A bridge between information theory, saddlepoint approximations, and measure transportation

Andrej Ilievski[1], Davide La Vecchia[2], Elvezio Ronchetti[2]

[1] Lafayette College, Department of Mathematics, 730 High St, Easton, PA 18042, USA
[2] Research Center for Statistics, Geneva School of Economics and Management, University of Geneva, Blv du Pont D'Arve 40, CH-1211, Switzerland

Abstract

We showcase some unexplored connections between information theory, saddlepoint approximations, and measure transportation. To bridge these different areas, we review selectively the fundamental results available in the literature and we draw the connections between them. First, in an abstract setting, we explain how the Esscher's tilting (which is a key result in information theory and lies at the heart of saddlepoint approximations) is connected to the solution to the dual Kantorovich problem (which lies at the heart of measure transportation theory) via the notion of Legendre transform. Then, we turn to statistics and we explain how these connections work for the sample mean and for more general statistics, whose Legendre transform is available in closed-form. Finally, we consider the broad class of M-estimators, whose Legendre transform is seldomly available in closed-form. A discussion on some topics for future research concludes the paper.

Keywords: Change of variable, Kullback-Leibler divergence, Geodesic, Legendre transform, Wasserstein distance.

AMS classification: 28E99, 41A60, 46N10, 49K99, 62F12

1 Introduction

1.1 Overview

A typical problem that arises in statistics is the need for approximating the distribution of some random quantities, depending on $n \in \mathbb{N}$ independent and identical distributed (i.i.d.) random variables.

For instance, assume we are given a sample of size n and we have to derive a confidence interval for the parameter in a one-dimensional location model. To achieve this goal, we need to know the exact distribution of an estimator, such as e.g. the sample mean or another M-estimator, of the location parameter. This distribution is known exactly only in very special settings, such as the case when the estimator is a linear function of underlying Gaussian observations. More often, only approximations to the exact distribution are available.

A common approach to tackle this inferential issue is to define a first-order approximation of the distribution, based on a linearization of the estimator. Then, the behaviour of the linearized statistic is studied, as the sample size diverges to infinity. This leads, through the central limit theorem, to many asymptotic normality proofs, and the resulting first-order asymptotic distribution can be used as an approximation to the exact distribution of the estimator; see e.g. Serfling (2009) for a book-length presentation.

Unfortunately, the accuracy of the first-order (Gaussian) asymptotic approximation deteriorates quickly in small samples. Moreover, the Gaussian approximation tends to be inaccurate in the tails of the distribution even in large samples.

To cope with the low accuracy of the Gaussian approximation, several techniques have been developed and are available to achieve higher-order accuracy. These techniques include the Edgeworth expansions, saddlepoint approximations, the jackknife, and the bootstrap, and provide various corrections to the first-order approximation. For a general survey, we refer to Barndorff-Nielsen and Cox (1989), Young (2009), and Small (2010).

In this paper, we focus on saddlepoint approximations, introduced in the seminal paper of Daniels (1954). Book-length presentations can be found in Field and Ronchetti (1990), Jensen (1995), Kolassa (2006), and Brazzale et al. (2007). These approximations are derived by deforming the contour of the integral defining the inverse Fourier transform of the density of the random variable (i.e. the estimator) of interest. Such a derivation *hinges on complex analysis results*; see e.g. Field and Ronchetti (1990), Ch. 2 for a book-length introduction, or Reid (1988) for a review.

We believe that this technical derivation buries some interesting theoretical aspects which link the saddlepoint theory to some recent advances in mathematics. It is our aim to unveil these connections, bridging the statistical theory of saddlepoint approximations to the mathematical theory of optimal measure transportation, as defined in Monge (1781) and Kantorovich (1942).

Considering the practical problem of finding the optimal way to move given piles of sand to fill up given holes of the same total volume as the sand, Gaspard Monge (1746-1818), one of the pioneers in the area of optimal transportation, initiated a profound theory anticipating different mathematical areas, including differential geometry, linear programming, and nonlinear partial differential equations. Monge's problem remained open until the 1940s, when it was revisited by Leonid Vitaliyevitch Kantorovich (1912-1986; Nobel Prize in Economics in 1975) in relation to the economic problem of optimal allocation of resources. We refer to Villani (2009) for a book-length introduction to the historical background and for some mathematical applications of measure transportation theory (e.g in differential geometry) and to Galichon (2016) for applications in economics (e.g the principal-agent models or the assignment problem of firms to managers). Furthermore, we refer to the same paper for a survey of some recent applications of optimal transport methods in econometrics.

As far as statistics is concerned, the use of measure transportation techniques is becoming more and more popular; see e.g. Chernozhukov et al. (2017) for recent research papers, and Panaretos and Zemel (2019) for a review. For instance, measure transportation theory is related to the so-called Wasserstein distance, which is applied in probability and statistics to derive weak convergence and convergence of moments, and which can be easily bounded to derive concentration inequalities. Moreover, the Wasserstein distance is useful for contrasting complex objects and can be applied to signal processes and engineering; see e.g. Kolouri et al. (2017) for a survey. Recently, Hallin (2017) and Hallin et al. (2019) propose the application of certain measure transportation results in defining multivariate version of ranks and signs, which are suitable for semi-parametric inference for multivariate time series. Moreover, measure transportation theory is also rapidly becoming pivotal for machine learning research. Many data analysis techniques in computer vision, imaging (e.g. for color/texture processing or histograms comparisons), and more general machine learning problems about regression, classification, and generative modeling are often based on the optimal transportation theory; see Peyré and Cuturi (2019) for a book-length discussion.

In spite of this growing body of literature on the statistical applications of measure transportation theory, there is no paper which identifies the close connection between the Monge-Kantorovich results, the information theory, and the theory of saddlepoint approximations. It is our aim to fill that gap.

We review the fundamental results available in the literature, with the purpose of drawing the theoretical connections between them. This effort has its motivation not only in the intrinsic intellectual challenge of relating different branches of mathematics and statistics, but also in the methodological added-value. Indeed, there is a potential transfer of existing knoweledge developed in one field to the other two fields. The selected review presented in this paper and the unveiled connections lay down the theoretical foundation to trigger that transfer. With this spirit, in §5 we itemize some topics for future research: the aim is to generate new results, using this paper as a step stone.

Finally, we believe that the unveiled connections can have a direct impact on some areas of statistical education. They offer novel approaches to introduce higher-order techniques for asymptotic analysis via saddlepoint approximations, giving the chance of looking at statistical and mathematical problems from different angles.

1.2 Structure of the paper

We consider the derivation of the saddlepoint approximation via the method of the conjugate density, which *hinges on convex analysis results*.

The key tool of our development is the Legendre transform, which allows us to shed light on the connections between saddlepoint theory, information theory, and optimal transportation. We first consider an abstract setting in §2 and we work with a generic random variable. We set up the basic notions related to information theory, convex analysis, and measure transportation. In §2.1, we recall the method of the conjugate density, the notion of Legendre transform, and its link to information theory. In §2.2.1, we briefly state the optimal transport problem, introducing the concept of optimal transportation mapping, and highlight the relation between the solution to the optimal transport problem and the notion of the saddlepoint. In §2.2.2 and §2.3, we explain how the saddlepoint and the Wasserstein distance (i.e. the saddlepoint and the optimal transportation mapping) are related. In §3 we turn to the statistical framework and explain how the links discussed in the abstract setting of §2 work for well-known statistics whose cumulant generating function is available in closed-form. Following the seminal paper Daniels (1954), we start our journey from the sample mean and we provide some graphical

illustrations of the key concepts (measure transportation and relative error). Then, in §4, we extend the connections to M-estimators, which are general tools for conducting parametric inference.

The relations discussed in the paper lie at the intersection of several disciplines. The connections discussed here illustrate only a few of the several unexplored links between measure transportation, information theory, and statistics. Some other statistical points remain open and researchers in other fields will have the possibility to explore them, using our paper as a step stone. In §5 we mention several possible future research directions.

2 Basic mathematical aspects

2.1 Esscher's tilting, saddlepoint, and conjugate density

Let us first recall the main points of the method of the conjugate density. Let $X \sim F_X$, where $dF_X(x) = f_X(x)dx$, or equivalently, let us assume that X has a measure μ (which is absolutely continuous w.r.t. the Lebesgue measure), whose support is $\mathcal{X} \subseteq \mathbb{R}$. Then, given $t \in \mathbb{R}$, we define the conjugate density,

$$h_t(x) = C(t) \exp\{\upsilon(t)(x - t)\} f_X(x), \tag{1}$$

where $\upsilon(t)$ is chosen such that $E_{h_t}[X] = t$, E_{h_t} represents the expected value taken w.r.t. h_t. Eq. (1) defines an embedding of f_X into an exponential family. The function $C : \mathbb{R} \to \mathbb{R}$ is defined as

$$C(t) = \exp\{\upsilon(t)t - K_X[\upsilon(t)]\},$$

with $K_X(\upsilon) = \log E_{f_X}[\exp\{\upsilon X\}]$ representing the cumulant generating function (c.g.f.) of X and $C(t)$ is such that h_t is a density—namely, it integrates to one. In fact, $\upsilon(t)$ is the saddlepoint at t, obtained by solving the equation $K'_X(\upsilon) = t$; see among the others Field and Ronchetti (1990) p. 34-35.

The conjugate density highlights a connection between the saddlepoint solution and information theory. Indeed, (1) illustrates that $\upsilon(t)$ defines a transformation of the original density f_X into its conjugate density h_t. Theorem 2.1 in Kullback (1997) shows that h_t is the solution to the following information theoretic problem:

$$\min_{g \in \mathcal{G}} \mathrm{KL}(g, f_X), \quad \text{s.t.} \quad g(x) \geq 0, \quad \int_{\mathcal{X}} g(x)dx = 1, \quad E_g[X] = t \ , \tag{2}$$

where $\mathcal{G}$ contains all the densities having support $\mathcal{X}$ and finite second moment, while KL represents the backward Kullback-Leibler divergence

$$\mathrm{KL}(g, f_X) = \int_{\mathcal{X}} g(x) \log \frac{g(x)}{f_X(x)} dx.$$

We refer also to Kremer (1982) p. 59 and to Esscher (1932) for a discussion on this result. For completeness the proof is provided in the Appendix.

The transformation $f_X \mapsto h_t$, commonly called Esscher transformation (Esscher (1932) and Kremer (1982)) or exponential tilting (Barndorff-Nielsen and Cox (1989), p. 105), is related to the Legendre transform (henceforth, denoted by the symbol *) of K_X, defined as

$$K_X^*(t) = \sup_{\upsilon} \{\upsilon t - K_X(\upsilon)\}, \tag{3}$$

where the maximising value $\upsilon(t)$ is the saddlepoint, namely the solution to

$$K_X'(\upsilon) = t; \tag{4}$$

see McCullagh (2018), Ch. 6. The function $-K_X^*(t)$ is called the (point) entropy of the density f_X. In the literature on convex analysis, the term convex conjugate is also applied; see Rockafellar (2015), Section 12.

We notice that

$$K_X^*(t) = \upsilon(t)t - K_X(\upsilon(t)) = \log[\exp\{\upsilon(t)t - K_X(\upsilon(t)\}] = \log C(t).$$

By the well-known relation between the derivatives of a convex function and the derivatives of its Legendre transform, we have $K_X^{*\prime}(t) = 1/K_X'(\upsilon(t))$. Moreover, exploiting the strict convexity of the c.g.f. of X, the following identity holds:

$$K_X^{**}(t) = \sup_{\upsilon} \{\upsilon t - K_X^*(\upsilon)\} = K_X(\upsilon);$$

see e.g. McCullagh (2018), p. 179.

2.2 Measure transportation

2.2.1 Kantorovich dual problem

The conjugate density method focuses on the transformation of the p.d.f. f_X into the new p.d.f. h_t. In fact, this transformation has in background a change of variables, or equivalently, a random variable transformation

$\mathcal{T} : \mathcal{X} \to \mathcal{X}$, which yields $X \mapsto Y$. Specifically, we can think of the conjugate density method as if we start from $X \sim \mu$, having density f, and obtain $Y = \mathcal{T}(X)$ with $Y \sim \nu_t$, where the measure ν_t is absolutely continuous w.r.t. the Lebesgue measure, it has support $\mathcal{X}$ and a c.d.f. H_t such that $dH_t(y) = h_t(y)dy$.

From this perspective, the saddlepoint induces a measure transport of μ onto ν_t. Therefore, using the standard notation of optimal transportation theory, we say that $\mathcal{T}_{\#}\mu = \nu_t$—to be read as $\mathcal{T}$ pushes μ forward to ν_t. Specifically, if μ is a Borel measure on $\mathcal{X}$ and $\mathcal{T}$ is a Borel map $\mathcal{X} \to \mathcal{X}$, then $\mathcal{T}_{\#}\mu = \nu_t$ stands for the image measure (or push-forward) of μ by $\mathcal{T}$: this is a Borel measure on $\mathcal{X}$, defined by $(\mathcal{T}_{\#}\mu)[A] = \mu[\mathcal{T}^{-1}(A)]$ for any Borelian A; see Villani (2009). The map $\mathcal{T}$ appearing in all these statements is called the transportation mapping: informally, one can say that $\mathcal{T}$ transports the mass represented by the measure μ, to the mass represented by the measure ν_t.

The transportation map which minimizes a given transportation cost function is called an optimal transportation mapping. To elaborate further, let us consider the following formulation of the optimal transportation problem. Let μ and ν_t belong to the family $\mathcal{G}$, and let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Borel-measurable function representing the cost of transporting X to Y. The objective is to find a measurable (transportation) mapping $\mathcal{T}$ solving the minimization problem (see Kantorovich (1942))

$$\mathrm{KP}(\mu, \nu_t) = \inf_{\gamma \in \Gamma(\mu, \nu_t)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\gamma(x, y), \tag{5}$$

where the infimum is over all pairs (X, Y) of (μ, ν_t), belonging to $\Gamma(\mu, \nu_t)$, the set of probability measures γ on $\mathcal{X} \times \mathcal{X}$, satisfying $\gamma(A \times \mathcal{X}) = \mu(A)$ and $\gamma(\mathcal{X} \times B) = \nu_t(B)$, for Borel sets $A, B \subset \mathcal{X}$. In statistical applications of the optimal transport theory, a typical choice for $c(x, y)$ in (5) is the quadratic transport cost $c(x, y) = (1/2)|x - y|^2$; see e.g. Chernozhukov et al. (2017), Hallin (2017), and Panaretos and Zemel (2018).

Now, we recall from §2.1 that the saddlepoint $\upsilon(t)$ defines the conjugate density h_t, which represents the density of the transported measure ν_t and, at the same time, is the density that minimizes $\mathrm{KL}(g, f_X)$ as in (2). Then, a question naturally arises: *"Is the saddlepoint $\upsilon(t)$ related to an optimal transportation problem as in (5) for some choice of $c(x, y)$?"*

To answer this question, let us consider Kantorvich's dual formulation to the primal optimization problem (5). For a cost function c, the dual

problem reads as

$$\begin{aligned} \mathrm{KD}(\mu,\nu_t) &= \sup_{\phi,\varphi}\left(\int_{\mathcal{X}}\varphi(y)d\nu_t(y)-\int_{\mathcal{X}}\phi(x)d\mu(x)\right) \\ &\text{s.t.} \quad \varphi(y)-\phi(x)\leq c(x,y), \quad \forall(x,y). \end{aligned} \tag{6}$$

Since the primal problem is convex, we have $\mathrm{KP}(\mu,\nu_t)=\mathrm{KD}(\mu,\nu_t)$ and the solution to $\mathrm{KD}(\mu,\nu_t)$ is given by the pair:

$$\begin{aligned} \varphi(y) &= \inf_x[\phi(x)+c(x,y)] \\ \phi(x) &= \sup_y[\varphi(y)-c(x,y)]. \end{aligned} \tag{7}$$

The equations in (7) imply that the functions ϕ and φ are related to each other through the so-called c-transform, whose form depends on the cost function c; see Villani (2009), p. 55, for the definition and more mathematical detail. Setting $c(x,y)=-xy$, the c-transform coincides with the *Legendre transform* and setting $\varphi\equiv -K_X$, we have that the second equation of (7) becomes (after an elementary change of variable and algebraic manipulations) $\phi(t)=\sup_v[vt-K_X(v)]$, which is the Legendre transform of K_X as discussed in §2.1. Therefore, using (3), we conclude that $\phi(t)\equiv K_X^*(t)$, which implies that the solution to $\mathrm{KD}(\mu,\nu_t)$, when setting $\varphi\equiv -K_X$, is obtained for $\phi\equiv K_X^*$, the Legendre transform of K_X. Due to the strict convexity of K_X, we also have $K_X^{**}=K_X=-\varphi$. Moreover, Theorem 5.10 and Remark 5.13 in Villani (2009) imply that the transportation map induced by the saddlepoint through (the gradient of) K_X is optimal, given that, by the definition of the saddlepoint $v(t)$, the pair (K_X,K_X^*) satisfies $K_X(v(t))+K_X^*(t)=v(t)t$, for all $(v(t),t)$; see Villani (2009), p. 59.

2.2.2 Wasserstein distance, Wasserstein spaces, and geodesics

The solution to $\mathrm{KP}(\mu,\nu_t)$ with $c(x,y)=(1/2)(x-y)^2$ defines the squared Wasserstein distance

$$W_2^2(\mu,\nu_t)=\inf_{\gamma\in\Gamma(\mu,\nu_t)}\int_{\mathcal{X}\times\mathcal{X}}(1/2)(x-y)^2d\gamma(x,y),$$

where the infimum is taken over all pairs (X,Y) of (μ,ν_t), belonging to $\Gamma(\mu,\nu_t)$. As noticed by Villani (2009), p. 55-56, solving the primal problem $\mathrm{KP}(\mu,\nu_t)$ with $c(x,y)=(1/2)(x-y)^2$ is equivalent to solving the same problem with $c(x,y)=-xy$. To see this, let us consider the elementary equality

$(1/2)(x-y)^2 = (1/2)(x^2+y^2-2xy)$, which implies that the interaction between x and y in the quadratic cost function is the same as in $c(x,y) = -xy$ (recall that in $\mathrm{KP}(\mu,\nu_t)$ the marginal of X and Y are fixed). As a result, solving the problem $\mathrm{KP}(\mu,\nu_t)$ is equivalent to finding the $\sup_\gamma E_\gamma(XY)$, where the supremum is over all coupling (X,Y) of (μ,ν_t), so the problem is to maximize the correlation between the random variables X and Y. From (7), we know that the solution to $\mathrm{KP}(\mu,\nu_t)$ with $c(xy) = -xy$ is related to the Legendre transform of K_X, which is further related to the saddlepoint $\upsilon(t)$, which in turn defines the solution of the information theoretic problem in (2). Thus, we conclude that the saddlepoint $\upsilon(t)$ is related to the optimal mass transportation in the sense of both $W_2^2(\mu,\nu_t)$ and backward Kullback-Leibler divergence. The latter consideration give insights into the geometric properties of the saddlepoint—an aspect almost completely ignored in the literature on saddlepoint approximations.

Some further considerations are in order. Let $\mathcal{X}$ be a convex subset of $\mathbb{R}$ and let us denote by $\mathcal{P}(\mathcal{X})$ the set of all probability measures defined on $\mathcal{X}$ and admit finite second moment. The resulting 2-Wasserstein space $(\mathcal{P}(\mathcal{X}), W_2)$ is a metric space with a geometric structure: for $\mu,\nu_t \in \mathcal{P}(\mathcal{X})$ there exists a continuous path going from μ to ν_t, such that its length is the distance between the two measures. Indeed, $(\mathcal{P}(\mathcal{X}), W_2)$ inherits the geometric properties of the ground space $(\mathcal{X}, \|x-y\|)$, such as being a geodesic space. We recall that a geodesic space refers to a metric space in which every pair of points $x, y \in \mathcal{X}$ is connected by a continuous curve $s \in [0,1] \to x(s) \in \mathcal{X}$ which satisfies

$$\|x - x(s)\| = s\|x-y\| \quad \text{and} \quad \|x(s)-y\| = (1-s)\|y-x\|,$$

Such a curve is called a geodesic; see e.g. McCann and Guillen (2011) for a survey in the context of optimal transportation.

In the space $(\mathcal{P}(\mathcal{X}), W_2)$, the geodesics are easily characterized and they are given by the so-called displacement interpolation (a.k.a. McCann interpolation); see Villani (2009) Ch. 7. Specifically, the geodesic in $(\mathcal{P}(\mathcal{X}), W_2)$ is obtained exploiting the geodesic properties of $(\mathcal{X}, \|x-y\|)$: for $s \in [0,1]$ and the optimal transport $\mathcal{T}$, we set $\mathcal{T}_s(x) = (1-s)x + s\mathcal{T}(x)$—uniqueness of $\mathcal{T}$ implies uniqueness of the geodesic. We interpret $\mathcal{T}_s$ as the position at time s of the mass initially at x. We remark that $\mathcal{T}_0 \equiv Id$ (the identity function), while $\mathcal{T}_1 \equiv \mathcal{T}$, where the latter is defined using the saddlepoint $\upsilon(t)$. In the 2-Wasserstein space, the geodesic from μ to ν_t is defined as

$$\mu_s = \mathcal{T}_s \# \mu, \tag{8}$$

which indicates that the (shape of the) geodesic depends on the optimal transport, which is related to the Legendre transform. The velocity of each particle is $\partial_s \mathcal{T}_s(x) = \mathcal{T}(x) - x$, while its acceleration is $\partial_s^2 \mathcal{T}_s(x) \equiv 0$. This implies that the mass of μ is transported to the mass of ν_t at a constant speed, along the geodesic μ_s.

2.3 Optimal transformation, Jacobian formula, and saddlepoint equation

Finally, it remains to understand the connection between the optimal transformation $\mathcal{T}$ and the saddlepoint technique. To this end, we remark that the mapping $\mathcal{T}$ associated with the quadratic cost function (hence also with $c(x, y) = -xy$, see §2.2.2) is $\mathcal{T} = H_t^{-1} \circ F_X$, where H_t^{-1} is the (generalized) inverse of the conjugate c.d.f. H_t. Therefore,

$$Y = \mathcal{T}(X) = H_t^{-1} \circ F_X(X); \tag{9}$$

see e.g. Panaretos and Zemel (2018) and reference therein for a survey on this result in the statistical framework. With this regard, Theorem 10.28 in Villani (2009) (formula (10.20)) implies that $\mathcal{T}$ is the gradient of a convex function; see also Galichon (2017) (p. C5) and Galichon (2016) Ch. 4. Eq. (9) illustrates that, in the case of the saddlepoint method, this convex function is related to K_X, whose gradient defines the saddlepoint equation in (4). We conclude that the saddlepoint $\upsilon(t)$ induces the optimal mapping $\mathcal{T}$, which is a deterministic coupling–see Villani (2009) p. 6–yielding $X \mapsto Y$.

Furthermore, (9) illustrates that $\mathcal{T}$ is a change of variable from μ to ν_t. Thus, for all ν_t-integrable functions ω,

$$\int_{\mathcal{X}} \omega(y) d\nu_t(y) = \int_{\mathcal{X}} \omega(\mathcal{T}(x)) d\mu(x).$$

In that sense, as remarked in Villani (2009) Ch. 11, optimal transport essentially depicts a change of variables. This aspect is typically overlooked in the literature on saddlepoint approximations, whose focus is mainly on the derivation of the order of the approximation. In contrast, our connection to measure transportation theory highlights the change of variable that is induced by the saddlepoint, opening the door for further links between these areas.

Indeed, in probability and statistics, it is customary to write the Jacobian equation for problems involving changes of variables. In the univariate case,

the Jacobian formula implies that the density of Y satisfies:

$$f_Y(y) = f_X\left[\mathcal{T}^{-1}(y)\right]\left|\frac{\partial \mathcal{T}^{-1}(y)}{\partial y}\right| = h_t(y). \tag{10}$$

where the term $\left|\partial \mathcal{T}^{-1}(y)/\partial y\right|$ is the absolute value of the Jacobian. Eq. (10) illustrates that, once the optimal transportation mapping $\mathcal{T}$ is known, the conjugate density is obtained by a standard application of Jacobian formula, computing $\mathcal{T}^{-1}$ and its derivative.

2.4 Example: exponential random variable

The arguments from §2.1 to §2.3 are better understood via the following simple example. Let us consider a random variable having an exponential p.d.f. with rate one, namely $X \sim \exp(1)$ so $X \sim \mu$, and define a target variable Y, such that $E[Y] = t$ and $Y \sim \nu_t$.

Conjugate density method. The c.g.f of X is

$$K_X(\upsilon) = \log(1/(1-\upsilon)),$$

which is defined for $0 < \upsilon < 1$, and solving

$$K'_X(\upsilon) = t$$

yields the saddlepoint $\upsilon(t) = 1 - 1/t$. From (1) it follows that the conjugate density is

$$h_t(y) = C(t)\exp\{\upsilon(t)(y-t) - y\},$$

where $C(t) = \exp\{t - 1 - \log t\} = \exp\{K^*_X(t)\}$. Simple algebraic manipulations yield

$$h_t(y) = (1/t)\exp\{-y/t\},$$

namely $Y \sim \exp(1/t)$. The c.d.f. H_t is related to the measure ν_t.

Optimal transportation. Now, let us verify that the result obtained via the conjugate density is coherent with the solution to the optimal transportation problem. To this end, we notice that $H_t(x) = 1 - \exp\{-x/t\}$ so $H_t^{-1}(u) = -t\log(1-u)$, for $u \in (0,1)$. As remarked in §2.3, the optimal transportation mapping is as in (9), namely

$$Y = \mathcal{T}(X) = H_t^{-1} \circ F_X(X) = tX,$$

which illustrates that $\mathcal{T}$ is related to the gradient of the convex function K_X. The Jacobian formula yields that the p.d.f. of Y is

$$f_Y(y) = (1/t)\exp\{-y/t\}.$$

Hence, $f_Y(y) \equiv h_t(y)$, in agreement with (10).

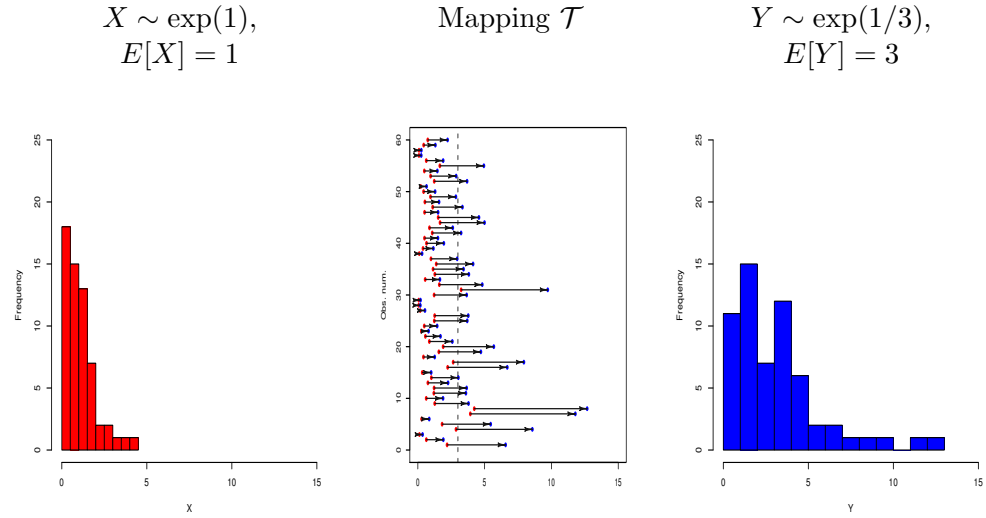

Figure 1: Optimal transportation problem. Left panel, histogram of 60 random observations drawn from an exponential with rate one (related to the measure μ). Middle panel: the optimal transportation map $\mathcal{T}(x) = tx$, that deforms μ into ν_t, for $t = 3$, where each arrow indicates the source and destination of the transported mass. The vertical dotted line at 3 represents the mean of the target Y. Right panel: histogram of 60 random observations drawn from an exponential with rate 1/3 (related to the measure ν_t with $t = 3$).

Graphical interpretation. In Figure 1, we provide a graphical illustration of the optimal transportation map related to the saddlepoint. In the left panel, we display the histogram of an observed sample $\{x_i\}_{i=1}^{60}$ drawn from an exponential with rate one (related to the measure μ). The middle panel illustrates the optimal map $y = \mathcal{T}(x) = 3x$, that deforms the original measure μ into ν_t, for $t = 3$, i.e., $\mathcal{T}_{\#}\mu = \nu_3$. The map is plotted in the form of vectors acting on the observed data, where each arrow indicates the source and destination of the mass being transported for each x_i. Reversing the direction of the arrows would produce the inverse map $\mathcal{T}_{\#}^{-1}\nu_3 = \mu$,

optimally deforming the measure ν_3 onto μ. The arrows illustrate that the map $\mathcal{T}$ acts more on those x_i's which are between the origin and the vertical dotted line, which represents the mean of the target random variable Y. In the right panel, we display the histogram of the sample $\{y_i\}_{i=1}^{60}$ as obtained applying $\mathcal{T}$ to $\{x_i\}_{i=1}^{60}$. Comparing the height of the histogram bars on the left and right plots, we realise the effect of the mass transportation induced by $\mathcal{T}$ and related to the saddlepoint.

3 The connections for some simple statistics

How are the connections unveiled in §2 related to statistics? In this section we answer this question, starting from the inferential problem of deriving an approximation to the sampling distribution of the sample mean. Then, we consider the case of a general statistics with a known c.g.f.

We start with the mean because it is a simple (linear) statistic. This was the case considered in the seminal paper by Daniels (1954), who introduced the saddlepoint density approximation of the mean of n i.i.d. random variables. In §3.1 we consider this case and we explain how the connections in §2 are related to Daniels' saddlepoint approximation. Then, in §3.2, we explain briefly how the same arguments extend to the case of general statistics.

In the remaining of this section, we let $X \sim f$ (for the ease-of-notation, we use f rather than f_X to denote the density of X), with f related to a measure μ which is absolutely continuous w.r.t. the Lebesgue measure. Assume we are given a random sample $X_1, ..., X_n$ of i.i.d. copies of X. Assume that the c.g.f. of X is well defined and call it K_X.

3.1 The sample mean

3.1.1 Saddlepoint approximation for the mean

Let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ be the mean of $X_1, ..., X_n$ and let us denote its density by f_n. Furthermore, let $K_{\bar{X}_n}(\upsilon) = \log E_{f_n}[\exp\{\upsilon \bar{X}_n\}]$ be the c.g.f. of $\bar{X}_n$. By standard Fourier inversion, the density f_n is obtained as

$$f_n(t) = \frac{n}{2\pi i} \int_{-i-\infty}^{i+\infty} \exp\{n(K_X(\upsilon) - \upsilon t)\} d\upsilon. \tag{11}$$

The integral in (11) is typically not available in a closed-form. However, an approximation to f_n can be obtained by applying the method of steepest descent to the integral in (11), followed by the Cauchy's theorem (to deform the path of the integral in the complex domain) and Watson's lemma (to

control the error of the approximation) accordingly; see Daniels (1954), p. 633 for the mathematical detail and Small (2010), Ch. 7, for book-length presentations. A review is available in Reid (1988).

The resulting approximation to f_n is the saddlepoint density approximation f_{sad}, namely

$$f_{\text{sad}}(t) \quad = \quad \left[\frac{n}{2\pi K_X''(\upsilon(t))}\right]^{1/2} \exp\{n[K_X(\upsilon(t)) - \upsilon(t)t]\}, \tag{12}$$

where $K_X''(\upsilon(t))$ is the second derivative of K_X computed at $\upsilon(t)$, the saddlepoint solution to the equation $K_X'(\upsilon) = t$. The approximation f_{sad} is such that

$$f_n(t) = f_{\text{sad}}(t)\{1 + O(n^{-1})\}. \tag{13}$$

See also Barndorff-Nielsen and Cox (1979) and Goutis and Casella (1999) for a review.

A few remarks are in order. The use of f_{sad} in (13) yields some well-known advantages over other routinely applied approximations. For instance, the Gaussian approximation is first-order accurate, with an absolute error of order $O(n^{-1/2})$, and it performs poorly in the tails. The combination of these two aspects entails large approximation errors in small samples and/or for large values of t in moderate samples. The Edgeworth expansion features higher-order accuracy, with an absolute error of order $o(n^{-1/2})$ or $O(n^{-1})$, and it can perform well in small samples. However, using the first (two or three) terms of the expansion usually provides good approximation in the center of the density, but it can be inaccurate in the tails, where the approximation can even become negative. The saddlepoint approximation f_{sad} overcomes these problems. By construction, f_{sad} is a density-like object which cannot become negative and it keeps its accuracy in the tails, providing accurate small sample approximations. With this regard, notice that the error in (13) is of order $O(n^{-1})$ and it is of relative type—to be contrasted with the absolute error entailed by the asymptotic theory and by the Edgeworth expansion. Moreover, Daniels (1954) proves that the size of the error holds uniformly in $t \in \mathbb{R}$. All these features are peculiar of f_{sad}.

3.1.2 Connection to measure transportation

Daniels (1954) provided an alternative derivation of f_{sad} using the method of the conjugate density and this establishes the connection of this construction with measure transportation theory.

The basic idea of the conjugate density method is to recenter the density of $\bar{X}_n$ at the point of interest t and use a normal approximation in the recentered problem which leads to the approximation to $f_n(t)$. To perform this construction, we first embed the original density f into an exponential family, and then look for the (conjugate) density h_t which is the closest to f in KL distance and has a mean of t. Then, we compute an Edgeworth expansion (see Bhattacharya and Rao (1986)) for the tilted density to obtain $f_{\text{sad}}(t)$. Finally, we repeat this procedure for every $t \in \mathbb{R}$. See also Reid (1988), p. 215, for mathematical details.

We now explain how this approach clarifies the connections between the saddlepoint approximation and the measure transportation. To this end, we proceed in three steps: (i) we make explicit the link between the saddlepoint and the recentering of $\bar{X}_n$ at t, pointing to the results in §2.1; (ii) we depict the link between the recentering and the Kantorovich dual problem, highlighting the role of the Legendre transform of the c.g.f. of $\bar{X}_n$ and using the analysis performed in §2.2.1; (iii) we rewrite $f_{\text{sad}}(t)$ in terms of the Legendre transform of $K_{\bar{X}_n}$.

(i) Eq. (1) gives the expression of the conjugate density h_t, which is defined by means of the saddlepoint $\upsilon(t)$, the solution to $K'_X(\upsilon) = t$. This implies that, by construction, h_t is such that $E_{h_t}[X] = t$ and we have

$$K'_X(\upsilon) = t \iff E_{h_t}[X] = t. \tag{14}$$

Moreover, simple algebra shows that

$$\begin{aligned} K''_X(\upsilon(t)) &= \frac{1}{E_f[e^{\upsilon(t)(X-t)}]} \int_{\mathbb{R}} (x-t)^2 e^{\upsilon(x-t)} f(x) dx \\ &= \int_{\mathbb{R}} (x-t)^2 h_t(x) dx \\ &= E_{h_t}(X^2) - [E_{h_t}(X)]^2 = var_{h_t}[X] =: \sigma^2(t). \end{aligned} \tag{15}$$

The statistic $\bar{X}_n$ is a linear combination of X_is and its c.g.f. is

$$K_{\bar{X}_n}(\upsilon) = \log E_{f_n}[\exp\{\upsilon \bar{X}_n\}] = nK_X(\upsilon/n). \tag{16}$$

Thus

$$K'_{\bar{X}_n}(\upsilon(t)) = t \iff E_{h_t}[\bar{X}_n] = t, \tag{17}$$

which illustrates that the saddlepoint $\upsilon(t)$ performs a re-centering of the measure of $\bar{X}_n, \mu^{(n)}$, at the point of interest t.

(ii) The combination of (14) with (17) yields the following implication:

$$K'_X(\upsilon(t)) = t \iff K'_{\bar{X}_n}(\upsilon(t)) = t. \tag{18}$$

The implications in (18) make explicit the measure transportation induced by the saddlepoint $\upsilon(t)$: the measure $\mu^{(n)}$ is transported into a new measure, say $\nu_t^{(n)}$, with mean t. To perform this measure transportation, (18) illustrates that we may fully exploit the linearity of $\bar{X}_n$ in the X_is; in particular, we make use of the fact that $E_{h_t}[X] = t$ implies $E_{h_t}[\bar{X}_n] = t$. Thus, to move $\mu^{(n)}$ into $\nu_t^{(n)}$, we write the Kantorovich dual problem directly in terms of μ and ν_t, obtaining the KD(μ, ν_t) as in (6) with cost function $c(x, y) = -xy$. In this formulation, each $X_i \sim \mu$ and the measure μ is transported to the target measure ν_t, which has density h_t. As discussed in §2.2.1, the solution to KD(μ, ν_t) is obtained via the Legendre transform $K_X^*(t) = \sup_\upsilon\{\upsilon t - K_X(\upsilon)\}$, which defines the saddlepoint $\upsilon(t)$, which in turn yields h_t via (1).

(iii) We now focus on the saddlepoint density approximation f_{sad}. The link between K_X and $K_{\bar{X}_n}$ in (16) induces the same link between their Legendre transforms, namely

$$\begin{aligned} K_X^*(t) &= \sup_\upsilon\{\upsilon t - K_X(\upsilon)\} = n \sup_\upsilon\{\upsilon t/n - K_X(\upsilon/n)\} \\ &= n \sup_\upsilon\{\upsilon t/n - K_{\bar{X}_n}(\upsilon)\} = nK_{\bar{X}_n}^*(t/n). \end{aligned} \tag{19}$$

Thus from (15) it follows the saddlepoint density approximation to f_n at t becomes

$$\begin{aligned} f_{\text{sad}}(t) &= \left(\frac{n}{2\pi\sigma^2(t)}\right)^{1/2} \exp\{n[K_X(\upsilon(t)) - \upsilon(t)t]\} \\ &= (2\pi)^{1/2} \left[K_X^{*\prime\prime}(t)\right]^{1/2} \exp\{-nK_X^*(t)\} \\ &= (2\pi)^{1/2} \left[n^{-1}K_{\bar{X}_n}^{*\prime\prime}(t/n)\right]^{1/2} \exp\{-n^2 K_{\bar{X}_n}^*(t/n)\}, \end{aligned} \tag{20}$$

where we make use of $C(t) = \exp\{\upsilon(t)t - K_X[\upsilon(t)]\} = \exp\{K_X^*(t)\}$ and of (19). This illustrates the link between the saddlepoint density approximation and the solution to the $KD(\mu, \nu_t)$, passing through the Legendre transform K_X^*—or equivalently through the Legendre transform $K_{\bar{X}_n}^*$.

3.1.3 Example (cont'd): mean of exponential random variables

To illustrate the concepts described in §3.1.1 and §3.1.2, we keep working on the the example of §2.4 and consider the mean $(\bar{X}_n)$ of n i.i.d. copies

of X having an exponential density. The R code to replicate the numerical exercises is available at https://github.com/dvdlvc/MyGitHub/tree/Saddlepoint_MeasureTransportation.

Density approximation. Let us start from the density density approximation. The plots in Figure 2 and Figure 3 complete the information provided in Figure 1. In Figure 2 we illustrate, for $n = 10$, the inaccuracy of the asymptotic theory, which performs poorly in the center and also in both tails of the distribution. The Edgeworth expansion yields accuracy improvements on the asymptotic theory. However, in the left tail the Edgeworth approximation becomes negative. In Figure 3 we depict how the Esscher tilting and the use of the saddlepoint density approximation, as obtained using the optimal transportation (see middle panel of Figure 1), overcomes this problem. To illustrate this aspect, we compare the Edgeworth and the saddlepoint approximations to the exact density of $\bar{X}_n$. Both approximations are available in the the statistical software R, R Core Team (2013); see functions `edgeworth` and `saddlepoint` in the package `EQL`. As far as the saddlepoint density is concerned, the formulae for the Legendre transform, the saddlepoint and the conjugate density are available in §2.4. Using these expressions in (20) we obtain f_{sad}. For each n, the exact density of $\bar{X}_n$ is known (the Gamma distribution).

We consider $n = 10, 50, 100, 250$. In Figure 3 we display the relative error, computed as 100*(approx density - true density)/(true density), as entailed by each approximation for different sample sizes. The plots illustrate the improvement that the tilting of the underlying distribution yields in terms of approximation accuracy, especially in the tails. For instance, when $n = 50$, the relative error entailed by the Edgeworth expansion and by the saddlepoint approximation are similar for $x \in [0.8, 1.2]$, but outside this central region the error of the Edgeworth is higher than the one entailed by the saddlepoint approximation. Looking at the plots, we see that the accuracy of Edgeworth improves when n increases, but for $n = 250$ the relative error entailed by the Edgeworth approximation in the left tail of the distribution (see bottom right plot) is still higher than the error entailed by the saddlepoint approximation.

Saddlepoint test via Legendre transform. Let us consider the mean of n iid random variables $X_i \sim \exp(\alpha)$. Assume we want to test the hypothesis $\mathcal{H}_0 : \alpha = 1$ versus $\mathcal{H}_1 : \alpha > 1$. Following Robinson et al. (2003), to perform the test, we make use of the saddlepoint test statistic based on the Legendre transform of X evaluated at $\bar{X}_n$, i.e. $2nK_X^*(\bar{X}_n)$, where $K_X^*(t) =$

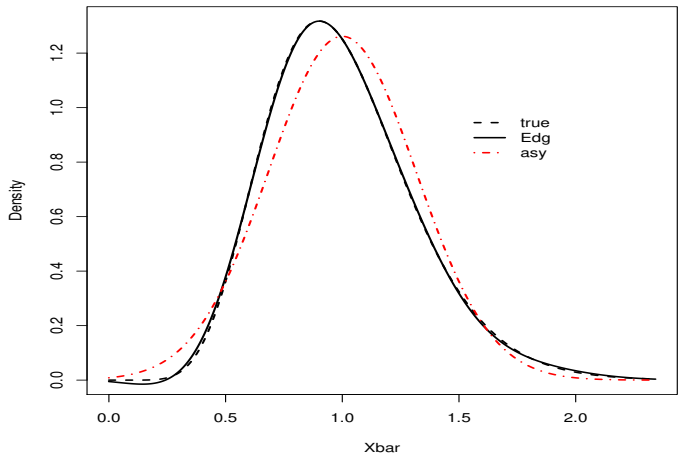

Figure 2: Density of the sample mean ($\bar{X}_n$) of $n = 10$ i.i.d. random variable distributed as an exponential one. True density (dashed line), Edgeworth expansion (continuous line) and Gaussian asymptotic approximation (dash-dotted line).

$\sup_\upsilon\{\upsilon t - K_X(t)\} = (t-1) - \log(t)$.

Robinson et al. (2003) prove that under $\mathcal{H}_0$ the distribution of the test can be approximated by a $\chi^2(1)$ distribution, with relative error of order $O(n^{-1})$. This yields a very accurate approximation of the level of the test, even for small sample sizes. The left panel of Figure 4 illustrates this aspect, depicting the accuracy of the approximation for $n = 10$. To obtain the plot, we simulate 5000 samples, drawing from exp(1) and for each sample we compute the test statistic. Finally, we do a QQplot, comparing the quantiles of the distribution of the test statistic with the quantiles of the $\chi^2(1)$ distribution. Inspecting the picture, we see that the quantiles of the test statistics are close to the ones of the $\chi^2(1)$: the approximation is remarkably accurate for the 0.95 and 0.975 quantiles, which are the quantiles typically applied for hypothesis testing. Similar pictures are available for $n = 50$ and 250, where the accuracy of the approximation is even better.

Beside the behaviour under the null, another key aspect of the testing problem is the power of the test based on $2nK_X^*(\bar{X}_n)$. To investigate this aspect, in the right panel of Figure 4 we plot the power curves for $n = 10, 50, 250$. To obtain the plots, we proceed as described in the paragraph

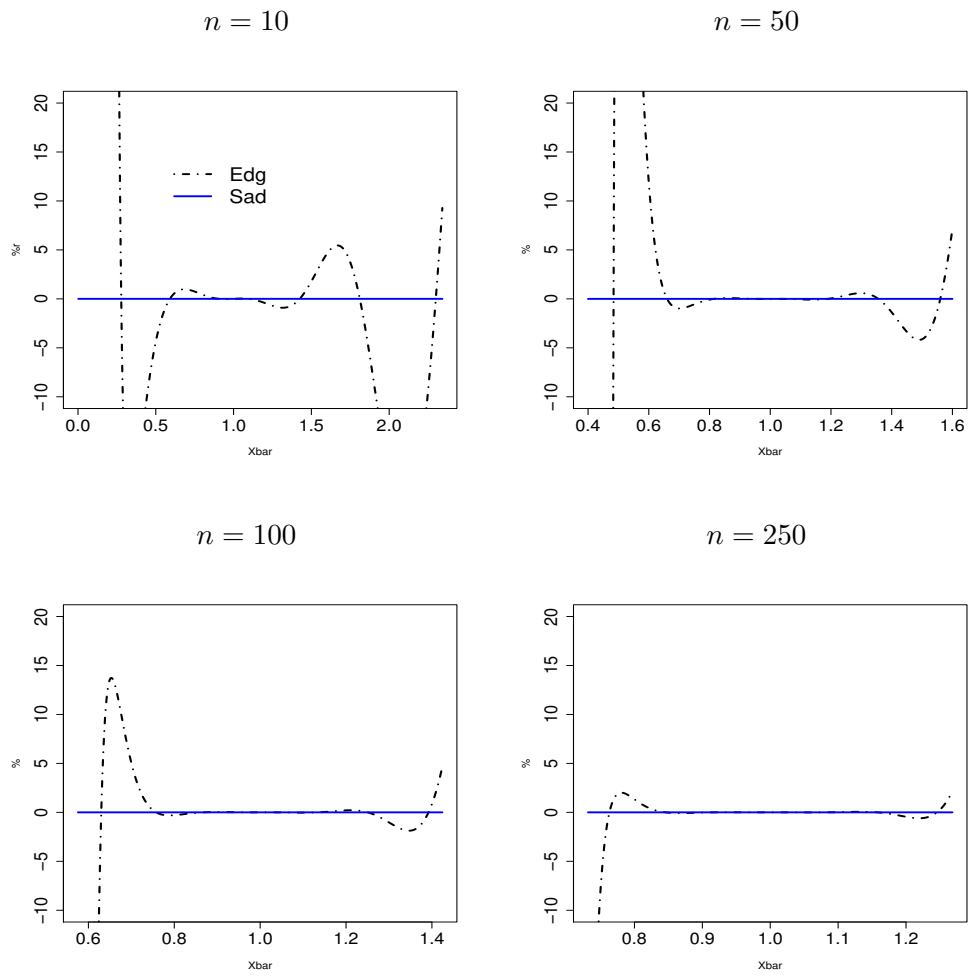

Figure 3: Relative error in percentage (y-axis) for the Edgeworth (dot-dashed line) and saddlepoint (continuous line) approximation to the density of the sample mean of n i.i.d. random variable distributed as an exp(1). The comparison is for different sample sizes.

about the QQplot. The main difference is that, for each sample size, we simulate 5000 samples using a sequence of alternative hypotheses and we consider the frequency of non acceptance of $\mathcal{H}_0$. For each sample size, we have random drawings of size n from an $\exp(1+\delta)$, where $1+\delta > 1$, represents the value of α under the alternative hypothesis and we consider

$\delta \in [0, 0.8]$. We see that the test has good power already for $n = 10$. Clearly, the larger the sample size, the higher the power.

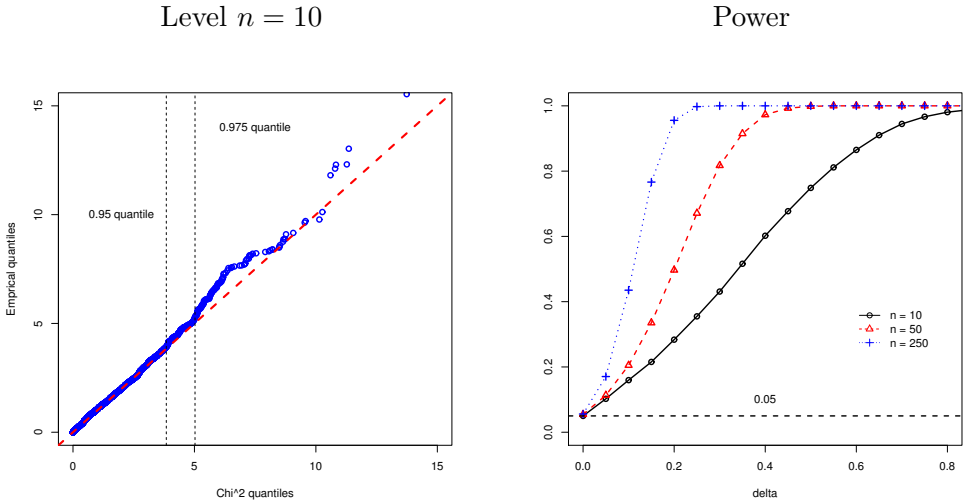

Figure 4: QQ plot for the test of the sample mean of n i.i.d. random variable distributed as an $\exp(\alpha)$. Left panel, $n = 10$ and $\alpha = 1$ (density under $\mathcal{H}_0$). Right panel, power for sample size $n = 10, 50, 250$.

3.2 General statistics with known cumulant generating function

Consider now a real-valued statistic $U_n = U(X_1, ..., X_n)$ having a density f_n and c.g.f. $K_{U_n}(\upsilon) = \log E_{f_n}[\exp\{\upsilon U_n\}]$, which is well-defined and known in a closed-form. It is important to note that U_n need not be linear in the X_i's as it was the case with the sample mean.

Then, by Fourier inversion we obtain:

$$f_n(t) = \frac{n}{2\pi i} \int_{-i-\infty}^{i+\infty} \exp\{nR_n(\upsilon) - n\upsilon t\}, \tag{21}$$

where

$$R_n(\upsilon) = K_{U_n}(n\upsilon)/n. \tag{22}$$

This is a generalization of (11): if $U_n = n^{-1}\sum_{i=1}^n X_i$, we have $R_n(\upsilon) \equiv K_X(\upsilon)$ and (21) coincides with (11). Following Easton and Ronchetti (1986),

the saddlepoint equation is defined by $R'_n(\upsilon) = t$ and the saddlepoint density approximation $f_{\text{sad}}(t)$ is as in (13) and (12), with K_X replaced by R_n.

The interpretation of the saddlepoint procedure in terms of measure transportation (as a solution to the Kantorovich dual problem) remains essentially the same as in §3.1.2. Briefly, we have $U_n \sim \mu^{(n)}$ and we move its mass into a new measure $\nu_t^{(n)}$ such that $E_{\nu_t^{(n)}}[U_n] = t$. This implies that the dual Kantorovich problem can be written as $\text{KD}(\mu_\theta^{(n)}, \nu_t^{(n)})$ as in (6), with a cost function $c(x, y) = -xy$. Defining

$$R_n^*(t) = \sup_{\upsilon} \{\upsilon t - R_n(\upsilon)\}, \tag{23}$$

as the Legendre transform of R_n, the pair (R_n, R_n^*) is a solution to the $\text{KD}(\mu_\theta^{(n)}, \nu_t^{(n)})$ problem, as discussed in §3.1.2.

The main difference with the construction in §3.1.2 lies in the (possible) non-linearity of U_n. We can no longer exploit the linearity of the functional in the X_is and work on the conjugate density of f_n in order to recenter the statistic at the point t. Rather, we have to perform the Esscher tilting directly on f_n and obtain its conjugate density, say $h_{n,t}$, via R_n^*. Doing so (see formula (2.9) in Easton and Ronchetti (1986)) we obtain

$$h_{n,t}(x) = C_n(t) f_n(x) \exp\{n[R_n(\upsilon(t) - \upsilon t]\},$$

with $\upsilon(t)$ is defined via $R'_{U_n}(\upsilon(t)) = t$ and $C_n(t)$ is an integration constant. Under $h_{n,t}$ we have the desired recentering, since

$$R'_{U_n}(\upsilon(t)) = t \iff E_{h_{n,t}}[U_n] = t., \tag{24}$$

which, similarly to (17), illustrates that the saddlepoint $\upsilon(t)$ performs a recentering of the measure of U_n at the point of interest t.

4 M-estimators

The previous sections emphasize the link between Esscher tilting and measure transportation, when the c.g.f. of the random variable we want to transform is available in closed-form. In this section, we consider a more general case in which the c.g.f. of the statistic we are working with is not available in closed-form. Among the possible statistics with this property, we focus on M-estimators, whose saddlepoint density approximations are derived in Field (1982). For the sake of exposition, in what follows, we confine ourselves to the case of a univariate parameter and a univariate random variable, but our arguments can be easily extended to the case of a vector parameter.

4.1 Setting and notation

Similarly to §2, let $X \sim F$, where $dF(x) = f(x)dx$, or equivalently, let us assume that X has measure μ, whose support is $\mathcal{X} \subseteq \mathbb{R}$. We are interested to make inference on some parameter $\theta(F)$. Given a random sample $X_1, ..., X_n$ of i.i.d. copies of X, an M-estimator of θ is the solution U_n to

$$\sum_{i=1}^{n} \psi(X_i; \theta) = 0, \tag{25}$$

where $\psi : \mathcal{X} \times \Theta$ and $\Theta \subseteq \mathbb{R}$. The population version of (25) is $E_f[\psi(X; \theta)] = 0$, where E_f represents the expected value taken w.r.t. to f.

M-estimators include the maximum likelihood estimator as a special case, when F belongs to a parametric family of distributions $\{F_\theta\}$ and $\psi(x; \theta) = \partial_\theta \log f_\theta(x)$. Moreover, setting $\psi(x; \theta) = x - \theta$ in (25), U_n becomes the sample mean. We refer to Huber (1981), Huber and Ronchetti (2009), Hampel et al. (1986), van der Vaart (1998), and Serfling (2009) for book-length presentations.

Standard first-order asymptotic results establish the asymptotic normality of M-estimators via the first-order von Mises expansion:

$$U_n - \theta(F) = \frac{1}{n} \sum_{i=1}^{n} IF_\psi(X_i; \theta, F) + o_p(n^{-1/2}), \tag{26}$$

where IF_ψ is the influence function (Hampel (1974))

$$IF_\psi(x; \theta, F) = \frac{\psi(x; \theta)}{M(\theta)},$$

where $M(\theta) = E_f[-\partial_\theta \psi(X; \theta)]$.
The expansion in (26) is the step stone for first-order asymptotics: under suitable assumptions (see e.g. Huber (1981) Ch. 6.3, Fernholz (1983), Welsh (1996) p. 194), the remainder in (26) is negligible and

$$n^{1/2}(U_n - \theta) \xrightarrow{D} \mathcal{N}(0, V(\theta)),$$

where $\xrightarrow{D}$ represents the convergence in distribution and $\mathcal{N}(0, V(\theta))$ is a Gaussian distribution with expectation zero and variance $V(\theta) = Q(\theta)/M^2(\theta)$, and $Q(\theta) = E_f[\psi^2(X; \theta)]$.

The expansion in (26) shows that the distributional properties of the M-estimator depend on the distributional properties of the estimating function. This provides the intuition for the fact that, to derive higher-order

asymptotic properties of U_n, in particular its f_{sad}, we need to work on ψ. Thus, let $K_\psi(\upsilon;\theta) = \log E_f[\exp\{\upsilon\psi(X;\theta)\}$ be the c.g.f. of $\psi(X;\theta)$. We use $K_\psi(\upsilon;\theta)$ to show how the arguments of §2 adapt to the case of M-estimators.

In §4.2 and in §4.3, we explain how to link the saddlepoint of U_n with the measure transportation theory following the same arguments as in §3. We start from the formula for the saddlepoint density approximation to f_n. Then, we follow the arguments in §3 and extend them for the general functional U_n. At the end, we recall the method of the conjugate density for M-estimators and its link to information theory. The main difference with the construction in §3 is that the c.g.f. of U_n is (typically) unavailable in closed-form. To overcome this problem we need to approximate K_{U_n} using (26): this leads to the definition of the saddlepoint equation based on K_ψ. In this way, we derive the conjugate density and we illustrate that it is the solution to a KL minimization problem as in (2) with a constraint expressed in terms of ψ. This illustrates the connections between saddlepoint and information theory. Then, we state the KD problem and highlight the relation between its solution, the Legendre transform of the approximation to K_{U_n}, and the saddlepoint. Finally, we explain why the Legendre transform of the approximation to K_{U_n} and the saddlepoint density approximation are related.

4.2 Saddlepoint approximation for M-estimators

Theorem 1 in Field (1982) shows that the density $f_n(t)$ of U_n can be approximated by an expansion of the form

$$f_n(t) = f_{\text{sad}}(t)\{1 + O(n^{-1})\},$$

where the saddlepoint density approximation is

$$f_{\text{sad}}(t) = (n/2\pi)^{1/2} C^{-n}(t) \left| E_{h_{\psi,t}}\left[\frac{\partial\psi(X;t)}{\partial t}\right]\right| \left[E_{h_{\psi,t}}\left[\psi^2(X;t)\right]\right]^{-1/2} , \quad (27)$$

where

$$h_{\psi,t}(x) = C(t)e^{\upsilon(t)\psi(x;t)} f(x), \quad (28)$$

is the conjugate density, with $C(t) = \exp\{-K_\psi(\upsilon(t);t)\}$ and $\upsilon(t)$ is the saddlepoint, i.e., the solution of

$$\partial_\upsilon K_\psi(\upsilon;t) = 0, \quad (29)$$

or equivalently $E_{h_{\psi,t}}[\psi(X;t)] = 0$.

4.3 Connections to measure transportation

Notice that $h_{\psi,t}$ in (28) is defined via the saddlepoint $\upsilon(t)$ which is the solution of the equation

$$\begin{aligned} \partial_\upsilon K_\psi(\upsilon;t)=0 \quad &\Longleftrightarrow \quad E_{h_{\psi,t}}[\psi(X;t)]=0 \\ &\Longleftrightarrow \quad E_f[\psi(X;t)\exp\{\upsilon\psi(X;t)\}]=0. \end{aligned} \tag{30}$$

In (30), $E_{h_{\psi,t}}$ and E_f represent the expected value computed w.r.t. the conjugate and the original density, respectively. As noticed in Field (1982), p. 678, $h_{\psi,t}(x)$ is the conjugate density of the linearized version of U_n, as obtained in (26). Field (1982) shows that even if we neglect the remainder in (26) the order of approximation error entailed by the saddlepoint approximation is $O(n^{-1})$. Moreover, $\upsilon(t)$ is such that the conjugate density is centering U_n at t, namely

$$\partial_\upsilon K_\psi(\upsilon;t)|_{\upsilon=\upsilon(t)}=0 \iff E_{h_{\psi,t}}[\psi(X;t)]=0 \iff E_{h_{\psi,t}}[U_n]=t+O(n^{-1}). \tag{31}$$

Starting from (30), we introduce a modified version of the Kullback-Leibler divergence problem in (2). For fixed $t\in\Theta$, we can prove (see Appendix) that $h_{\psi,t}(x)$ in (28) is a solution of the following information-theoretic problem:

$$\min_{g\in\mathcal{G}}\left\{\int_{\mathcal{X}} g(x)\log\frac{g(x)}{f(x)}dx\right\}, \quad \text{s.t.} \quad g(x)\geq 0, \textstyle\int_{\mathcal{X}} g(x)dx=1, E_g[\psi(X;t)]=0, \tag{32}$$

where $\mathcal{G}$ contains all the densities having support $\mathcal{X}$ and such that ψ has finite second moment. Notice that the constraint in (32)

$$E_g[\psi(X;t)]=\int_{\mathcal{X}}\psi(x;t)g(x)dx=0$$

is set up in terms of the distribution properties of ψ and it implies that, under g, the M-estimator U_n is centered at t—up to the remainder in (31).

Therefore, the saddlepoint solution to (30) yields an optimal property in terms of information theory, illustrating the link between saddlepoint and information theory as discussed in §2.2.1.

To make the connection with measure transportation as in §2.2, we parallel the argument in §3.2. Let $U_n\sim\mu^{(n)}$: the saddlepoint induces a measure transport of $\mu^{(n)}$ into a new measure $\nu_t^{(n)}$, which is such that

$$E_{\nu_t^{(n)}}[U_n]=t+O(n^{-1}).$$

As in §3.2, we set the dual Kantorovich problem $\mathrm{KD}(\mu_\theta^{(n)}, \nu_t^{(n)})$ using as a cost function $c(x, y) = -xy$. Since K_{U_n} is not available in closed-form, we proceed as in Easton and Ronchetti (1986) and we approximate it using K_ψ. Let us define $\tilde{K}_{U_n}(\upsilon) = nK_\psi(\upsilon/n; \theta)$. Then, by Fourier inversion we have

$$\tilde{f}_n(t) = \frac{n}{2\pi i} \int_{-i-\infty}^{i+\infty} \exp\{n\tilde{R}_n(\upsilon) - n\upsilon t\}. \tag{33}$$

where

$$\tilde{R}_n(\upsilon) = \tilde{K}_{U_n}(n\upsilon)/n. \tag{34}$$

The saddlepoint is defined by

$$\tilde{R}_n'(\upsilon) = t \tag{35}$$

which is the maximizer υ of

$$\tilde{R}_n^*(t) = \sup_\upsilon \{\upsilon t - \tilde{R}_n(\upsilon)\}, \tag{36}$$

the Legendre transform of the approximated c.g.f. of U_n. Hence, we may consider the pair $(\tilde{R}_n, \tilde{R}_n^*)$ as the solution to $\mathrm{KD}(\tilde{\mu}^{(n)}, \nu_t^{(n)})$, where $\tilde{\mu}^{(n)}$ is the "measure" related to the "density" $\tilde{f}_n$, as in (33). This remark leads to the consideration that the problem $\mathrm{KD}(\tilde{\mu}_\theta^{(n)}, \nu_t^{(n)})$ is peculiar: we have to perform a measure transportion in which we know only approximately the original measure and we know that the target measure $\nu_t^{(n)}$ is such that $E_{h_{\psi,t}}[\psi(X; t)] = 0$. On the other hand, this can be viewed from the point of view of the robustness framework, where the original measure is only known to lies in a neighborhood of some "model" measure. Perhaps well known tools from the statistical robustness theory can be used here to analyze the impact of this uncertainty.

5 Discussion and outlook

In this paper we highlight the connections between information theory, measure transportation, and saddlepoint approximations. The links between these areas form a remarkable picture, which relates well-known elements from diverse mathematical fields to data science. We believe that many other connections can be established and we are planning to work on them in the near future. Here, we provide some possible developments.

5.1 General Legendre transform

The rigorous proof of our results for M-estimators remains the object of future research. In particular, the remainder term in (26) has to be carefully addressed, and suitable assumptions have to be introduced. Nevertheless, (34) and (35) provide an interesting research direction. Indeed, (35) is equivalent to solving $\partial_\upsilon K_\psi(\upsilon;t)|_{\upsilon=\upsilon(t)} = 0$, which in turn is equivalent to finding $\sup_\upsilon\{-K_\psi(\upsilon;t)\}$. This leads to the definition of

$$K_\psi^\dagger(t) = \sup_\upsilon\{-K_\psi(\upsilon;t)\}, \tag{37}$$

which is a "Legendre-type transform" of K_ψ.

Simple calculations show that $\partial_t K_\psi(\upsilon;t) = \tilde{M}(t)\ \upsilon$, where $\tilde{M}(t) = E_{h_{\psi,t}}\left[\partial\psi(X;t)/\partial t\right]$ and

$$\begin{aligned}\frac{d}{dt}K_\psi^\dagger(t) &= -\frac{d}{dt}K_\psi(\upsilon(t);t)\\ &= -\partial_\upsilon K_\psi(\upsilon(t);t)\ \frac{d}{dt}\upsilon(t) - \partial_t K_\psi(\upsilon(t);t) = -\tilde{M}(t)\ \upsilon(t).\end{aligned}$$

Therefore,

$$\frac{d}{dt}K_\psi^\dagger(\partial_\upsilon K_\psi(\upsilon(t);t)) = \frac{d}{dt}K_\psi^\dagger(0) = 0,$$

which generalizes the characterization between the derivatives of a function and its Legendre transform.

The function $K_\psi^\dagger$ has been used in the definition of the saddlepoint test based on M-estimators introduced by Robinson et al. (2003); see the numerical illustration for the mean in §3.1.3. Moreover, Ronchetti and Welsh (1994) apply $K_\psi^\dagger$ (as obtained using the empirical measure of the oberva-tions) to define the empirical saddlepoint approximation of M-estimators: Monti and Ronchetti (1993) apply it to unveil the connection between the empirical saddlepoint techniques and the empirical likelihood; Gatto (2017) uses $K_\psi^\dagger$ to define two tests on the mean direction of the von Mises–Fisher distribution: the tests perform well even in large dimensional settings, with small sample sizes. Additional applications of this and related tests can be found in Toma and Leoni-Aubin (2010), Toma and Broniatowski (2011), Aeberhard et al. (2017), and Holcblat and Sowell (2019).

These papers illustrate that the use of $K_\psi^\dagger$ is well-established in the statistical literature. As far as the mathematical literature is concerned, the study of $K_\psi^\dagger$ might offer new perspectives in convex analysis and/or measure transportation theory. With this regard, notice that in the case of the sample

mean, where the estimating function is $\psi(x;t) = x - t$, $K^{\dagger}_{\psi}(t) \equiv K^{*}_{X}(t)$, so it coincides with the Legendre transform; see Figure 4 for an illustration.

For other (complicated, non linear) estimating functions, we conjecture that $K^{\dagger}_{\psi}$ is a c-transform, namely a transform that includes the Legendre transform as a special case; see §2.2.1. In other words, we conjecture that we may set up a KD problem for the measure of ψ, where the original measure under f is transported into a new measure having a density $h_{\psi,t}$. The function $K^{\dagger}_{\psi}$ characterises the solution to that KD problem—in the same way as the Legendre transform characterizes the solution to (6) when $c(x,y) = -xy$. The fundamental difference between the aforementioned measure transportation problem dealing with the measure of ψ and those discussed in §2 and §3 is that the cost function is not $-xy$, rather, it is a function depending non-linearly on t (e.g. via the constraint $E_{h_{\psi,t}}[\psi(X;t)] = 0$). What is the functional form of this cost function? What is the optimal transportation map within this setting? These are some open questions for which further investigation is required.

5.2 Change of variable

Measure transportation is essentially a change of variable. In §2.1 we illustrates that the conjugate density is obtained via the Jacobian formula in (10). This connection unveils a link between information theory and the theory of nonlinear partial differential equations (PDEs). As pointed out in Villani (2009) p. 282, the Jacobian formula (10) is related to the Monge-Ampère partial differential equation, which is a nonlinear PDE arising in several problems, such as the Gaussian curvature equation and affine geometry; see De Philippis and Figalli (2014) for a recent review. Using our notation related to saddlepoint techniques, we flag that the Monge-Ampère equation, for the cost $c(x,y) = -xy$ in $\mathbb{R}$, becomes

$$f_Y(T(x)) = f_X(x) \left(\left| \frac{\partial \mathcal{T}(y)}{\partial y} \right| \right)^{-1}, \tag{38}$$

which is equivalent to (10); see Villani (2009) Eq. (12.4). Thus, we have that h_t can be characterized via the solution to an elliptical PDE. To our knowledge this connection is unexplored and can be studied to relate the saddlepoint to the notion of Gaussian curvature in Riemanian geometry—or more generally to the Ricci's curvature; see Villani (2009) Ch. 14, where the Jacobian determinant appearing in (38) plays a pivotal role. For a related discussion, see also §5.3. Moreover, as noticed in Villani (2009), p.

272, in general there are two points that one should check before writing the Jacobian formula in (10). Both points are related to the regularity of $\mathcal{T}$: first, $\mathcal{T}$ should be injective on its domain of definition, and second, $\mathcal{T}$ should possess some minimal regularity that allows for the existence and differentiability of its inverse. These requirements are the typical assumptions cited in the statistical literature; see e.g. Mood et al. (1974)—where the derivation of (10) is obtained using standard theorems from calculus on change of variables. Unfortunately, for measure transportation problems these requirements are too stringent. Nevertheless, Theorem 11.1 in Villani (2009) proves that if the map $\mathcal{T}$ is differentiable almost everywhere (equivalently, $\mathcal{T}$ is the gradient of a convex function related to the c.g.f.), (10) holds. This result suggests that the Jacobian formula can be obtained under some assumptions which are weaker than those commonly applied in statistics. In the same spirit, Theorem 11.1 in Villani (2009) provides a set of assumptions alternative to those usually applied in information theory for the derivation of the conjugate density; see Field and Ronchetti (1990) and reference therein. The statistical interpretation of these assumptions already available in the mathematical literature remains an open question.

5.3 Geodesics and saddlepoints

In Section 2.2.2, we discuss the notion of geodesic and we explain how it is related to the saddlepoint. We present only some aspects, but we believe that the topic is much broader and it deserves further investigation, which may shed light on the geometry of the saddlepoint. This research direction should contribute to the literature on the interplay between geometry, asymptotics, and inference; see the seminal book by Fraser (1968), and e.g. Kass (1989), Amari (1989), Amari (2016).

The link between saddlepoint and information geometry (which studies the properties of a manifold of probability distributions) can be useful for various applications in statistics, machine learning, signal processing, and optimization. For a book-length description, see Amari (2016), Ch. 11-13, and for recent developments see Amari et al. (2018). As in our Section 2, also Amari's construction hinges on the Legendre transform (see Th. 4 in Amari et al. (2018)) and on the exponential family (see Amari (2016) Ch. 2). We feel that this is a challenging, promising, and almost unexplored research area. For instance, the expression in (8) is well-known in the literature on measure transportation theory; see e.g. Villani (2009). However, in the statistical literature, its interpretation and its role are unknown.

Additionally, one may consider the Bregman divergence induced by the

c.g.f. (a convex function) and investigate its connection to the 2-Wasserstein distance. As in Section 3.1, one may start from the case of $\bar{X}_n$, where the X_i are i.i.d. copies of $X \sim f_X$ and f_X belongs to the exponential family. Ch. 1 and Ch. 2 in Amari (2016) illustrate that the manifold of exponential families can be regarded as a primal manifold $\mathbb{M}$, say, whose coordinate system is obtained via the canonical parameter. The Legendre transform of K_X defines a dual manifold (via the normal vector), having a coordinate system which is coupled with the coordinate system of $\mathbb{M}$. Then, the c.g.f. yields a Bregman divergence on $\mathbb{M}$ (see Eq. (1.57) in Amari (2016)) and its Legendre transform yields a divergence on the dual of $\mathbb{M}$ (see Eq. (1.67) in Amari (2016)). Within this context, one may study the Riemannian structure (e.g. the curvature) of $\mathbb{M}$ and of its dual, working on the notion of geodesic. One may dig more into the subject and consider statistics other than the sample mean. For instance, a possible research direction is related to the study of the links between the curvature and the Bregman divergence in the case of a general statistic (like the statistics of Section 3.2). In the same spirit, one may explore the connections between the curvature and the Fisher information matrix (an invariant Riemannian metric in the exponential family, see Kass (1989), Amari et al. (2018) and reference therein) in the setting of M-estimators.

5.4 Concentration inequalities and large deviations

We did not explore the links between concentration inequalities (as obtained using the Wasserstein distance) and the theory of large deviations—as derived using the Legendre transform defined by the saddlepoint, as in the Chernoff bound and the Bahadur half slope; see Chernoff (1952), Bahadur (1971). These connections are more involved than those mentioned in this paper and they require a separate investigation; some results are available in Arcones (2006). Here, we say that a step stone for new research in this direction may be related to Theorem 5.28 and Example 5.29 (p. 80) in Villani (2009), where dual transport inequalities are obtained linking the Wasserstein distance and the Kullback information. Working on the same lines as in the proof on p. 81-83 of Villani (2009), we conjecture that concentration inequalities can be obtained exploiting the relation between the Wasserstein distance, the Legendre transform, and the Kullback information derived in §2.1-2.2.2. Then, these inequalities can be helpful in the large deviation analysis of the saddlepoint density approximation, as in section 7 of Daniels (1954)—see Eq. (7.2) in that paper.

5.5 Connections to machine learning

Information theory (via entropy and Kullback-Leibler minimization) plays a pivotal role in the literature on machine learning, e.g. for patterns recognition; see Grenander et al. (2007) and Murphy (2012), among the others. Moreover, measure transportation is advocated by machine learners for the analysis of large data sets; see Cuturi (2013) and Peyré and Cuturi (2019). In contrast, the use of saddlepoint techniques has been overlooked by the machine learning community and more computationally intensive methods (typically, the bootstrap) are preferred; see e.g. Murphy (2012) and James et al. (2013). We hope that this paper may draw the attention of machine learners toward saddlepoint techniques, stimulating their use to solve some open problems. For instance, we mention two possible research directions, where the saddlepoint techniques can be of help: the need for speeding up machine learning algorithms, replacing resampling methods by saddlepoint approximations, without compromising accuracy—see Davison and Hinkley (1988); the problem of conducting inference on generative models where the available data set is such that the ratio between the number of parameters and the sample size is large and the problem related to post-selection inference.

Acknowledgement

Andrej Ilievski thanks the Office of Science, Technology and Higher Education (OSTHE) at the Embassy of Switzerland in Washington, D.C. for the financial support through the ThinkSwiss program. Davide La Vecchia acknowledges the Swiss National Science Foundation grant number 100018_169559 for financial support.

References

Aeberhard, W. H., Cantoni, E., and Heritier, S. (2017). Saddlepoint tests for accurate and robust inference on overdispersed count data. *Computational Statistics & Data Analysis*, 107:162–175.

Amari, S. (1989). The geometry of asymptotic inference: Comment. *Statistical Science*, 4(3):220–222.

Amari, S. (2016). *Information geometry and its applications*, volume 194. Springer.

Amari, S., Karakida, R., and Oizumi, M. (2018). Information geometry connecting Wasserstein distance and Kullback-Leibler divergence via the entropy-relaxed transportation problem. *Information Geometry*, 1(1):13–37.

Arcones, M. A. (2006). Large deviations for M-estimators. *Annals of the Institute of Statistical Mathematics*, 58(1):21–52.

Bahadur, R. (1971). *Some Limit Theorems in Statistics*. Soc. Ind. Appl. Math., Philadelphia.

Barndorff-Nielsen, O. and Cox, D. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman and Hall London.

Barndorff-Nielsen, O. and Cox, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(3):279–299.

Bhattacharya, R. N. and Rao, R. R. (1986). *Normal approximation and asymptotic expansions*, volume 64. Siam.

Brazzale, A., Davison, A. C., and Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge University Press.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on sums of observations. *Annals of Mathematical Statistics*, 23:493–507.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, 25:631–650.

Davison, A. C. and Hinkley, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika*, 75(3):417–431.

De Philippis, G. and Figalli, A. (2014). The Monge–Ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51(4):527–580.

Easton, G. S. and Ronchetti, E. (1986). General saddlepoint approximations with applications to L statistics. *Journal of the American Statistical Association*, 81(394):420–430.

Esscher, F. (1932). On the probability function in the collective theory of risk. *Skand. Aktuarie Tidskr.*, 15:175–195.

Fernholz, L. T. (1983). *von Mises Calculus For Statistical Functionals*. Springer Science & Business Media.

Field, C. (1982). Small sample asymptotic expansions for multivariate M-estimates. *The Annals of Statistics*, 10:672–689.

Field, C. A. and Ronchetti, E. (1990). *Small Sample Asymptotics*. IMS, Lecture notes-monograph series.

Fraser, D. A. S. (1968). *The Structure of Inference*. Wiley, New York.

Galichon, A. (2016). *Optimal Transport Methods in Economics*. Princeton University Press.

Galichon, A. (2017). A survey of some recent applications of optimal transport methods to econometrics. *Econometrics Journal*, 20:C1–C11.

Gatto, R. (2017). Multivariate saddlepoint tests on the mean direction of the von Mises–Fisher distribution. *Metrika*, 80(6-8):733–747.

Goutis, C. and Casella, G. (1999). Explaining the saddlepoint approximation. *The American Statistician*, 53(3):216–224.

Grenander, U., Miller, M. I., Miller, M., et al. (2007). *Pattern Theory: From Representation to Inference*. Oxford University Press.

Hallin, M. (2017). On distribution and quantile functions, ranks and signs in $\mathbb{R}^d$: a measure transportation approach. Available at ideas.repec.org/p/eca/wpaper/2013-258262.html.

Hallin, M., La Vecchia, D., and Liu, H. (2019). Center-outward R-estimation for semiparametric VARMA models. arXiv:1910.08442.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393.

Holcblat, B. and Sowell, F. (2019). The empirical saddlepoint estimator. *arXiv preprint arXiv:1905.06977*.

Huber, P. J. (1981). *Robust Statistics*. Wiley.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. Wiley, 2nd edition.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 112. Springer.

Jensen, J. L. (1995). *Saddlepoint Approximations*. Oxford University Press.

Kantorovich, L. V. (1942). On the translocation of masses. *(Dokl.) Acad. Sci. URSS*, 37(3):199–201.

Kass, R. (1989). The geometry of asymptotic inference. *Statistical science*, 4(3):188–219.

Kolassa, J. (2006). *Series Approximation Methods in Statistics*, volume 88. Springer.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59.

Kremer, E. (1982). A characterization of the Esscher-transformation. *ASTIN Bulletin: The Journal of the IAA*, 13(1):57–59.

Kullback, S. (1997). *Information Theory and Statistics*. Dover Publications.

McCann, R. J. and Guillen, N. (2011). Five lectures on optimal transportation: geometry, regularity and applications. *Analysis and geometry of metric measure spaces: lecture notes of the séminaire de Mathématiques Supérieure (SMS) Montréal*, pages 145–180.

McCullagh, P. (2018). *Tensor Methods in Statistics*. Dover Publications.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Monti, A. C. and Ronchetti, E. (1993). On the relationship between empirical likelihood and empirical saddlepoint approximation for multivariate M-estimators. *Biometrika*, 80(2):329–338.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Panaretos, V. and Zemel, Y. (2018). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and its Application*, 6:405–431.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science*, 3:213–227.

Robinson, J., Ronchetti, E., and Young, G. (2003). Saddlepoint approximations and tests based on multivariate M-estimates. *The Annals of Statistics*, 31:1154–1169.

Rockafellar, R. T. (2015). *Convex Analysis*. Princeton University Press.

Ronchetti, E. and Welsh, A. (1994). Empirical saddlepoint approximations for multivariate M-estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56:313–326.

Serfling, R. J. (2009). *Approximation Theorems of Mathematical Statistics*, volume 162. Wiley.

Small, C. G. (2010). *Expansions and Asymptotics for Statistics*. Chapman and Hall/CRC.

Toma, A. and Broniatowski, M. (2011). Dual divergence estimators and tests: robustness results. *Journal of Multivariate Analysis*, 102:20–36.

Toma, A. and Leoni-Aubin, S. (2010). Robust tests based on dual divergence estimators and saddlepoint approximations. *Journal of Multivariate Analysis*, 101:1143–1155.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics.

Villani, C. (2009). *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media.

Welsh, A. H. (1996). *Aspects of Statistical Inference.* Wiley.

Young, G. A. (2009). Routes to higher-order accuracy in parametric inference. *Australian & New Zealand Journal of Statistics*, 51(2):115–126.

Appendix

A Proof of the Kullback-Leibler optimization problems in §2.1 and §4.3

We show that for a fixed t, the conjugate density h_t is the solution to the following information theoretic problem:

$$\min_{g\in\mathcal{G}} \int_{\mathcal{X}} g(x)\log\frac{g(x)}{f_X(x)}dx, \quad \text{s.t.} \quad g(x)\geq 0, \quad \int_{\mathcal{X}} g(x)dx = 1, \quad \int_{\mathcal{X}} xg(x)dx = t.$$

The Lagrangian of this problem is

$$L[g(x)] = g(x)\log\tfrac{g(x)}{f_X(x)} - \lambda_1 g(x) - \lambda_2(x-t)g(x),$$

and the Euler-Lagrange equation

$$\frac{\partial L}{\partial g} - \frac{d}{dx}\frac{\partial L}{\partial g'} = 0.$$

Since $\frac{\partial L}{\partial g'} = 0$ we have

$$0 = \tfrac{\partial L}{\partial g} = \tfrac{f_X(x)}{g(x)}\tfrac{1}{f_X(x)}g(x) + \log\tfrac{g(x)}{f_X(x)} - \lambda_1 - \lambda_2(x-t),$$

i.e.

$$\log\tfrac{g(x)}{f_X(x)} = \lambda_1 - 1 + \lambda_2(x-t).$$

If we let $c_1 = e^{\lambda_1 - 1}$, the last expression implies that

$$g(x) = c_1 e^{\lambda_2(x-t)} f_X(x).$$

Notice a slight abuse of notation: in fact, the function g depends on x and t, but we drop the last argument and we write $g(x)$.

We are now left to find the values of c_1 and λ_2, based on the initial conditions. Namely,

$$0 = E_g[(X-t)] = \int (x-t)c_1 e^{\lambda_2(x-t)} f(x)dx$$

which implies that

$$t = \frac{\int xe^{\lambda_2 x} f(x)dx}{\int e^{\lambda_2 x} f_X(x)dx} = K^{'}(\lambda_2).$$

Hence, $\lambda_2 = \upsilon(t)$. In addition

$$1 = \int g(x)dx = c_1 \int e^{\lambda_2(x-t)} f(x)dx$$

implies that

$$1/c_1 = e^{-t\upsilon(t)} \int e^{\upsilon(t)x} f_X(x)dx = e^{-t\upsilon(t)+K(\upsilon(t))} = 1/C(t).$$

Hence

$$g(x) = C(t)\exp\{\lambda_2(x-t)\}f(x) = C(t)\exp\{\upsilon(t)(x-t)\}f(x).$$

Since $C(t) \geq, 0$ the condition $g(x) \geq 0$ is indeed satisfied and it follows that $g(x) \equiv h_t(x)$. Thus, the conjugate density h_t is the solution to the Kullback-Leibler optimization problem.

The same arguments apply to prove the more general optimization problem (32), where the solution is given by (28).