.47	-0.34	3.47
	1-3	normal	84	32	16	12	9.50	57.45	0.31	4.59	30	15	12	15.20	95.04	-0.13	4.30
	1-3	percent	84	32	16	12	9.93	57.75	-0.14	2.81	31	16	13	15.39	95.84	-0.26	3.27
TRUE	1-2	basic	77	36	18	12	6.95	40.53	-0.08	4.65	37	19	12	9.25	52.46	-0.04	3.95
	1-2	BCa	77	36	20	12	6.90	42.68	-0.23	5.60	37	20	14	9.55	56.50	-0.17	3.21
	1-2	normal	77	36	19	11	6.98	40.67	-0.11	4.70	37	19	12	9.28	52.68	-0.07	3.56
	1-2	percent	77	36	20	12	7.02	40.67	-0.14	4.69	37	19	12	9.31	52.78	-0.10	3.00
	1-4	basic	73	18	8	6	8.96	57.36	0.14	1.41	19	7	4	13.57	96.64	0.05	2.29
	1-4	BCa	73	21	7	5	9.35	61.52	-0.06	2.01	20	8	4	14.78	105.12	-0.15	2.01
	1-4	normal	73	17	7	6	9.02	57.56	0.08	1.29	18	7	3	13.60	96.77	0.01	1.96
	1-4	percent	73	18	7	4	9.09	57.66	0.01	1.57	16	7	4	13.71	97.07	-0.08	2.13

Table C.24: IPD batch III (191 data-sets). Results for 95% empirical CIs of the MLE simulation. Mean and standard deviation (± s.d.) of original IPD CIs (any of the type percent, normal Bca and basic IPD bootstrap quantiles) and of difference between reconstructed and IPD quantile (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Type: IPD reference bootstrap CI type. Size: number of IPD examples relative to total batch size (%).

				2.5th Percentile							97.5th Percentile						
				Biased (%)			*IPD Value*		*Bias*		*Biased (%)*			*IPD Value*		*Bias*	
Similar data	*Method*	*Type*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1	basic	99	48	33	26	1.92	17.45	0.19	6.77	46	33	24	6.50	45.82	-0.04	6.11
	1-1	BCa	99	48	35	28	1.32	15.67	0.19	7.57	48	35	26	6.84	50.46	0.06	5.85
	1-1	normal	99	49	33	25	1.85	17.72	0.39	7.47	47	32	26	6.57	46.34	-0.29	7.12
	1-1	percent	99	49	34	26	1.95	17.46	0.15	6.95	48	34	26	6.75	46.80	0.01	6.78
	1-3	basic	98	36	18	12	2.65	23.38	0.49	4.80	35	19	13	9.37	70.10	-0.28	4.86
	1-3	BCa	98	37	21	14	2.41	26.05	0.26	5.92	36	20	15	11.67	89.36	-0.60	4.79
	1-3	normal	98	35	17	12	2.65	24.24	0.52	4.89	34	19	12	10.80	80.53	-0.41	5.87
	1-3	percent	98	35	19	14	2.97	25.24	0.21	5.58	34	19	14	10.70	80.35	-0.37	5.41
TRUE	1-2	basic	17	39	23	19	0.96	10.10	0.03	0.57	38	25	18	2.33	11.35	0.09	0.76
	1-2	BCa	17	37	26	19	1.31	10.62	-0.03	0.59	40	26	19	3.15	12.20	0.06	0.73
	1-2	normal	17	38	24	18	0.97	10.12	0.02	0.58	36	23	17	2.34	11.37	0.08	0.73
	1-2	percent	17	38	22	18	1.00	10.16	-0.01	0.66	38	25	18	2.37	11.41	0.05	0.69
	1-4	basic	55	26	11	8	1.35	34.23	0.07	4.11	27	12	8	8.29	66.74	0.00	4.36
	1-4	BCa	55	26	13	8	1.87	23.30	0.11	2.12	26	13	8	6.18	46.33	-0.10	1.33
	1-4	normal	55	23	11	8	1.45	34.33	-0.03	4.07	26	12	8	8.32	67.50	-0.06	4.22
	1-4	percent	55	24	11	8	1.51	34.33	-0.08	4.10	24	11	8	8.41	67.66	-0.14	4.62

Table C.25: Data batch IV (194 data-sets). Results for 95% empirical CIs of the MLE simulation. Mean and standard deviation (± s.d.) of original IPD CIs (any of the type percent, normal Bca and basic IPD bootstrap quantiles) and of difference between reconstructed and IPD quantile (Bias). Biased (%): percentage of differences exceeding ±0.1, ±0.3 and ±0.5. Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Table 2.8, page 46. Type: IPD reference bootstrap CI type. Size: number of IPD examples relative to total batch size (%).

				2.5th Percentile							97.5th Percentile						
				Biased (%)			*IPD Value*		*Bias*		*Biased (%)*			*IPD Value*		*Bias*	
Similar data	*Method*	*Type*	*Size (%)*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*	*> .1*	*> .3*	*> .5*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1	basic	100	54	39	29	2.03	24.98	-0.11	7.46	54	38	29	6.48	37.45	-0.49	8.28
	1-1	BCa	100	54	41	33	2.80	30.83	-0.20	9.32	54	40	31	8.90	46.17	-0.90	9.78
	1-1	normal	100	54	40	30	2.01	25.58	-0.00	8.12	53	39	29	6.62	38.13	-0.64	8.64
	1-1	percent	100	54	40	31	2.01	25.73	-0.10	8.38	53	39	30	6.40	38.15	-0.48	7.88
	1-3	basic	94	35	20	14	2.43	30.65	0.29	2.96	36	20	15	7.93	48.65	-0.21	4.39
	1-3	BCa	94	42	26	19	3.21	37.46	0.28	4.67	42	25	19	11.19	59.74	-0.56	5.07
	1-3	normal	94	34	19	14	2.34	31.26	0.39	4.25	35	20	14	8.13	49.35	-0.40	4.90
	1-3	percent	94	36	19	14	2.39	31.41	0.24	4.28	34	19	14	7.87	49.53	-0.23	3.32
TRUE	**1-3**	basic	6	35	13	1	4.78	19.62	0.01	0.17	30	11	4	7.01	20.85	0.02	0.19
	1-3	BCa	6	30	13	3	4.79	19.58	0.01	0.18	30	10	4	7.01	20.82	0.02	0.20
	1-3	normal	6	30	13	3	4.77	19.61	0.02	0.25	29	10	4	7.02	20.84	-0.00	0.26
	1-3	percent	6	29	12	4	4.80	19.60	-0.00	0.18	33	11	4	7.02	20.84	0.00	0.20
	1-4	basic	41	30	16	12	4.16	30.60	0.13	2.41	29	16	12	8.95	46.39	-0.02	2.30
	1-4	BCa	41	32	17	13	5.10	34.46	-0.01	2.86	32	18	13	10.75	52.61	-0.19	2.98
	1-4	normal	41	29	15	12	4.19	30.91	0.10	2.27	29	16	12	9.00	46.80	-0.07	2.08
	1-4	percent	41	27	13	10	4.27	31.19	0.02	2.23	27	15	11	9.06	47.24	-0.12	1.99

Table C.26: IPD batch I (111 data-sets). Results on the marginal x-axis (event-time line on log scale) of the average Nelson-Aalen simulation in control and treatment group. Mean and standard deviation (± s.d.) of the difference between simulated and IPD original mean x-axis (Bias). Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: see Section 2.2.3, page 38. Size: number of IPD examples relative to total batch size (%).

			Bias			
			Control		Treatment	
Similar data	*Method*	*Size (%)*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1	99	0.18	0.18	0.20	0.26
	2	74	0.18	0.28	0.26	0.51
	3	87	0.08	0.13	0.11	0.16
	4	12	0.10	0.11	0.15	0.20
TRUE	1	1	0.13	0.03	0.10	0.02
	2	26	0.11	0.09	0.16	0.21
	3	13	0.07	0.06	0.03	0.03
	4	88	0.06	0.05	0.08	0.12

Table C.27: IPD batch III (106 data-sets). Results on the marginal x-axis (event-time line on log scale) of the average Breslow simulation in control and treatment group. Mean and standard deviation (± s.d.) of the difference between simulated and IPD original mean x-axis (Bias). Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

			Bias	
Similar data	*Method*	*Size (%)*	*Mean*	*± s.d.*
FALSE	1-1	100	0.09	0.11
	2-1	99	0.11	0.17
	3-1	100	0.09	0.11
	1-2	79	0.09	0.13
	2-2	78	0.11	0.23
	3-2	79	0.09	0.12
	1-3	99	0.02	0.02
	2-3	99	0.02	0.02
	3-3	99	0.02	0.02
	1-4	35	0.03	0.05
	2-4	35	0.03	0.05
	3-4	35	0.03	0.06
TRUE	1-2	21	0.07	0.07
	2-2	21	0.07	0.07
	3-2	21	0.06	0.07
	1-3	1	0.02	
	2-3	1	0.02	
	3-3	1	0.03	
	1-4	65	0.01	0.01
	2-4	65	0.01	0.01
	3-4	65	0.01	0.01

Table C.28: IPD batch IV (195 data-sets). Results on the marginal *x*-axis (event-time line on log scale) of the average Breslow simulation in control and treatment group. Mean and standard deviation (± s.d.) of the difference between simulated and IPD original mean *x*-axis (Bias). Similar data: Boolean for similarity of IPD simulation relative to original IPD (see Table 3.1, page 50). Method: Table 2.8, page 46. Size: number of IPD examples relative to total batch size (%).

			Bias			
			Control		Treatment	
Similar data	*Method*	*Size (%)*	*Mean*	*± s.d.*	*Mean*	*± s.d.*
FALSE	1-1	100	0.11	0.15	0.10	0.11
	2-1	99	0.11	0.15	0.11	0.18
	3-1	100	0.11	0.12	0.11	0.14
	1-2	98	0.08	0.08	0.09	0.15
	2-2	98	0.08	0.12	0.13	0.34
	3-2	98	0.12	0.40	0.09	0.11
	1-3	94	0.04	0.12	0.03	0.09
	2-3	94	0.04	0.12	0.03	0.09
	3-3	94	0.05	0.12	0.04	0.09
	1-4	58	0.03	0.02	0.02	0.02
	2-4	58	0.03	0.02	0.02	0.02
	3-4	58	0.03	0.03	0.02	0.02
TRUE	1-2	2	0.07	0.06	0.28	0.40
	2-2	2	0.07	0.06	0.28	0.40
	3-2	2	0.10	0.10	0.31	0.46
	1-3	6	0.01	0.00	0.01	0.00
	2-3	6	0.01	0.00	0.01	0.00
	3-3	6	0.02	0.01	0.02	0.01
	1-4	42	0.02	0.04	0.02	0.04
	2-4	42	0.02	0.04	0.02	0.04
	3-4	42	0.02	0.04	0.02	0.05

C.3 Practical examples

Below we give details related to Section 3.3, page 59.

C.3.1 Sampling under sub-optimal settings

We give further details from the example of Section 3.3.1, page 59. Below we compare the original IPD sample moments and correlation entries against their corresponding simulated averages (column mc.mean)

```
> Xspacelist[[1]]$is.data.similar
$first.moment
             ipd    mc.mean
time   35.610457 35.6313057
status  0.393401  0.3929442
treat   0.500000  0.5023942
agedx  20.781726 20.7291342

$second.moment
              ipd    mc.mean
time   21.3574963 21.3630638
status  0.4891256  0.4881107
treat   0.5006357  0.4999266
agedx  14.8120737 14.7634833

$third.moment
              ipd      mc.mean
time   -0.1127469  1.168552162
status  0.4364299  0.440889975
treat   0.0000000 -0.009620705
agedx   0.8081132  1.393291449

$fourth.moment
            ipd  mc.mean
time   1.748925 4.941795
status 1.190471 1.210235
treat  1.000000 1.011398
agedx  2.538233 5.811978

$lower.triangular.Rx
          ipd      mc.mean
```

```
1 -0.62474958 -0.622961692
2  0.15274712  0.153134265
3  0.01945774  0.019078008
4 -0.24419291 -0.105002862
5  0.03286381 -0.011180187
6  0.00000000 -0.001342778

$bool
[1] FALSE
```

C.3.2 Predictive meaning of MaxEntBoot sample

We give furhter output excerpts from example of Section 3.4.1, page 62. We print the IPD excerpt below

```
> head(datalist$wh.4$ipd.raw.data)
     X_t X_d all10
1 24.665   1     0
2 17.832   1     0
3 27.157   0     0
4 11.135   0     0
5 26.344   0     0
6 27.121   0     0
```

and

```
> tail(datalist$wh.4$ipd.raw.data)
         X_t X_d all10
17255 25.774   0     0
17256 27.168   0     0
17257 19.611   0     0
17258 19.822   1     0
17259 10.029   0     0
17260 12.156   0     0
```

where from left to right we have the survival time, the time-event, and the binary treatment variable respectively. The original IPD moment correlation matrix is

```
> cor(datalist$wh.4$ipd.raw.data)
              X_t        X_d      all10
X_t    1.0000000 -0.4103116 -0.7654625
X_d   -0.4103116  1.0000000  0.2517678
all10 -0.7654625  0.2517678  1.0000000
```

Below we report output of the comparison between simulated and original IPD distributional features.

```
> datalist$wh.4$is.data.similar

$first.moment
             ipd     mc.mean
X_t   21.8067821 21.80938443
X_d    0.1492468  0.14935400
all10  0.0967555  0.09656199

$second.moment
            ipd   mc.mean
X_t   6.8373173 6.8362970
X_d   0.3563419 0.3564295
all10 0.2956331 0.2953475

$third.moment
            ipd   mc.mean
X_t   -1.419183 -1.420870
X_d    1.968690  1.967801
all10  2.728085  2.732431

$fourth.moment
           ipd  mc.mean
X_t   3.878021 3.884749
X_d   4.875740 4.873031
all10 8.442449 8.467847

$lower.triangular.Rx
         ipd    mc.mean
1 -0.4103116 -0.4096435
2 -0.7654625 -0.7632126
3  0.2517678  0.2505669

$bool
[1] TRUE
```

We see a remarkable agreement indicating the reconstructed IPD is overall similar to the original IPD.

Below the output extract of the original IPD Cox analysis,

```
> coxph(Surv(X_t, X_d)~all10,
```

```
data = datalist$wh.4$ipd.raw.data)

          coef exp(coef) se(coef)    z    p
all10 2.31e+01  1.05e+10 4.30e+02 0.05 0.96

Likelihood ratio test=4619  on 1 df, p=0
n= 17260, number of events= 2576
Warning message:
In fitter(X, Y, strats, offset, init,
control, weights = weights,  :
  Loglik converged before variable  1 ;
beta may be infinite.
```

showing problems during partial likelihood maximization. The IPD MLE is singular here.

C.3.3 Replacement of singular IPD statistic

IPD `wh.11` from batch IV is an extension of data `wh.4` (see Section 3.4.1, page 62), allowing for inclusion of an additional variable, `chol`, cholesterol level. Inclusion of this confounder does not resolve the sparse data bias (Greenland et al., 2016) in the original IPD. We generate $B = 100$ NORTAmax repetitions of IPD `wh.11` and obtain MaxEnt bootstrap estimates for Cox parameters and Breslow curves. Below we print the expected log HR vector simulation (row `mc.mean`) against the original IPD value (row `ipd.true`). Column `beta.1` (`vbeta.1`) respectively `beta.2` (`vbeta.2`) refer to the log HR (r.f.i.d.) for cholesterol level and treatment effect.

```
> inf11$raw.results$wh.11$summary$model.summary

              beta.1     beta.2      vbeta.1      vbeta.2      low.1
ipd.true 0.16119223 23.0502300 2.282666e-04 1.664779e+05 0.13157958
mc.mean  0.15490236  4.3016970 2.352515e-04 5.674528e-03 0.12488521
mc.error 0.01741187  0.0667925 8.025577e-06 2.545140e-04 0.01741871
mc.range 0.07529235  0.3599100 3.507468e-05 1.292625e-03 0.07695093
                low.2       up.1         up.2         max        aic
ipd.true -776.66322066 0.19080488 822.76368076 -22068.3111 44140.6222
mc.mean     4.15408806 0.18491951   4.44930588 -22544.4912 45092.9824
mc.error    0.06452077 0.01742805   0.06914852    456.3728   912.7456
mc.range    0.35112331 0.07376655   0.36869676   2500.3129  5000.6259
```

The average log HR (r.f.i.d.) for cholesterol level well agrees with the original IPD estimate. As already seen in Section 3.4.1 the original IPD log HR (r.f.i.d.) is singular but the corresponding MaxEnt bootstrap average is interpretable. Figure C.1 shows the expected Breslow estimate versus its original IPD counterpart in group treatment and control. In group treatment the original IPD estimate is exploding at time 10 while the expected Breslow curve remains interpretable. Next we compare MaxEntBoot 95% empirical quantiles (type `maxent`) against various types of ordinary IPD bootstrap intervals. Intervals for the log HR of cholesterol levels are comparable between MaxEntBoot and IPD reference. IPD bootstrap intervals for the treatment effect are all exploding. Conversely MaxEntBoot intervals are interpretable and comparable to the Normally approximated ones printed above. Original IPD Bca intervals always fail for each effect.

```
> inf11$raw.results$wh.11$summary$model.boot

       2.5th.perc/maxent 97.5th.perc/maxent 2.5th.perc/normal
beta.1          0.123311           0.181509         0.1311305
beta.2          4.156431           4.412071        22.1605762
       97.5th.perc/normal 2.5th.perc/basic 97.5th.perc/basic 2.5th.perc/percent
beta.1          0.1910264        0.1307171         0.1912057          0.1311787
beta.2         23.9925159       22.0626289        23.8825513         22.2179088
       97.5th.perc/percent 2.5th.perc/bca 97.5th.perc/bca
```

```
beta.1          0.1916674            NA             NA
beta.2         24.0378312            NA             NA
```

C.3.4 Long run alternative to IPD statistic

Here we consider IPD `abo2` (batch I) that is a variables' sub-selection of data `abortion` (see Appendix B.6.1, page 96). The effect variable of interest is `group`. We draw $B = 100$ highly honest data samples using NORTAmax resampling with Johnson continuous marginals. The original IPD time variable is tied since visits are scheduled at fixed weekly intervals. However we model this tied time as a continuous variable which amounts to artificially smooth the weekly visit pace. We can argue this choice as an attempt to interpolate the original discrete time approximation. Below we compare outputs between the MaxEnt bootstrap log HR average for `group` effect and the original IPD estimate.

```
> inf2$raw.results$abo2$summary$model.summary
               beta       vbeta         low         up         max        aic
ipd.true -0.17607822 0.011972323 -0.39053771 0.03838128 -6024.68405 12051.3681
mc.mean   0.31962787 0.012405002  0.09993491 0.54151492 -6030.10231 12062.2046
mc.error  0.08050089 0.001170603  0.08233410 0.07708943    79.02521   158.0504
mc.range  0.37964282 0.006760082  0.39142467 0.36262502   421.93378   843.8676
```

We see an appreciable difference between the MaxEntBoot average and the original IPD log HR. This is the single instance of a difference greater than 0.3 in absolute value in our experiments on batch I under method 1-4 (NORTAmax with four moments) when the simulated IPD is similar to reference (see Table C.4, page 107). The MaxEntBoot average log HR indicates the group effect is slightly significantly positive on the long-run. The expected r.f.i.d. estimate (one-to-one to Fisher Information here) well agrees to the IPD original value indicating a good degree of log-likelihood information conservation here. Figure C.2 compares the expected Nelson-Aalen estimate to the original IPD estimate, for each group. In group treatment the MaxEntBoot curve grows slightly faster than the IPD reference. Below we compare MaxEntBoot 95% quantiles versus their IPD ordinary bootstrap counterparts. The former too detect a positive effect (more skewed than the above Normal approximation) as opposed to the latter where no effect is detected.

```
> inf2$raw.results$abo2$summary$model.boot
     2.5th.perc/maxent 97.5th.perc/maxent 2.5th.perc/normal 97.5th.perc/normal
beta          0.154057           0.455637        -0.3706267         0.01158204
     2.5th.perc/basic 97.5th.perc/basic 2.5th.perc/percent 97.5th.perc/percent
beta       -0.3723502        0.01064885         -0.3628053          0.02019379
     2.5th.perc/bca 97.5th.perc/bca
beta     -0.3759239      0.00637373
```

C.3.5 Replacement of singular IPD bootstrap 95% CIs

Here original IPD `rats.2` (batch I) produces singular ordinary bootstrap 95% intervals while MaxEntBoot intervals give an interpretable alternative. This IPD is a variables' sub-selection of data `rats` (see Appendix B.6.1, page 96) and has three-hundreds records with only about 2 observed events among males, as opposed to about 40/42 events among females. This could be a warning on possible sparse data bias (Greenland et al., 2016). We focus on the MLE for rat gender, `sex`. We produce $B = 300$ highly honest IPD simulations via NORTAmax resampling with Johnson marginlas. Below we report the expected MaxEnt bootstrap log HR (r.f.i.d.) estimate versus its original IPD value. We see high agreement everywhere.

```
> inf2$raw.results$rats.2$summary$model.summary
               beta     vbeta       low         up        max       aic
ipd.true -3.0378788 0.5252662 -4.458394 -1.6173638 -203.45262 408.90524
mc.mean  -3.0341921 0.6114286 -4.516205 -1.5564342 -195.93343 393.86686
mc.error  0.6238479 0.3100391  1.002998  0.2728784   29.42033  58.84066
mc.range  2.7713927 0.8701178  3.876740  1.5507519  164.72598 329.45196
```

However, the extremely small number of events among male rats is problematic as shown in Figure C.3, where the IPD cumulative hazard among males (treatment group) is nearly singular. There is good agreement between the MaxEntBoot average Nelson-Alen curve and its IPD counterpart in control group. Below we compare MaxEntBoot 95% CIs with various IPD ordinary bootstrap counterparts. Original IPD intervals comes from 10000 bootstrap realizations. (results with 50000 realizations were materially equal and are not shown).

```
> inf2$raw.results$rats.2$summary$model.boot
     2.5th.perc/maxent 97.5th.perc/maxent 2.5th.perc/normal 97.5th.perc/normal
beta         -3.979393          -1.753119         -12.29472           10.78592
     2.5th.perc/basic 97.5th.perc/basic 2.5th.perc/percent 97.5th.perc/percent
beta        -4.081935          14.29264          -20.36839           -1.993823
     2.5th.perc/bca 97.5th.perc/bca
beta      -20.24085       -1.662684
```

All IPD intervals are exploding while MaxEntBoot intervals are quite reasonable, also if compared with the above Normal approximation.

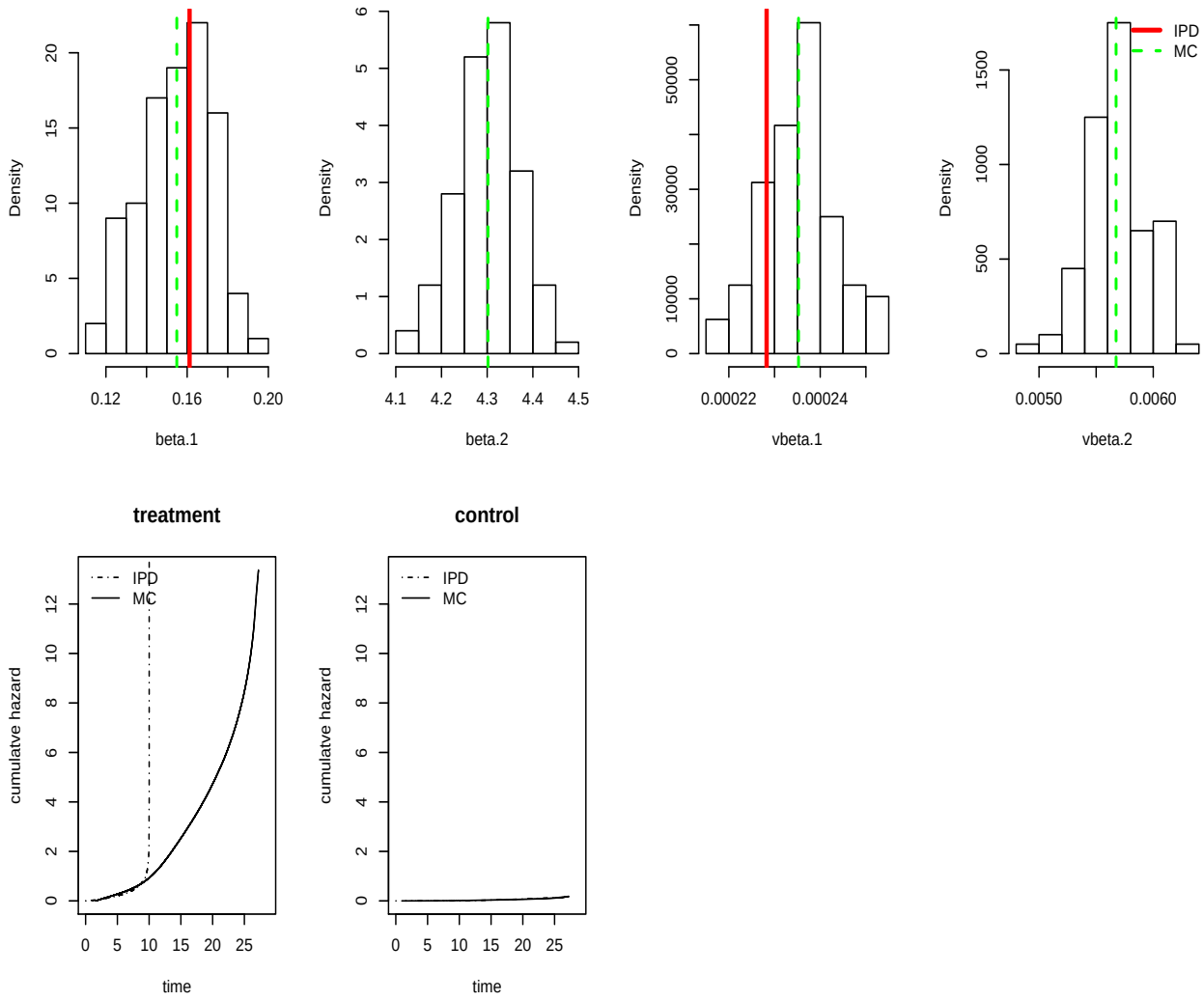

Figure C.1: The `wh.11` data: MaxEntBoot samples for the log HR (`beta.`) and respective reciprocal Fisher Information diagonal (`vbeta.`) alongside Breslow estimates in group treatment and control. MC = Monte Carlo average. IPD = reference estimate. Note some IPD point estimate are outside the plotting range because singular.

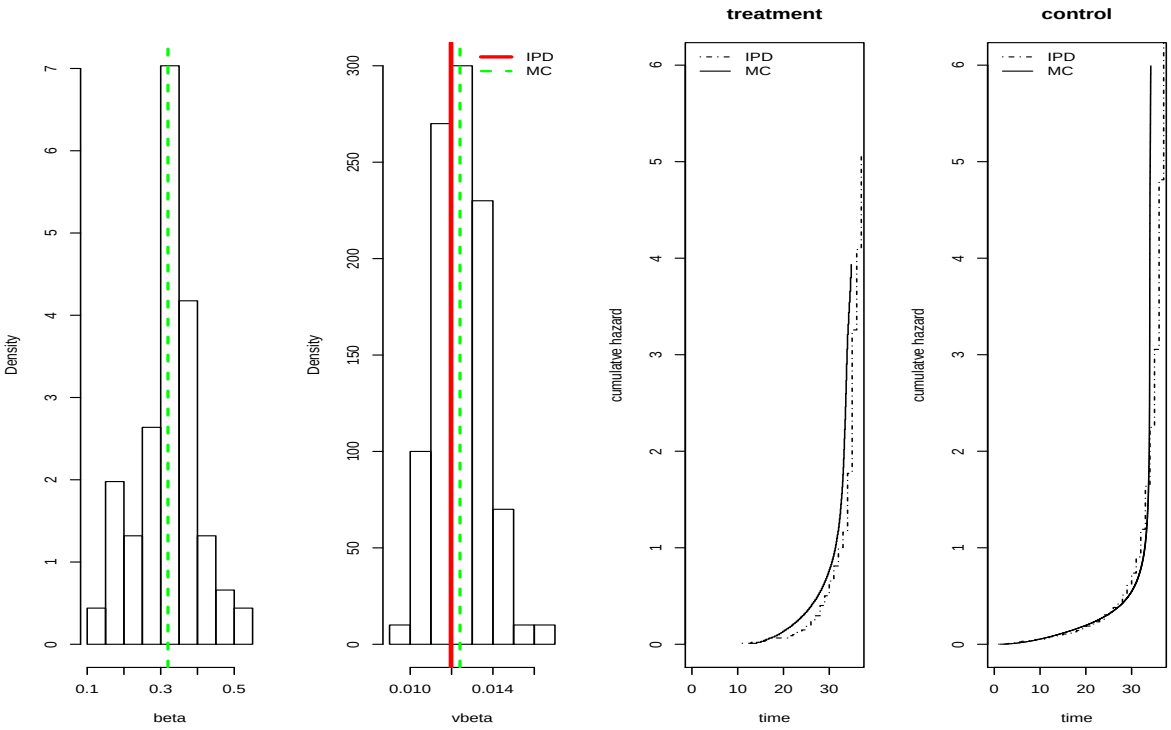

Figure C.2: The `abo2` data: MaxEntBoot samples for the log HR (`beta.`) and respective reciprocal Fisher Information diagonal (`vbeta.`) alongside Nelson-Aalen estimates in group treatment and control. MC = Monte Carlo average. IPD = reference estimate.

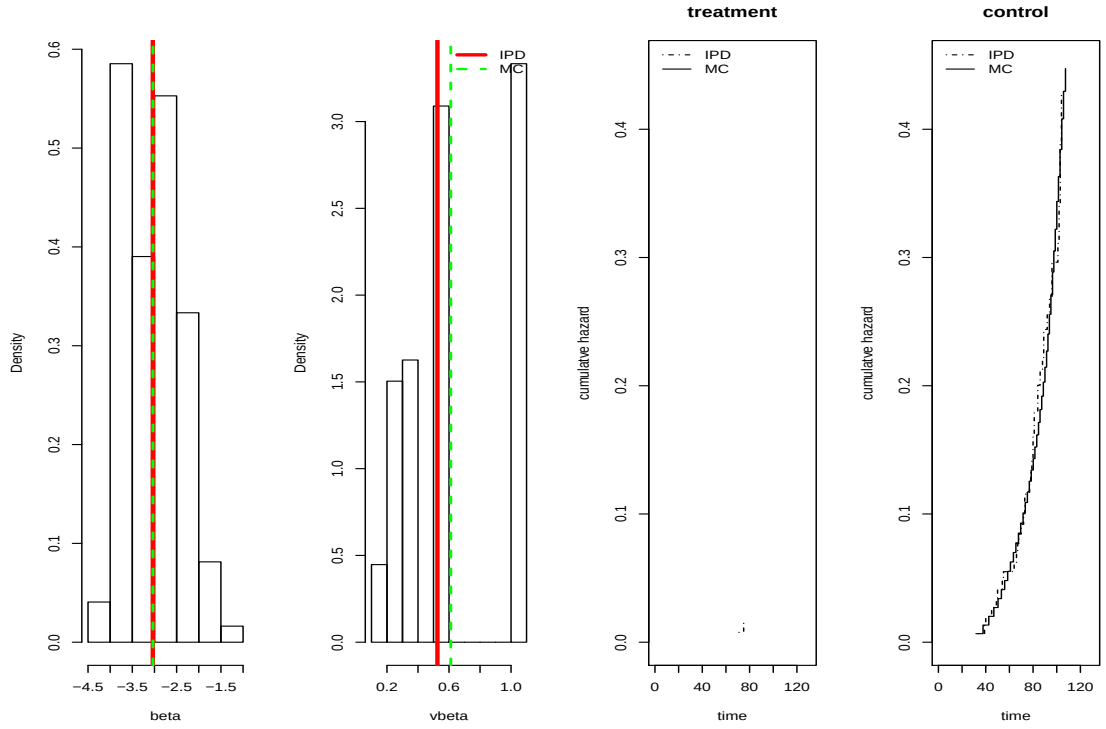

Figure C.3: The `rats.2` data: MaxEntBoot samples for the log HR (`beta.`), and respective reciprocal Fisher Information diagonal (`vbeta.`) alongside Nelson-Aalen estimates in group treatment and control. Note group treatment has originally only two events. MC = Monte Carlo average. IPD = reference estimate.

Appendix D

Programming details

All calculations are performed in R (R Core Team, 2017) and C++. Extensive coding was needed to implement the whole methodology. C++ routines were imported and ran into R via Rcpp and RcppArmadillo packages.

D.1 List of R dependencies

In order to develop our routines we make use of the following R packages:

mvtnormRpack (Genz et al., 2014)

JohnsonDistribution (McLeod and King, 2012)

moments (Komsta and Novomestky, 2015)

doParallel (Analytics and Weston, 2014)

doRNG (Gaujoux, 2014)

Rcpp (Eddelbuettel and François, 2011)

RcppArmadillo (Eddelbuettel and Sanderson, 2014)

maxLik (Henningsen and Toomet, 2011)

microbenchmark (Mersmann, 2015)

cubature (Johnson and Narasimhan, 2013)

boot (Canty and Ripley, 2016)

plyr (Wickham, 2011)

ggplot2 (Wickham, 2009)

D.2 Code extracts

Code excerpt of example from Section 3.3.1, 59. The code generates IPD copies using an incomplete lower triangular (with rank correlation entries) and Gamma real-valued marginals.

```
> m <- 1
> setting.comb.matrix()[, m]
[1] "incomplete" "gamma"      "rank.corr"

> Xspacelist <- Simulate.many.datasets(
+             datalist[2], H = NULL, method = m,
+                SBjohn.correction = T, checkdata = TRUE,
+                  tabulate.similar.data = TRUE )
```

The routine `Simulate.many.datasets` performs the simulating procedure for several different data-sets. Here it acts upon a single data-set (number 2 in the data list). Index `m` defines the re-construction main settings, printed out for clarity. Argument `H` denotes the number B of re-samplings. When set to `NULL` it takes value 300 if the data has fewer than 400 records, and value 100 otherwise. The argument `SBjohn.correction` is needed to fix minor issues in the package `JohnsonDistribution`, as mentioned in Appendix A.1.1, page 79. An extract of the routine output looks like

```
Number of data-set(s): 1
 simulated with following settings
 MC REPLICATES: check by data-set source
 CORRELATION GENERATION: incomplete
 CORRELATION FUNCTION: rank.corr
 MARGINAL DISTRIBUTION: gamma (and bernoulli for binary variables)

Warning messages:
1: In is.data.similar(Xlist, correlation.matrix, moments, corrtype,  :
  some MC third moments differ from reference on average
2: In is.data.similar(Xlist, correlation.matrix, moments, corrtype,  :
  some MC fourth moments differ from reference on average
3: In is.data.similar(Xlist, correlation.matrix, moments, corrtype,  :
  some correlation entries differ from reference on average
```

where MC is generic for Monte Carlo. The routine warns that some simulated moments and correlation entries differ on average from the respective IPD references (see Section 2.2.4, page 38). We should expect such warning because we operate under sub-optimal conditions.

Code excerpt generating HR reconstructions from simulated IPDs from the same example above. Setting `fsm = 1` indicates the first option, ordinary re-sampling, among those described in Section 2.3.3, page 41.

```
> modtype <- c("gaussian", "poisson", "binomial", "cox")
> im <- 4
> fsm <- 1
> fixed.stat.meth.matrix()[ ,fsm]
set.fixed.stat.2.null   expected.fixed.stat
                 TRUE                 FALSE

> RES <- Return.MC.Evidence.and.IPD.estimates(
+       Xspacelist[[1]], fsm, modtype[im], compute.bias = T,
+                     ipd.bootstrap = T)
```

The routine `Return.MC.Evidence.and.IPD.estimates` performs Cox and Breslow estimation for all MC replicates and compare results with those ran on the `diab.2` reference. With the option `ipd.bootstrap = T` we also indicate we want to compare simulated percentiles with those of a IPD bootstrap. All reference IPD bootstraps have size equal to 10000 versus $B = 300$ in our simulation. Once concluded the program outputs the message

```
MC replicates: 300
Model performed: cox regression (and cumul. hazard estimation)

 outcome: status
 covariates: time treat agedx
Handling of fixed stat:
 set.fixed.stat.2.null = TRUE
 expected.fixed.stat = FALSE
```

Below we print code excerpt to reconstruct IPD under the near-optimal settings of Section 3.3.2, 61.

```
> m <- 4
> setting.comb.matrix()[, m]
[1] "norta"       "johnson"     "moment.corr"
```

The NORTA method use moment correlations. An output message of the routine looks like

```
Number of data-set(s): 1
 simulated with following settings
```

```
 MC REPLICATES: check by data-set source
 CORRELATION GENERATION: norta
 CORRELATION FUNCTION: moment.corr
 MARGINAL DISTRIBUTION: johnson (and bernoulli for binary variables)
Stochastic integration for corr. matrix: FALSE
Warning message:
In is.data.similar(Xlist, correlation.matrix, moments, corrtype,  :
  some correlation entries differ from reference on average
```

Here the warning message is caused by the NORTA routine failing to detect a sign of the zero.

D.3 Computations run-time

NORTAmax implementation (Algorithm 2.1.1) bares most computational cost on marginal and dependence structure optimization. For these two steps routines are written in R and run time can vary between few to several seconds. Translation of those routines into a more speedy language is recommended.

Bibliography

Aalen, O. O., Borgan, Ø., and Gjessing, H. K. (2008). *Survival and Event History Analysis. A Process Point of View*. Springer.

Abo-Zaid, G., Sauerbrei, W., and Riley, R. D. (2012). Individual participant data meta-analysis of prognostic factor studies: state of the art? *BMC medical research methodology*, 12(1):1.

Ades, A. and Sutton, A. (2006). Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(1):5–35.

Altman, D. G. (2015). Making research articles fit for purpose: structured reporting of key methods and findings. *Trials*, 16(1):1.

Analytics, R. and Weston, S. (2014). *doParallel: Foreach parallel adaptor for the parallel package*. R package version 1.0.8.

Andersen, P., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics. New York, NY: Springer.

Arends, L. R., Hunink, M. G. M., and Stijnen, T. (2008). Meta-analysis of summary survival curve data. *Statist. Med.*, 27(22).

Baker, S. G. and Kramer, B. S. (2005). Simple maximum likelihood estimates of efficacy in randomized trials and before-and-after studies, with implications for meta-analysis. *Statistical Methods in Medical Research*, 14(4):349–367.

Barron, A. R. (1986). Entropy and the central limit theorem. *The Annals of Probability*, pages 336–342.

Basu, D. (1975). Statistical information and likelihood [with discussion]. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 1–71.

Basu, P. and Templeman, A. (1984). An efficient algorithm to generate maximum entropy distributions. *International journal for numerical methods in engineering*, 20(6):1039–1055.

Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., and Greene, C. S. (2017). Privacy-preserving generative deep neural networks support clinical data sharing. *bioRxiv*.

Becker, B. J. and Wu, M.-J. (2007). The synthesis of regression slopes in meta-analysis. *Statistical Science*, pages 414–429.

Bedford, T. and Wilson, K. J. (2014). On the construction of minimum information bivariate copula families. *Annals of the Institute of Statistical Mathematics*, 66(4):703–723.

Berger, J. O., Wolpert, R. L., Bayarri, M., DeGroot, M., Hill, B. M., Lane, D. A., and LeCam, L. (1988). The likelihood principle. *Lecture notes-Monograph series*, pages iii–199.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306.

Birnbaum, A. (1964). The anomalous concept of statistical evidence: axioms, interpretations, and elementary exposition. Technical Report IMM-NYU-332, New York University Courant Institute of Mathematical Sciences, New York.

Birnbaum, A. (1972). More on concepts of statistical evidence. *Journal of the American Statistical Association*, 67(340):858–861.

Bonofiglio, F., Beyersmann, J., Schumacher, M., Koller, M., and Schwarzer, G. (2015). Meta-analysis for aggregated survival data with competing risks: a parametric approach using cumulative incidence functions. *Research Synthesis Methods*.

Borgan, Ø. and Keogh, R. (2015). Nested case–control studies: should one break the matching? *Lifetime Data Analysis*, 21(4):517–541.

Bowden, J., Tierney, J. F., Simmonds, M., Copas, A. J., and Higgins, J. (2011). Individual patient data meta-analysis of time-to-event outcomes: one-stage versus two-stage approaches for estimating the hazard ratio under a random effects model. *Research synthesis methods*, 2(3):150–162.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Butucea, C., Delmas, J.-F., Dutfoy, A., Fischer, R., et al. (2018). Maximum entropy distribution of order statistics with given marginals. *Bernoulli*, 24(1):115–155.

Cafri, G., Banerjee, S., Sedrakyan, A., Paxton, L., Furnes, O., Graves, S., and Marinac-Dabic, D. (2015). Meta-analysis of survival curve data using distributed health data networks: application to hip arthroplasty studies of the international consortium of orthopaedic registries. *Research Synthesis Methods*, 6(4):347–356.

Cai, T., Gerds, T. A., Zheng, Y., and Chen, J. (2011). Robust prediction of t-year survival with data from multiple studies. *Biometrics*, 67(2):436–444.

Canty, A. and Ripley, B. D. (2016). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-18.

Cario, M. C. and Nelson, B. L. (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Chalmers, I. (1993). The cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Annals of the New York Academy of Sciences*, 703(1):156–165.

Chamandy, N., Muralidharan, O., and Wager, S. (2015). Teaching statistics at google-scale. *The American Statistician*, 69(4):283–291.

Chan, A.-W., Song, F., Vickers, A., Jefferson, T., Dickersin, K., Gøtzsche, P. C., Krumholz, H. M., Ghersi, D., and Van Der Worp, H. B. (2014). Increasing value and reducing waste: addressing inaccessible research. *The Lancet*, 383(9913):257–266.

Chan, A.-W., Tetzlaff, J. M., Gøtzsche, P. C., Altman, D. G., Mann, H., Berlin, J. A., Dickersin, K., Hróbjartsson, A., Schulz, K. F., Parulekar, W. R., et al. (2013). Spirit 2013 explanation and elaboration: guidance for protocols of clinical trials. *Bmj*, 346:e7586.

Chu, B. (2011). Recovering copulas from limited information and an application to asset allocation. *Journal of Banking & Finance*, 35(7):1824–1842.

Clemen, R. T. and Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science*, 45(2):208–224.

Commenges, D. and Hejblum, B. P. (2013). Evidence synthesis through a degradation model applied to myocardial infarction. *Lifetime Data Analysis*, 19(1):1–18.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. John Wiley & Sons, second edition.

Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29(2):357–372.

Crowther, M. J., Riley, R. D., Staessen, J. A., Wang, J., Gueyffier, F., and Lambert, P. C. (2012). Individual patient data meta-analysis of survival data using poisson regression models. *BMC medical research methodology*, 12(1):1.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158.

Csiszár, I. (1984). Sanov property, generalized i-projection and a conditional limit theorem. *The Annals of Probability*, pages 768–793.

Csiszár, I. (1998). The method of types [information theory]. *IEEE Transactions on Information Theory*, 44(6):2505–2523.

Csiszár, I. (2006). A simple proof of sanovs theorem. *Bulletin of the Brazilian Mathematical Society*, 37(4):453–459.

Cuadras, C. (1992). Probability distributions with given multivariate marginals and given dependence structure. *Journal of multivariate analysis*, 42(1):51–66.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer.

Dawid, A. P. (2014). Discussion of on the birnbaum argument for the strong likelihood principle. *Statistical Science*, 29(2):240–241.

de Amo, E., Carrillo, M. D., and Fernández-Sánchez, J. (2012). Characterization of all copulas associated with non-continuous random variables. *Fuzzy Sets and Systems*, 191:103–112.

Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

Debray, T., Moons, K. G., Ahmed, I., Koffijberg, H., and Riley, R. D. (2013a). A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine*, 32(18):3158–3180.

Debray, T. P., Moons, K. G., Abo-Zaid, G. M. A., Koffijberg, H., and Riley, R. D. (2013b). Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? *PLoS One*, 8(4):e60650.

Dembo, A. and Zeitouni, O. (1996). Refinements of the gibbs conditioning principle. *Probability theory and related fields*, 104(1):1–14.

Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. Applications of mathematics. Springer.

Dempster, M. A., Medova, E. A., and Yang, S. W. (2007). Empirical copulas for cdo tranche pricing using relative entropy. *International Journal of Theoretical and Applied Finance*, 10(04):679–701.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177 – 188.

Dias, S., Sutton, A. J., Ades, A., and Welton, N. J. (2013). Evidence synthesis for decision making 2 a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making*, 33(5):607–617.

Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105(492):1347–1357.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201.

Durante, F., Fernández-Sánchez, J., and Sempi, C. (2012). Sklars theorem obtained via regularization techniques. *Nonlinear Analysis: Theory, Methods & Applications*, 75(2):769–774.

Ebrahimi, N., Soofi, E. S., and Soyer, R. (2008). Multivariate maximum entropy identification, transformation, and dependence. *Journal of Multivariate Analysis*, 99(6):1217–1231.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063.

Eddy, D. M., Hasselblad, V., and Shachter, R. (1990). An introduction to a bayesian method for meta-analysis the confidence profile method. *Medical Decision Making*, 10(1):15–23.

Efron, B. (1996). Empirical bayes methods for combining likelihoods. *Journal of the American Statistical Association*, 91(434):538–550.

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *The annals of applied statistics*, 6(4).

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–483.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45(4):302–304.

Fang, H.-B., Fang, K.-T., and Kotz, S. (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82(1):1–16.

Faugeras, O. P. (2013). Sklars theorem derived using probabilistic continuation and two consistency results. *Journal of Multivariate Analysis*, 122:271–277.

Faugeras, O. P. (2015). Maximal coupling of empirical copulas for discrete vectors. *Journal of Multivariate Analysis*, 137:179–186.

Faugeras, O. P. (2017). Inference for copula modeling of discrete data: a cautionary tale and some facts. *Dependence Modeling*, 5(1):121–132.

Fiocco, M., Putter, H., and Van Houwelingen, J. (2009). Meta-analysis of pairs of survival curves under heterogeneity: A poisson correlated gamma-frailty approach. *Statistics in medicine*, 28(30):3782–3797.

Fiocco, M., Stijnen, T., and Putter, H. (2012). Meta-analysis of time-to-event outcomes using a hazard-based approach: Comparison with other models, robustness and meta-regression. *Computational Statistics and Data Analysis*, 56(5):1028 – 1037.

Fisher, C. K., Smith, A. M., and Walsh, J. R. (2018). Boltzmann encoded adversarial machines.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, pages 309–368.

Fraser, D. et al. (2004). Ancillaries and conditional inference. *Statistical Science*, 19(2):333–369.

Gaujoux, R. (2014). *doRNG: Generic Reproducible Parallel Backend for foreach Loops*. R package version 1.6.

Gaye, A., Marcon, Y., Isaeva, J., LaFlamme, P., Turner, A., Jones, E. M., Minion, J., Boyd, A. W., Newby, C. J., Nuotio, M.-L., et al. (2014). Datashield: taking the analysis to the data, not the data to the analysis. *International journal of epidemiology*, 43(6):1929–1944.

Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA*, 37(2):475–515.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2014). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-2.

Ghosh, S. and Henderson, S. G. (2002). Chessboard distributions and random vectors with specified marginals and covariance matrix. *Operations Research*, 50(5):820–834.

Ghosh, S. and Henderson, S. G. (2003). Behavior of the norta method for correlated random vector generation as the dimension increases. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 13(3):276–294.

Goodman, S. N. (1989). Meta-analysis and evidence. *Controlled Clinical Trials*, 10(2):188–204.

Greenland, S., Mansournia, M. A., and Altman, D. G. (2016). Sparse data bias: a problem hiding in plain sight. *bmj*, 352:i1981.

Grimmett, G. R. (1973). A theorem about random fields. *Bulletin of the London Mathematical Society*, 5(1):81–84.

Grünwald, P. (2001). *Strong Entropy Concentration, Game Theory, and Algorithmic Randomness*, pages 320–336. Springer Berlin Heidelberg, Berlin, Heidelberg.

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics*, pages 1367–1433.

Guolo, A. (2012). Higher-order likelihood inference in meta-analysis and meta-regression. *Statistics in Medicine*, 31(4):313–327.

Guolo, A. (2013). Flexibly modeling the baseline risk in meta-analysis. *Statistics in Medicine*, 32(1):40–50.

Guyot, P., Ades, A., Ouwens, M. J., and Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves. *BMC medical research methodology*, 12(1):9.

Henningsen, A. and Toomet, O. (2011). maxlik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3):443–458.

Hiai, F. and Petz, D. (1998). Maximizing free entropy. *Acta Mathematica Hungarica*, 80(4):335–356.

Higgins, J. and Whitehead, A. (1996). Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine*, 15(24):2733–2749.

Hill, I. (1976). Algorithm as 100: Normal-johnson and johnson-normal transformations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 25(2):190–192.

Hill, I., Hill, R., and Holder, R. (1976). Algorithm as 99: Fitting johnson curves by moments. *Journal of the royal statistical society. Series C (Applied statistics)*, 25(2):180–189.

Holly, A., Monfort, A., and Rockinger, M. (2011). Fourth order pseudo maximum likelihood methods. *Journal of Econometrics*, 162(2):278–293.

Hrynaszkiewicz, I., Norton, M. L., Vickers, A. J., and Altman, D. G. (2010). Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers.

Huang, C., Kairouz, P., Chen, X., Sankar, L., and Rajagopal, R. (2017). Context-aware generative adversarial privacy. *Entropy*, 19(12):656.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and De Wolf, P.-P. (2012). *Statistical disclosure control*. John Wiley & Sons.

Iman, R. L. and Conover, W.-J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics-Simulation and Computation*, 11(3):311–334.

Jackson, D., Riley, R., and White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, 30(20):2481–2498.

Jansen, M. J. (1997). Maximum entropy distributions with prescribed marginals and normal score correlations. In *Distributions with Given Marginals and Moment Problems*, pages 87–92. Springer.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4):620.

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952.

Jaynes, E. T. (1996). *Probability theory: the logic of science*. Washington University St. Louis, MO.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1/2):149–176.

Johnson, S. G. and Narasimhan, R. B. (2013). *cubature: Adaptive multivariate integration over hypercubes*. R package version 1.1-2.

Jones, A. P., Riley, R. D., Williamson, P. R., and Whitehead, A. (2009). Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. *Clinical Trials*, 6(1):16–27.

Joshua, S., M., R. G., Beata, N., Chris, D., and Aleksandra, S. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3):663–688.

Kalbfleisch, J. D. (1975). Sufficiency and conditionality. *Biometrika*, 62(2):251–259.

Katsahian, S., Latouche, A., Mary, J.-Y., Chevret, S., and Porcher, R. (2008). Practical methodology of meta-analysis of individual patient data using a survival outcome. *Contemporary Clinical Trials*, 29(2):220–230.

Keogh, R. H., Seaman, S. R., Bartlett, J. W., and Wood, A. M. (2018). Multiple imputation of missing data in nested case-control and case-cohort studies. *Biometrics*.

Kohnen, C. N. and Reiter, J. P. (2009). Multiple imputation for combining confidential data owned by two agencies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(2):511–528.

Komsta, L. and Novomestky, F. (2015). *moments: Moments, cumulants, skewness, kurtosis and related tests*. R package version 0.14.

Kotz, S. and Seeger, J. (1991). A new approach to dependence in multivariate distributions. In *Advances in probability distributions with given marginals*, pages 113–127. Springer.

Larralde, H. (2012). Maximum-entropy distributions of correlated variables with prespecified marginals. *Physical Review E*, 86(6):061117.

Latouche, A., Allignol, A., Beyersmann, J., Labopin, M., and Fine, J. P. (2013). A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of clinical epidemiology*, 66(6):648–653.

Li, S. T. and Hammond, J. L. (1975). Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Transactions on Systems, Man, and Cybernetics*, (5):557–561.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P., Kleijnen, J., and Moher, D. (2009). The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Annals of Internal Medicine*, 151(4):W–65.

Lin, D. and Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, pages 321–332.

Liu, D., Liu, R. Y., and Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association*, 110(509):326–340.

Liu, Z., Rich, B., and Hanley, J. A. (2014). Recovering the raw data behind a non-parametric survival curve. *Systematic reviews*, 3(1):1.

MacKenzie, G. (1994). *Approximately Maximum-entropy Multivariate Distributions with Specified Marginal and Pairwise Correlations*. PhD thesis, University of Oregon.

Madan, J., Chen, Y.-F., Aveyard, P., Wang, D., Yahaya, I., Munafo, M., Bauld, L., and Welton, N. (2014). Synthesis of evidence on heterogeneous interventions with multiple outcomes recorded over multiple follow-up times reported inconsistently: a smoking cessation case-study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(1):295–314.

Manomaiphiboon, K., Park, S.-K., and Russell, A. G. (2008). Accounting for high-order correlations in probabilistic characterization of environmental variables, and evaluation. *Stochastic Environmental Research and Risk Assessment*, 22(2):159–168.

Mansoury, S. and Pasha, E. (2008). Determination of maximum entropy probability distribution via burgs measure of entropy. *Applied Mathematical Sciences*, 2(57):2851–2858.

McLeod, A. and King, L. (2012). *JohnsonDistribution: Johnson Distribution*. R package version 0.24.

Mersmann, O. (2015). *microbenchmark: Accurate Timing Functions*. R package version 1.4-2.1.

Miller, D. J. and Liu, W.-h. (2002). On the recovery of joint distributions from limited information. *Journal of Econometrics*, 107(1):259–274.

Moodie, P. F., Nelson, N. A., and Koch, G. G. (2004). A non-parametric procedure for evaluating treatment effect in the meta-analysis of survival data. *Statistics in medicine*, 23(7):1075–1093.

Mukherjee, D. and Ratnaparkhi, M. V. (1986). On the functional relationship between entropy and variance with related applications. *Communications in Statistics-Theory and Methods*, 15(1):291–311.

Nature (2017). Challenges in irreproducible research.

Nelsen, R. (2006). *An Introduction to Copulas*. Springer Series in Statistics. Springer.

Oertel, F. (2015). An analysis of the rüschendorf transform-with a view towards sklars theorem. *Dependence Modeling*, 3(1).

Oikonomou, K. N. and Grünwald, P. D. (2016). Explicit bounds for entropy concentration under linear constraints. *IEEE Transactions on Information Theory*, 62(3):1206–1230.

Olkin, I. and Sampson, A. (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics*, pages 317–322.

Ormoneit, D. and White, H. (1999). An efficient algorithm to compute maximum entropy densities. *Econometric Reviews*, 18(2):127–140.

O'Rourke, K. (2007). *The Combining of Information: Investigating and Synthesizing What is Possibly Common in Clinical Observations or Studies Via Likelihood*. PhD thesis, University of Oxford.

O'Rourke, K. and Altman, D. G. (2005). Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales by de warn, sg thompson and dj spiegelhalter, statistics in medicine 2002; 21: 1601–1623. *Statistics in Medicine*, 24(17):2733–2742.

ORourke, K. (2001). Meta-analysis: Conceptual issues of addressing apparent failure of individual study replication or inexplicable heterogeneity. In *Empirical Bayes and Likelihood Inference*, pages 161–183. Springer.

ORourke, K., Shea, B., and Wells, G. A. (2001). Meta-analysis of clinical trials. In *Applied Statistics in the Pharmaceutical Industry*, pages 397–424. Springer.

Parmar, M. K. B., Torri, V., and Stewart, L. (1998). Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine*, 17(24):2815–2834.

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance*, 12(3):30–32.

Piantadosi, J., Howlett, P., and Borwein, J. (2012). Copulas with maximum entropy. *Optimization Letters*, 6(1):99–125.

Ponomareva, K., Roman, D., and Date, P. (2015). An algorithm for moment-matching scenario generation with application to financial portfolio optimisation. *European Journal of Operational Research*, 240(3):678–687.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. R version 3.3.3 "Another Canoe".

Raftery, A. E., Givens, G. H., and Zeh, J. E. (1995). Inference from a deterministic population dynamics model for bowhead whales. *Journal of the American Statistical Association*, 90(430):402–416.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of official statistics*, 19(1):1.

Reiter, J. P. (2005). Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21(3):441.

Riley, R. D., Lambert, P. C., and Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*, 340:c221.

Riley, R. D. and Steyerberg, E. W. (2010). Meta-analysis of a binary outcome using individual participant data and aggregate data. *Research Synthesis Methods*, 1(1):2–19.

Robert, C. (1990). An entropy concentration theorem: applications in artificial intelligence and descriptive statistics. *Journal of applied probability*, 27(02):303–313.

Rockinger, M. and Jondeau, E. (2002). Entropy densities with an application to autoregressive conditional skewness and kurtosis. *Journal of Econometrics*, 106(1):119–142.

Rosenkrantz, R. D. (2012). Concentration of distributions at entropy maxima. In *ET Jaynes: Papers on probability, statistics and statistical physics*, volume 158. Springer Science & Business Media.

Royston, P. and Sauerbrei, W. (2008). *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*, volume 777. John Wiley & Sons.

Rüschendorf, L. (1985). Construction of multivariate distributions with given marginals. *Annals of the Institute of Statistical Mathematics*, 37(1):225–233.

Rüschendorf, L. (2009). On the distributional transform, sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139(11):3921–3927.

Salakhutdinov, R. and Larochelle, H. (2010). Efficient learning of deep boltzmann machines. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 693–700.

Sanov, I. N. (1958). On the probability of large deviations of random variables. Technical report, North Carolina State University. Dept. of Statistics.

Schweder, T. and Hjort, N. L. (1996). Bayesian synthesis or likelihood synthesis: What does borel's paradox say? *Report of the International Whaling Commission*, 46:475–480.

Schweizer, B. (1991). Thirty years of copulas. In *Advances in probability distributions with given marginals*, pages 13–50. Springer.

Shao, J. and Tu, D. (1995). *The jackknife and bootstrap*. Springer Science & Business Media.

Siddall, J. and Diab, Y. (1975). The use in probabilistic design of probability curves generated by maximizing the shannon entropy function constrained by moments. *Journal of Engineering for Industry*, 97(3):843–852.

Singer, H. (2010). Maximum entropy inference for mixed continuous-discrete variables. *International Journal of Intelligent Systems*, 25(4):345–364.

Singh, V. P. and Zhang, L. (2018). Copula–entropy theory for multivariate stochastic modeling in water engineering. *Geoscience Letters*, 5(6).

Stewart, G. B., Altman, D. G., Askie, L. M., Duley, L., Simmonds, M. C., and Stewart, L. A. (2012). Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. *PLoS One*, 7(10):e46042.

Stewart, L. A. and Tierney, J. F. (2002). To ipd or not to ipd? advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the health professions*, 25(1):76–97.

Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application*, 2:1–19.

Tagliani, A. (1993). On the application of maximum entropy to the moments problem. *Journal of mathematical physics*, 34(1):326–337.

Templ, M. (2017). *Statistical Disclosure Control for Microdata*. Springer International Publishing AG.

Tiit, E.-M. (2002). Existence of multivariate distributions with given marginals. In *Distributions With Given Marginals and Statistical Modelling*, pages 229–241. Springer.

Van Campenhout, J. and Cover, T. (1981). Maximum entropy and conditional probability. *IEEE Transactions on Information Theory*, 27(4):483–489.

van der Vaart, A. W. (2007). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vasicek, O. A. (1980). A conditional law of large numbers. *The Annals of Probability*, 8(1):142–147.

Vickers, A. J. (2006). Whose data set is it anyway? sharing raw data from randomized trials. *Trials*, 7(1):15.

Vinod, H. D. (2004). Ranking mutual funds using unconventional utility theory and stochastic dominance. *Journal of Empirical Finance*, 11(3):353–377.

Vinod, H. D. (2006). Maximum entropy ensembles for time series inference in economics. *Journal of Asian Economics*, 17(6):955–978.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29.

Williamson, P. R., Smith, C. T., Hutton, J. L., and Marson, A. G. (2002). Aggregate data meta-analysis with time-to-event outcomes. *Statistics in Medicine*, 21(22):3337–3351.

Wu, X. (2003). Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics*, 115(2):347–354.

Xiao, Q. (2014). Generating correlated random vector by johnson system. *Journal of Data Science*, 12(2):217–234.

Yoneoka, D. and Henmi, M. (2016). Synthesis of linear regression coefficients by recovering the within-study covariance matrix from summary statistics. *Research Synthesis Methods*.

Yu, B. et al. (2013). Stability. *Bernoulli*, 19(4):1484–1500.

Zellner, A. and Highfield, R. A. (1988). Calculation of maximum entropy distributions and approximation of marginalposterior distributions. *Journal of Econometrics*, 37(2):195–209.

Zhao, N. and Lin, W. T. (2011). A copula entropy approach to correlation measurement at the country level. *Applied Mathematics and Computation*, 218(2):628–642.

Zografos, K. (1999). On maximum entropy characterization of pearson's type ii and vii multivariate distributions. *Journal of Multivariate Analysis*, 71(1):67–75.

Author Index

Index