Codivergences and information matrices

Alexis Derumigny[1] · **Johannes Schmidt-Hieber[2]**

Abstract

We propose a new concept of codivergence, which quantifies the similarity between two probability measures P_1, P_2 relative to a reference probability measure P_0. In the neighborhood of the reference measure P_0, a codivergence behaves like an inner product between the measures $P_1 - P_0$ and $P_2 - P_0$. Codivergences of covariance-type and correlation-type are introduced and studied with a focus on two specific correlation-type codivergences, the χ^2-codivergence and the Hellinger codivergence. We derive explicit expressions for several common parametric families of probability distributions. For a codivergence, we introduce moreover the divergence matrix as an analogue of the Gram matrix. It is shown that the χ^2-divergence matrix satisfies a data-processing inequality.

Keywords Divergence · Chi-square divergence · Hellinger affinity · Gram matrix

Mathematics Subject Classification 62B11 · 46E27 · 15A63

1 Introduction

One of the objectives of information geometry is to measure distances or angles in statistical spaces, usually for parametric models. This is often done by the use of a divergence, generating a Riemannian manifold structure on the considered space of distributions, see [1, 3, 19]. Divergences between probability measures quantify a certain notion of difference between them. Divergences are in general not symmetric, as

✉ Alexis Derumigny
a.f.f.derumigny@tudelft.nl

Johannes Schmidt-Hieber
a.j.schmidt-hieber@utwente.nl

[1] Department of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

[2] University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands

opposed to distances. Famous examples of divergences include the Kullback–Leibler divergence, the χ^2-divergence, and the Hellinger distance.

In this article, we are interested in defining a local notion of inner product between two probability measures in the neighborhood of a given reference probability measure P_0. This allows us to identify different directions relative to P_0, and to give some meaning to the "angle" between these directions. Contrary to most of the previous work on finite-dimensional Riemannian manifolds spanned by specific parametric statistical models, we do not require any parametric restrictions on the considered probability measures.

Motivation and application of our approach is the recently considered generic framework to derive lower bounds for the trade-off between bias and variance in nonparametric statistical models [9]. The key ingredient in this lower bound strategy are so-called change of expectation inequalities that relate the change of the expected value of a random variable with respect to different distributions to the variance and also involve the divergence matrices examined in this work. Another possible area of application are the lower bounds for statistical query algorithms, see e.g. [10, 11]

Regarding work on infinite-dimensional information geometry, [5, 20] studied the manifold generated by all probability densities connected to a given probability density. [21] reviews a more general theory on infinite-dimensional statistical manifolds given a reference density. Another line of work [12, 16–18] seeks to define infinite-dimensional manifolds, with applications to Bayesian estimation and the choice of priors. [24] consider different possible structures on the set of probability densities on [0, 1].

The article is structured as follows. In Sect. 2, we introduce a general concept of codivergence, study specific properties of codivergences on the space of probability measures, and discuss specific (classes of) codivergences. Section 3 considers the construction of divergence matrices from a given codivergence. Section 4 is devoted to the data-processing inequality that holds for the χ^2-divergence matrix introduced in Sect. 3, thereby generalizing the usual data-processing inequality for the χ^2-divergence. Section 5 provides derivations of explicit expressions for a class of codivergences applied to common parametric models. Elementary facts on ranks from linear algebra are collected in Sect. 6.

Notation: If P is a probability measure and $\mathbf{X}$ a random vector, we write $E_P[\mathbf{X}]$ and $\mathrm{Cov}_P(\mathbf{X})$ for the expectation vector and covariance matrix with respect to P, respectively.

2 Codivergences

2.1 Abstract framework and definition

We start by recalling the definition of a divergence [1, Definition 1.1]. This definition is situated within the framework of a d-dimensional differentiable manifold $\mathcal{X}$ with an atlas (U_i, φ_i). Formally, this means that the (U_i) are an open cover of the topological space $\mathcal{X}$, and $\varphi_i : U_i \rightarrow \mathbb{R}^d$ are isomorphisms such that $\varphi_j \circ \varphi_i^{-1}$ is C^1 on $\varphi_i(U_i \cap U_j) \rightarrow \varphi_j(U_i \cap U_j)$ [14].

Definition 2.1 A divergence D on a d-dimensional differentiable manifold $\mathcal{X}$ is a function $\mathcal{X}^2 \to \mathbb{R}_+$ satisfying

(i) $\forall P, Q \in \mathcal{X}$, $D(P|Q) = 0$ if and only if $P = Q$.
(ii) For all $P \in \mathcal{X}$, for any chart (U, φ) with $P \in U$, there exists a matrix $G = G(P)$ such that for any $Q \in U$,

$$D(P|Q) = \frac{1}{2}\big(\varphi(Q) - \varphi(P)\big)^T G\big(\varphi(Q) - \varphi(P)\big) + O\Big(\|\varphi(Q) - \varphi(P)\|^3\Big). \quad (1)$$

The matrix $G = G(P)$ may depend on the choice of coordinates φ. For the most common divergences, G is symmetric, positive-definite and thus defines a scalar product on the tangent space at P. Whereas a divergence measures the similarity between two elements $P, Q \in \mathcal{X}$, we want to define codivergences measuring the angle $\sphericalangle P_1 P_0 P_2$ of $P_1, P_2 \in \mathcal{X}$ relative to $P_0 \in \mathcal{X}$.

Equation (1) states that the divergence $D(P|Q)$ is a quadratic form in terms of the local coordinates $\varphi(Q) \in \mathbb{R}^d$, whenever P and Q are close. Generalizing to the infinite-dimensional case requires to work with bilinear forms instead. Moreover for the infinite-dimensional setting, imposing an expansion of the form (1) in every possible direction around $P \in \mathcal{X}$ is restrictive. We therefore allow the quadratic expansion to hold in a possibly smaller bilinear expansion domain. Furthermore, we allow codivergences to attain the value $+\infty$. This is inspired by existing statistical divergences (such as χ^2- or Kullback–Leibler divergences) that can also take the value $+\infty$. Therefore, imposing an expansion of the form (1) globally may not be possible as the codivergence on the left-hand side of (1) may take the value $+\infty$ in some directions away from P, while the right-hand side of (1) is always finite.

We now provide the definition of a codivergence if $\mathcal{X}$ is a subset of a real vector space.

Definition 2.2 Let $\mathcal{X}$ and $(E_u)_{u \in \mathcal{X}}$ be a subset and a family of subspaces of a real vector space E, respectively. A function $(u, v, w) \in \mathcal{X}^3 \mapsto D(u|v, w) \in \mathbb{R} \cup \{+\infty\}$ defines a *codivergence* on $\mathcal{X}$ with *bilinear expansion domain E_u* at u, if for any $u, v, w \in \mathcal{X}$,

(i) $D(u|v, w) = D(u|w, v)$;
(ii) $D(u|v, v) \geq 0$, with equality if $u = v$;
(iii) there exists a bilinear map $\langle \cdot, \cdot \rangle_u$ defined on E_u, such that, for any $h, g \in E_u$ and for any scalars s, t in some sufficiently small open neighborhood of $(0, 0)$ (that may depend on h and g) with respect to the Euclidean topology in $\mathbb{R}^2$, we have $(u + th, u + sg) \in \mathcal{X}^2$, $D\big(u|u + th, u + sg\big) < +\infty$, and $D\big(u|u + th, u + sg\big) = ts\langle h, g \rangle_u + o(t^2 + s^2)$ as $(s, t) \to (0, 0)$.

The last part of the definition imposes that, locally around each u, the codivergence $(v, w) \mapsto D(u|v, w)$ is finite and behaves like a bilinear form in the centered variables $(v - u, w - u)$. As a consequence, for a given u, the mapping $(v, w) \mapsto D(u|v, w)$ is Gateaux-differentiable on $\mathcal{X}^2$ at (u, u) with Gateaux derivative 0 in every direction $(h, g) \in E_u^2$. Condition (iii) can moreover be understood as a second-order Taylor

expansion at (u, u) in the direction (h, g). The mapping $(v, w) \mapsto D(u|v, w)$ needs, however, not to be twice Gateaux-differentiable at (u, u) for (iii) to hold. This is analogous to usual counter-examples in analysis where a function may admit a second-order Taylor expansion at a given point without being twice differentiable. Nevertheless, if $D(u|u+th, u+sg)$ is twice differentiable in (t, s) at $(0, 0)$, then the partial derivative $\partial^2 D(u|u + th, u + sg)/\partial t \partial s$ at $(0, 0)$ must be equal to $\langle h, g \rangle_u$. We refer to [2] for a discussion on higher-order functional derivatives.

We also provide a definition of codivergence if $\mathcal{X}$ is a differentiable Banach manifold, see [4, 14] for an introduction to Banach manifolds. Let B be a Banach space and $\mathcal{X}$ be a Banach manifold modeled on B. This guarantees existence of a B-atlas (U_i, φ_i) with U_i an open cover of $\mathcal{X}$ and $\varphi_i : U_i \to B$ such that $\varphi_j \circ \varphi_i^{-1}$ is C^1 (with respect to the norm on B).

This generalization can be useful in the case where the space $\mathcal{X}$ is not flat. Indeed, part (iii) of Definition 2.2 imposes that for $h \in E_u$, we must have $u + th \in \mathcal{X}$ for t small enough. On the contrary, in the following definition we consider a more subtle case where the point u may be approached on a smooth curve (not necessarily affine), under the assumption that $\mathcal{X}$ is a B-manifold.

We first recall the construction of the tangent space via curves following [4, Definition 2.21] and [15, Section 2.1.1]: for a fixed $u \in \mathcal{X}$, let i be such that $u \in U_i$ and let $\mathscr{C}_u$ be the set of smooth curves c such that $c : [-1, 1] \to U_i$ and $c(0) = u$. We define an equivalence relation $\sim$ on $\mathscr{C}_u$ by $c_1 \sim c_2$ if for all smooth real-valued functions f on U_i, we have $(f \circ c_1)'(0) = (f \circ c_2)'(0)$. We define the tangent space at u as the quotient set $T_u := \mathscr{C}_u/\sim$, which can be given a vector space structure isomorphic to B.

We give a short outline of the main ideas to obtain this property. Let D denote the Fréchet differential operator. For $\overline{c} \in T_u$ and c a representative of the equivalence class $\overline{c}$, note that $\varphi_i \circ c : [-1, 1] \to B$ is differentiable (by assumption on c); the mapping $D(\varphi_i \circ c)(0)$ is linear from $\mathbb{R}$ to B and can therefore be identified with an element of B itself; this element $D(\varphi_i \circ c)(0)$ also does not depend on the representative c. This defines a mapping $\theta_u : T_u \mapsto B$ by $\theta_u(\overline{c}) := D(\varphi_i \circ c)(0)$. It can be shown that θ_u is bijective. Through its inverse θ_u^{-1} one can transport the vector space structure of B on T_u, making it a real vector space too.

Definition 2.3 Let $\mathcal{X}$ be a B-manifold. A function $(u, v, w) \in \mathcal{X}^3 \mapsto D(u|v, w) \in \mathbb{R} \cup \{+\infty\}$ defines a *codivergence* on $\mathcal{X}$ with *bilinear expansion domain* E_u at u, if for any $u, v, w \in \mathcal{X}$,

(i) $D(u|v, w) = D(u|w, v)$;
(ii) $D(u|v, v) \geq 0$, with equality if $u = v$;
(iii) E_u is a subspace of the tangent space T_u of $\mathcal{X}$ at u;
(iv) there exists a bilinear map $\langle \cdot, \cdot \rangle_u$ defined on E_u. For any $\overline{g}, \overline{h} \in E_u$, for any representatives g and h of the respective equivalence classes $\overline{g}$ and $\overline{h}$, and for any scalars s, t in some sufficiently small open neighborhood of $(0, 0)$ with respect to the Euclidean topology in $\mathbb{R}^2$ (the neighborhood may depend on the choice of the representatives g and h), we have $D(u|h(t), g(s)) < +\infty$, and $D(u|h(t), g(s)) = ts \langle \overline{h}, \overline{g} \rangle_u + o(t^2 + s^2)$ as $(s, t) \to (0, 0)$.

From a codivergence $D(u|v, w)$ that takes finite values on a finite-dimensional manifold and with bilinear expansions domains the tangent spaces, we can always construct a divergence by setting $v = w$. Then $D(u|v, v)$ behaves like a quadratic form in v whenever v is close to u.

If $\mathcal{X}$ is a B-manifold and a closed subspace of a vector space E, then the notions of codivergences in Definitions 2.2 and 2.3 coincide. This is because differentiable curves are, in first order, linear functions in a small enough neighborhood of 0.

For both definitions, a given space $\mathcal{X}$ and a given family of bilinear expansion domains $(E_u)_{u \in \mathcal{X}}$, the set of codivergences on $\mathcal{X}$ is a convex cone.

For an example covered by Definition 2.3 but not by Definition 2.2 assume that $\mathcal{X}$ is the unit circle. No codivergence can exist in the sense of Definition 2.2 with non-trivial bilinear expansion domains (E_u). An example of a codivergence on $\mathcal{X} = \{e^{i\theta}, \theta \in \mathbb{R}\}$ in the sense of Definition 2.3 is

$$
D(u|v, w) = \begin{cases} e^{(\theta_1 - \theta_0)(\theta_2 - \theta_0)} - 1, & \text{if } v, w \in U_u, \\ +\infty, & \text{else,} \end{cases}
$$

where $u, v, w \in \mathcal{X}^3$, $U_u := \{ue^{i\theta}, \theta \in (-\pi/2, \pi/2)\}$, $u = e^{i\theta_0}$, $v = e^{i\theta_1}$ and $w = e^{i\theta_2}$ for some $\theta_0 \in \mathbb{R}$, $\theta_1, \theta_2 \in [\theta_0 - \pi, \theta_0 + \pi)$. Such a representation of v and w always exists and is unique since $[\theta_0 - \pi, \theta_0 + \pi)$ is a half-open interval of length 2π. In this case, the tangent space T_u of the circle at any point $u = e^{i\theta_0}$ is diffeomorphic to $\mathbb{R}$ and we will use this identification (denoted by the symbol "$\simeq$"). Let $g, h \in T_u \simeq \mathbb{R}$ and assume $s, t \in \mathbb{R}$. Then $g(s) = ue^{igs} \in U_u$ for s small enough. Similarly, $h(t) = ue^{iht} \in U_u$ for t small enough. So, $D(u|h(t), g(s))$ is finite for all (s, t) in a small enough neighborhood of $(0, 0)$, and, whenever this is the case, we have

$$
D(u|h(t), g(s)) = D(u|ue^{iht}, ue^{igs}) = e^{htgs} - 1 = ts\langle h, g\rangle_u + o(t^2 + s^2),
$$

where $\langle h, g\rangle_u = gh$ is the local bilinear form (which in this example is independent of u) and the bilinear expansion domain can be taken to be $E_u = T_u \simeq \mathbb{R}$.

2.2 Codivergences on the space of probability measures

For the application to statistics, E is the space of all finite signed measures on a measurable space $(\mathcal{A}, \mathcal{B})$, and $\mathcal{X}$ is the space of all probability measures on $(\mathcal{A}, \mathcal{B})$. Probability measures form a convex subset of all signed measures E. Since E is a vector space, the natural definition of a codivergence on $\mathcal{X}$ is Definition 2.2. A visual representation of such a codivergence is provided in Fig. 1.

In a next step, we characterize the bilinear expansion domains of a codivergence for $\mathcal{X}$ the space of probability measures. Given a probability measure $P_0 \in \mathcal{X}$, we say that a function $h : \mathcal{X} \to \mathbb{R}$ is P_0-essentially bounded by a constant $C > 0$ if $P_0(\{x \in \mathcal{A} : |h(x)| \leq C\}) = 1$ and define $\operatorname{ess\,sup}_{P_0} |h| := \inf\{C > 0 : |h| \text{ is } P_0\text{-essentially bounded by } C\}$. We will show that
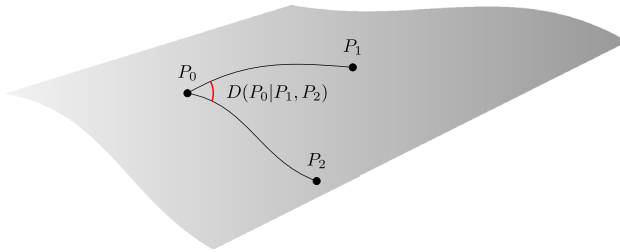
Fig. 1 The codivergence between P_1 and P_2 at P_0 measures the position of P_1 and P_2 relative to P_0

$$\mathcal{M}_{P_0} := \left\{ \mu \in E : \mu \ll P_0, \int d\mu = 0, \operatorname{ess\,sup}_{P_0} \left| \frac{d\mu}{dP_0} \right| < +\infty \right\}$$

is the largest bilinear expansion domain that any codivergence on $\mathcal{X}$ can have at P_0. The rationale is that $P_0 + t\mu$ is otherwise not a probability measure. Indeed if $\mu \in \mathcal{M}_{P_0}$ has a density h with respect to P_0, then the P_0-density $1 + th$ is non-negative for given $t > 0$ if and only if h is larger than $-1/t$. Conversely, the density $1 - th$ is non-negative for given $t < 0$ if and only if h is smaller than $1/t$. This gives a link between a bound on $h = d\mu/dP_0$ and the non-negativity of the probability measure $P_0 + t\mu$.

For every measurable set A,

$$(P_0 + t\mu)(A) = \int_{x \in A} d(P_0 + t\mu)(x) = \int_{x \in A} \big(1 + th(x)\big) dP_0(x). \tag{2}$$

The value of an integral is unchanged if the function to be integrated is modified on a P_0-null set. Therefore we only need the function $1 + th$ to be positive P_0-almost everywhere for $P_0 + t\mu$ to be a positive measure.

Proposition 2.4 *For any codivergence D on the space of probability measures $\mathcal{X}$, the bilinear expansion domain of D at any probability measure $P_0 \in \mathcal{X}$ must be included in $\mathcal{M}_{P_0}$. Furthermore, every $\mu \in \mathcal{M}_{P_0}$ has a density $d\mu/dP_0$ with respect to P_0 such that* $\operatorname{ess\,sup}_{P_0} |d\mu/dP_0| = 1/a_*$ *with* $a_* := \sup\{a > 0 : P_0 + t\mu \in \mathcal{X} \text{ for all } t \in [-a, a]\} \in (0, +\infty]$ *and the convention* $1/+\infty = 0$.

Proof of Proposition 2.4 We begin by proving the first part. Let μ be a finite signed measure belonging to the bilinear expansion domain at P_0 of some codivergence D on the space of probability measures $\mathcal{X}$. For $a > 0$, we write $\mu \in R(a)$ if and only if $P_0 + t\mu \in \mathcal{X}$ for all $-a \le t \le a$. Since μ belongs to the bilinear expansion domain of D at P_0, Definition 2.2(iii) implies existence of an open neighborhood T of 0 such that for any $t \in T$, $P_0 + t\mu \in \mathcal{X}$. Therefore, there exists $a > 0$ with $\mu \in R(a)$.

We now show that $\mu \in R(a)$, for some $a > 0$, implies $\mu \ll P_0$. The proof relies on the Jordan decomposition theorem for finite signed measures (e.g. Corollary 4.1.6 in [7]). It states that every finite signed measure μ on a measurable space $(\mathcal{A}, \mathcal{B})$ can be decomposed as

$$\mu = \alpha_+ \mu_+ - \alpha_- \mu_-, \quad \text{with } \alpha_+, \alpha_- \ge 0, \tag{3}$$

and μ_-, μ_+ orthogonal probability measures on $(\mathcal{A}, \mathcal{B})$. By the Lebesgue decomposition theorem (see Theorem 4.3.2 in [7]), μ can always be decomposed as $\mu = \mu_A + \mu_S$, where μ_A is a signed measure that is absolutely continuous with respect to P_0, μ_S is a signed measure that is singular with respect to P_0, and μ_A and μ_S are orthogonal. By the Jordan decomposition (3), we decompose the signed measure $\mu_S = \alpha_+ \mu_{S,+} - \alpha_- \mu_{S,-}$ into its positive and negative part $\mu_{S,+}$ and $\mu_{S,-}$. These two measures are orthogonal and $\alpha_+, \alpha_- \geq 0$. Then, $P_0 + a\mu = P_0 + a\mu_A + a\alpha_+ \mu_{S,+} - a\alpha_- \mu_{S,-}$ can be a probability measure only if $\alpha_- = 0$. This is because we can find a set U such that $P_0(U) = \mu_A(U) = \mu_{S,+}(U) = 0$ and $\mu_{S,-}(U) = 1$. Therefore $(P_0 + a\mu)(U) = -a\alpha_- \mu_{S,-}(U) = -a\alpha_- \leq 0$. In the same way, $P_0 - a\mu$ can be a probability measure only if $\alpha_+ = 0$. Therefore, if $\mu \in R(a)$ for some $a > 0$, then $\alpha_+ = \alpha_- = 0$, and $\mu = \mu_A$ is absolutely continuous with respect to P_0.

Let h be the density of μ with respect to P_0. Then

$$\frac{d(P_0 + t\mu)}{dP_0} = 1 + t\frac{d\mu}{dP_0} = 1 + th.$$

Note that $P_0 + t\mu$ is a signed measure integrating to 1 if and only if $\int d\mu = \int h\, dP_0 = 0$.

We now show that, for any $a > 0$, $\mu \in R(a)$ implies $\operatorname{ess\,sup}_{P_0} |h| \leq 1/a$. If $\mu \in R(a)$, then for any $A \in \mathcal{B}$, $(P_0 + a\mu)(A) \geq 0$ and $(P_0 - a\mu)(A) \geq 0$. Let us define the sets $A_+ := \{x \in \mathcal{A} : 1 + ah(x) \geq 0\}$ and $A_- := \{x \in \mathcal{A} : 1 - ah(x) \geq 0\}$. Let A^C denote the complement of a set A. We have $(P_0 + a\mu)(A_+^C) = \int_{A_+^C} 1 + ah(x)\, dP_X(x) \leq 0$ since this is the integral of a negative function. Therefore $P_0(A_+^C) = 0$ and then $P_0(A_+) = 1$. Similarly, $(P_0 + a\mu)(A_-^C) = \int_{A_-^C} 1 - ah(x)\, dP_X(x) \leq 0$. Hence, $P_0(A_-^C) = 0$ and $P_0(A_-) = 1$.

Therefore, $P_0(A_+ \cap A_-) = 1$. This means that for P_0-almost every $x \in \mathcal{A}$, $1 + ah(x) \geq 0$ and $1 - ah(x) \geq 0$. Therefore, for P_0-almost every $x \in \mathcal{A}$, $|h(x)| \leq 1/a$. Therefore, h is P_0-essentially bounded by $C := 1/a$. We have finally shown that $\mu \in R(a)$ implies $\operatorname{ess\,sup}_{P_0} |h| \leq 1/a$ and $\mu \in \mathcal{M}_{P_0}$, proving the first part of Proposition 2.4.

Conversely, note that $\mu \in \mathcal{M}_{P_0}$ is a sufficient condition for $P_0 + t\mu$ to be a probability measure for all t in a sufficiently small open neighborhood of 0.

We now show the second part of Proposition 2.4. Remember that $a_* := \sup\{a > 0 : \mu \in R(a)\} \in (0, +\infty]$. Let $(a_n)_{n \in \mathbb{N}}$ be an increasing sequence of real numbers strictly smaller than a_* and converging to a_*. For every positive integer n, we have $\mu \in R(a_n)$. Therefore, by the previous reasoning, $\operatorname{ess\,sup}_{P_0} |h| \leq 1/a_n$, meaning that $P_0(\{x \in \mathcal{A} : |h(x)| \leq 1/a_n\}) = 1$. By a union bound, we obtain $P_0(\cap_{n \geq 0}\{x \in \mathcal{A} : |h(x)| \leq 1/a_n\}) = 1$. Therefore $P_0(\{x \in \mathcal{A} : |h(x)| \leq 1/a_*\}) = 1$, and by definition $\operatorname{ess\,sup}_{P_0} |h| \leq 1/a_*$.

We now show the reverse version of this inequality. Let $C > \operatorname{ess\,sup}_{P_0} |h|$. Then $P_0(\{x \in \mathcal{A} : |h(x)| \leq C\}) = 1$. Hence, for any $t \in [-1/C, 1/C]$, and for P_0-almost every x, $-1 \leq th(x) \leq 1$. Consequently, for any $t \in [-1/C, 1/C]$, and for P_0-almost every x, $1 + th(x) \geq 0$ and $1 - th(x) \geq 0$. For any $t \in [-1/C, 1/C]$, $P_0 + t\mu$ is a finite signed measure with a density that is non-negative P_0-almost everywhere and

integrates to 1. These are sufficient conditions for $P_0 + t\mu$ to be a probability measure on $\mathcal{A}$, proving $\mu \in R(1/C)$. Therefore, $1/C \leq a_*$ and thus $1/a_* \leq C$. This holds for any $C > 0$ such that $C > \operatorname{ess\,sup}_{P_0} |h|$, proving that $1/a_* \leq \operatorname{ess\,sup}_{P_0} |h|$. Together with the inequality $\operatorname{ess\,sup}_{P_0} |h| \leq 1/a_*$, the claim $1/a_* = \operatorname{ess\,sup}_{P_0} |h|$ follows. $\qquad \square$

2.3 Examples of codivergences

For any real $a \geq 0$, set $a/0 := +\infty$. For a non-negative function $\phi : [0, \infty) \to [0, \infty)$, we can define two codivergences. The first one will be referred to as covariance-type codivergence between three probability measures P_0, P_1, P_2 and is defined by

$$
V_\phi(P_0|P_1, P_2) := \int \phi\left(\frac{dP_1}{dP_0}\right)\phi\left(\frac{dP_2}{dP_0}\right)dP_0 - \int \phi\left(\frac{dP_1}{dP_0}\right)dP_0 \int \phi\left(\frac{dP_2}{dP_0}\right)dP_0,
\tag{4}
$$

and the second one will be called correlation-type codivergence and is defined by

$$
R_\phi(P_0|P_1, P_2) := \frac{V_\phi(P_0|P_1, P_2)}{\int \phi\left(\frac{dP_1}{dP_0}\right)dP_0 \int \phi\left(\frac{dP_2}{dP_0}\right)dP_0} = \frac{\int \phi\left(\frac{dP_1}{dP_0}\right)\phi\left(\frac{dP_2}{dP_0}\right)dP_0}{\int \phi\left(\frac{dP_1}{dP_0}\right)dP_0 \int \phi\left(\frac{dP_2}{dP_0}\right)dP_0} - 1,
\tag{5}
$$

if $P_1, P_2 \ll P_0$. Otherwise, we define both $V_\phi(P_0|P_1, P_2)$ and $R_\phi(P_0|P_1, P_2)$ to be equal to $+\infty$.

Obviously, both codivergences V_ϕ and R_ϕ are symmetric in P_1 and P_2. By Jensen's inequality we see that $V_\phi(P_0|P_1, P_1) \geq 0$ and $R_\phi(P_0|P_1, P_1) \geq 0$. If $\phi(1) = 0$, then $R_\phi(P_0|P_0, P_0) = +\infty$. For $\phi(1) > 0$, the functions ϕ and $t\phi$ with positive scalar t give the same codivergence R_ϕ and simply rescale V_ϕ. Without loss of generality, we therefore can (and will) assume that $\phi(1) = 1$.

We say that a function f admits a second order Taylor expansion around 1 if $f(1 + y) = f(1) + yf'(1) + \frac{y^2}{2}f''(1) + o(y^2)$ for all y in an open neighborhood of zero. The following proposition is proved in Section A.

Proposition 2.5 *Assume that $\phi(1) = 1$ and ϕ admit a second order Taylor expansion around 1. Then the V_ϕ and the R_ϕ codivergences are codivergences in the sense of Definition 2.2 with bilinear expansion domains $\mathcal{M}_{P_0}$ and bilinear maps $\phi'(1)^2 \langle \mu, \widetilde{\mu} \rangle_{P_0}$, where*

$$
\langle \mu, \widetilde{\mu} \rangle_{P_0} := \int \frac{d\mu}{dP_0} d\widetilde{\mu} = \int hg \, dP_0, \quad \text{with } h = \frac{d\mu}{dP_0} \text{ and } g = \frac{d\widetilde{\mu}}{dP_0}.
$$

If ν is a measure dominating P_0, the bilinear map can be written as

$$
\langle \mu, \widetilde{\mu} \rangle_{P_0} = \int \frac{hg}{p_0} d\nu, \quad \text{for the densities } h = \frac{d\mu}{d\nu}, \ g = \frac{d\widetilde{\mu}}{d\nu}, \ p_0 = \frac{dP_0}{d\nu}.
\tag{6}
$$

A consequence of Proposition 2.5 is the following: locally, all V_ϕ and R_ϕ codivergences (that satisfies the regularity conditions) define the same structure. This scalar product is the nonparametric Fisher information metric. The name originates from the identity [12, Equation (8)]

$$[I(\theta)]_{ij} = \int \frac{p_i(x|\theta)p_j(x|\theta)}{p(x|\theta)} d\nu(x), \tag{7}$$

where $[I(\theta)]_{ij}$ is the (i, j)-th entry of the Fisher information matrix for a parametric model of ν-densities $p(\cdot|\theta)$ indexed by a finite dimensional parameter θ and $p_i(x|\theta) := \partial p(x|\theta)/\partial\theta_i$. Eqs. (7) and (6) have the same structure. One of the earliest reference to the nonparametric Fisher information metric is [8]. The concept has been applied in several frameworks, such as computer vision [24] or shape data analysis [25]. The geometry of the nonparametric Fisher information metric has been studied by [6, 12] in the context of Bayesian inference.

An interesting subclass of codivergences is obtained by choosing $\phi_\alpha(x) = x^\alpha$. To ease the notation, we set

$$V_\alpha := V_{\phi_\alpha} \quad \text{and} \quad R_\alpha := R_{\phi_\alpha}. \tag{8}$$

Although the resulting codivergences seem related to the well-known Rényi divergence $(1-\alpha)^{-1}\log(\int p(x)^\alpha q(x)^{1-\alpha}d\nu(x))$ between probability measures P and Q with densities p and q [23], the term $\int(p_1(x)p_2(x))^\alpha p_0(x)^{1-2\alpha}d\nu(x)$ occurring in the definitions of V_α and R_α is of a different nature.

In the case $\alpha = 1$, that is, $\phi(x) = x$, both notions of codivergence agree. Denoting by p_0, p_1, p_2 the respective ν-densities of P_0, P_1, P_2, where ν is a measure dominating P_0, the corresponding codivergence

$$\chi^2(P_0|P_1, P_2) := \begin{cases} \int \frac{dP_1}{dP_0}dP_2 - 1 = \int \frac{p_1 p_2}{p_0}d\nu - 1, & \text{if } P_1 \ll P_0 \text{ and } P_2 \ll P_0, \\ +\infty, & \text{else,} \end{cases}$$

will be called χ^2-codivergence. The (usual) χ^2-divergence is defined as $\chi^2(P, Q) := \int(dP/dQ - 1)^2 dQ = \int(dP/dQ)^2 dQ - 1$, if P is dominated by Q and $+\infty$ otherwise. Therefore, the χ^2-codivergence $\chi^2(P_0|P_1, P_1)$ coincides with the usual χ^2-divergence $\chi^2(P_1, P_0)$ for any P_0 and P_1.

Another interesting codivergence is R_α with $\alpha = 1/2$. The resulting codivergence

$$\rho(P_0|P_1, P_2) := \frac{\int \sqrt{p_1 p_2}d\nu}{\int \sqrt{p_1 p_0}d\nu \int \sqrt{p_2 p_0}d\nu} - 1, \tag{9}$$

is called Hellinger codivergence. We can (and will) define the Hellinger codivergence as $\int \sqrt{p_1 p_2}d\nu/(\int \sqrt{p_1 p_0}d\nu \int \sqrt{p_2 p_0}d\nu)$ whenever the denominator is positive. This is considerably weaker than P_1, $P_2 \ll P_0$, as it is only required that the support of p_0 intersects with non-zero ν-mass the support of p_1 and the support of p_2. Note

that $\rho(P_0|P_1, P_2)$ is independent of the choice of the dominating measure ν (and potentially $+\infty$ if the denominator is 0).

The name Hellinger codivergence is motivated by the representation

$$\rho(P_0|P_1, P_2) = \frac{\alpha(P_1, P_2)}{\alpha(P_0, P_1)\alpha(P_0, P_2)} - 1,$$

where $\alpha(P, Q) := \int \sqrt{pq}\,d\nu$ is the Hellinger affinity between two positive measures P, Q with densities p, q taken with respect to a common dominating measure.

The χ^2- and Hellinger codivergence are of interest as they can be used to control changes of expectation between probability measures, see Section 2.2 of [9].

We always have

$$\rho(P_0|P_1, P_1) \le \chi^2(P_0|P_1, P_1). \tag{10}$$

To see this, observe that Hölder's inequality with $p = 3/2$ and $q = 3$ gives for any non-negative function f, $1 = \int p_1 \le (\int f^{3/2}p_1)^{2/3}(\int f^{-3}p_1)^{1/3}$. The choice $f = (p_0/p_1)^{1/3}$ yields $1 \le (\int \sqrt{p_1 p_0})^2 \int p_1^2/p_0$. Therefore $1/(\int \sqrt{p_1 p_0})^2 \le \int p_1^2/p_0$. Subtracting one on each side of this expression yields (10).

Proposition 2.5 implies that the χ^2-codivergence and the Hellinger codivergence are codivergences with respective bilinear maps $\langle \mu, \widetilde{\mu} \rangle_{P_0}$ for the χ^2-codivergence and $\langle \mu, \widetilde{\mu} \rangle_{P_0}/4$ for the Hellinger codivergence.

For the Hellinger codivergence, the expansion in Proposition 2.5 can be generalized. Assume that P_0 is dominated by some positive measure ν. Define $\mathrm{Supp}(\mu) := \{x \in \mathcal{A} : d\mu/d\nu(x) \ne 0\}$ for any signed measure μ dominated by ν. If μ_1 and μ_2 are signed measures dominated by ν such that (i) $\mathrm{Supp}(\mu_i) \cap \mathrm{Supp}(P_0)$ has a positive ν-measure, and (ii) their densities h_i are positive on $\mathrm{Supp}(\mu_i) \backslash \mathrm{Supp}(P_0)$, then

$$\rho(P_0|P_0 + t\mu_1, P_0 + s\mu_2) = \sqrt{ts} \int_{\mathrm{Supp}(P_0)^C} \sqrt{h_1 h_2}\,d\nu$$
$$+ ts \int_{\mathrm{Supp}(P_0)} \frac{h_1 h_2}{2p_0}\,d\nu + o(t^2 + s^2). \tag{11}$$

Compared to Definition 2.2 (iii), there is thus an additional term for probability measures that have mass outside of the support of P_0. Consequently, this expansion cannot be linked to one local bilinear form and the mapping $(t, s) \in \mathbb{R}_+^2 \mapsto \rho(P_0|P_0 + t\mu_1, P_0 + s\mu_2)$ is not differentiable at $(0, 0)$. This is in line with Proposition 2.4: for perturbations μ that do not belong to $\mathcal{M}_{P_0}$, the measures $P_0 + t\mu$ cannot be probability measures for all t in any open neighborhood of 0.

The R_α codivergences admit convenient expressions for product measures and for exponential families. The first proposition is proved in Sect. A.2.

Proposition 2.6 *Let $P_{j\ell}$ be probability measures for any $j = 0, 1, 2$ and for any $\ell = 1, \dots, d$ satisfying $P_{1\ell}, P_{2\ell} \ll P_{0\ell}$. Then*

Table 1 Closed-form expressions for the R_α codivergence for some parametric distributions

Distribution	$R_\alpha(P_0|P_1, P_2)$
$P_j = \mathcal{N}(\theta_j, \sigma^2 I_d), \theta_j \in \mathbb{R}^d, \sigma > 0$	$\exp\left(\alpha^2 \dfrac{\langle \theta_1 - \theta_0, \theta_2 - \theta_0 \rangle}{\sigma^2}\right) - 1$
$P_j = \otimes_{\ell=1}^d \text{Pois}(\lambda_{j\ell}), \lambda_{j\ell} > 0$	$\exp\left(\sum_{\ell=1}^d \lambda_{0\ell}^{1-2\alpha}(\lambda_{1\ell}^\alpha - \lambda_{0\ell}^\alpha)(\lambda_{2\ell}^\alpha - \lambda_{0\ell}^\alpha)\right) - 1$
$P_j = \otimes_{\ell=1}^d \text{Exp}(\beta_{j\ell}), \beta_{j\ell} > 0$	$\displaystyle\prod_{\ell=1}^d \dfrac{\big(\beta_{0\ell} + \alpha(\beta_{1\ell} - \beta_{0\ell})\big)\big(\beta_{0\ell} + \alpha(\beta_{2\ell} - \beta_{0\ell})\big)}{\beta_{0\ell}\big(\beta_{0\ell} + \alpha(\beta_{1\ell} + \beta_{2\ell} - 2\beta_{0\ell})\big)} - 1,$ if all the involved quantities are positive, and $+\infty$ else
$P_j = \otimes_{\ell=1}^d \text{Ber}(\theta_{j\ell}), \theta_{j\ell} \in (0,1)$	$\displaystyle\prod_{\ell=1}^d \dfrac{\theta_{0\ell}^{1-2\alpha}\theta_{1\ell}^\alpha\theta_{2\ell}^\alpha + (1-\theta_{0\ell})^{1-2\alpha}(1-\theta_{1\ell})^\alpha(1-\theta_{2\ell})^\alpha}{r(\theta_{0\ell}, \theta_{1\ell})r(\theta_{0\ell}, \theta_{2\ell})} - 1,$ $r(\theta_0, \theta_1) := \theta_0^{1-\alpha}\theta_1^\alpha + (1-\theta_0)^{1-\alpha}(1-\theta_1)^\alpha$

Proofs can be found in Sect. 5

$$R_\alpha\left(\bigotimes_{\ell=1}^d P_{0\ell} \,\Big|\, \bigotimes_{\ell=1}^d P_{1\ell}\bigotimes_{\ell=1}^d P_{2\ell}\right) = \prod_{\ell=1}^d \left(R_\alpha(P_{0\ell}|P_{1\ell}, P_{2\ell}) + 1\right) - 1.$$

Proposition 2.7 *Let Θ be a subset of a real vector space and let $(P_\theta : \theta \in \Theta)$ be an exponential family with ν-densities $p_\theta(x) = h(x)\exp(\theta^\top T(x) - A(\theta))$ for some dominating measure ν. Then, for any $\theta_0, \theta_1, \theta_2 \in \Theta$ satisfying*

$$\theta_0 + \alpha(\theta_1 + \theta_2 - 2\theta_0), \theta_0 + \alpha(\theta_1 - \theta_0), \theta_0 + \alpha(\theta_2 - \theta_0) \in \Theta, \qquad (12)$$

we have

$$R_\alpha(P_{\theta_0}|P_{\theta_1}, P_{\theta_2}) = \exp\Big(A\big(\theta_0 + \alpha(\theta_1 + \theta_2 - 2\theta_0)\big) - A\big(\theta_0 + \alpha(\theta_1 - \theta_0)\big)$$
$$- A\big(\theta_0 + \alpha(\theta_2 - \theta_0)\big) + A(\theta_0)\Big) - 1.$$

This proposition is proved in Sect. A.3. (12) is satisfied if Θ is a vector space or if $0 < \alpha \leq 1$ and Θ is convex. In the case of the Gamma distribution the parameter space is $\Theta = (-1, +\infty) \times (-\mathbb{R})$ and in this case the constraints in (12) are sufficient and necessary for the statement of Proposition 2.7 to hold, see Sect. 5.4 for details.

For the most common families of distributions, closed-form expressions for the $R_\alpha(P_{\theta_0}|P_{\theta_1}, P_{\theta_2})$ codivergences are reported in Table 1. Derivations for these expressions are given in Sect. 5. This section also contains expressions for the Gamma distribution. As mentioned before, these codivergences quantify to which extent the measures P_1 and P_2 represent different directions around P_0. The explicit formulas show this in terms of the parameters and reveal significant similarity between the different families. For the multivariate normal distribution the R_α codivergence vanishes if and only if the vectors $\theta_1 - \theta_0$ and $\theta_2 - \theta_0$ are orthogonal.

3 Divergence matrices

Definition 3.1 Let $M \geq 1$. For a given codivergence $D(\cdot|\cdot, \cdot)$ on a space $\mathcal{X} \subset E$ and $u, v_1, \ldots, v_M$ elements of $\mathcal{X}$, we define the divergence matrix $D(u|v_1, \ldots, v_M)$ as the $M \times M$ matrix with (j, k)-th entry $D(u|v_1, \ldots, v_M)_{j,k} := D(u|v_j, v_k)$, for all $1 \leq j, k \leq M$.

If $v_1, \ldots, v_M$ are all in a neighborhood of u, the divergence matrix D can be related to the Gram matrix of the bilinear form $\langle \cdot, \cdot \rangle_u$. Formally, for $\mathbf{t} = (t_1, \ldots, t_M) \in \mathbb{R}^M$ such that for any $i = 1, \ldots, M$, $u + t_i h_i \in \mathcal{X}$, we have

$$D(u|u + t_1 h_1, \ldots, u + t_M h_M) = \mathbf{t} \mathbb{G}_u \mathbf{t}^\top + o(\|\mathbf{t}\|^2),$$

with Gram matrix $\mathbb{G}_u := (\langle h_i, h_j \rangle_u)_{1 \leq i, j \leq M}$.

Based on the codivergences $V_\phi(P_0|P_1, P_2)$, $R_\phi(P_0|P_1, P_2)$, one can now define corresponding $M \times M$ divergence matrices with (j, k)-th entry

$$
V_\phi(P_0|P_1, \ldots, P_M)_{j,k} := V_\phi(P_0|P_j, P_k)
$$
$$
= \int \phi\left(\frac{dP_j}{dP_0}\right)\phi\left(\frac{dP_k}{dP_0}\right)dP_0 - \int \phi\left(\frac{dP_j}{dP_0}\right)dP_0 \int \phi\left(\frac{dP_k}{dP_0}\right)dP_0, \qquad (13)
$$

and

$$
R_\phi(P_0|P_1, \ldots, P_M)_{j,k} := R_\phi(P_0|P_j, P_k) = \frac{\int \phi\left(\frac{dP_j}{dP_0}\right)\phi\left(\frac{dP_k}{dP_0}\right)dP_0}{\int \phi\left(\frac{dP_j}{dP_0}\right)dP_0 \int \phi\left(\frac{dP_k}{dP_0}\right)dP_0} - 1,
$$
$$
\qquad (14)
$$

provided that $P_1, \ldots, P_M \ll P_0$. The codivergence matrices are linked by the relationship

$$R_\phi(P_0|P_1, \ldots, P_M) = D \cdot V_\phi(P_0|P_1, \ldots, P_M) \cdot D, \qquad (15)$$

where D denotes the $M \times M$ diagonal matrix with j-th diagonal entry $1/\int \phi\left(\frac{dP_j}{dP_0}\right)dP_0$, $j = 1, \ldots, M$.

Similarly as $\text{Cov}(X_1, X_2)$ can denote either the covariance between the random variables X_1 and X_2 or the 2×2 covariance matrix of the random vector (X_1, X_2), the expressions $V_\phi(P_0|P_1, P_2)$ and $R_\phi(P_0|P_1, P_2)$ can also denote either codivergences or 2×2 divergence matrices. Within the context, it is always clear which of the two interpretations is meant.

The divergence matrices with function $\phi_\alpha(x) := x^\alpha$ are denoted by $V_\alpha(P_0|P_1, \ldots, P_M)$ and $R_\alpha(P_0|P_1, \ldots, P_M)$. Similarly, the χ^2-divergence matrix $\chi^2(P_0|P_1, \ldots, P_M)$

and the Hellinger affinity matrix $\rho(P_0|P_1, \ldots, P_M)$ are the $M \times M$ divergence matrices of the χ^2-codivergence and the Hellinger codivergence with (j, k)-th entry

$$\chi^2(P_0|P_1, \ldots, P_M)_{j,k} := \int \frac{dP_j}{dP_0} dP_k - 1, \quad \text{and,}$$

$$\rho(P_0|P_1, \ldots, P_M)_{j,k} := \frac{\int \sqrt{p_j p_k} d\nu}{\int \sqrt{p_j p_0} d\nu \int \sqrt{p_k p_0} d\nu} - 1,$$

for all $1 \leq j, k \leq M$. As in the previous section, the condition for finiteness of the Hellinger codivergence matrix is weaker than for general R_ϕ and V_ϕ codivergences. Instead of domination $P_1, \ldots, P_M \ll P_0$, it is only required that the integrals $\int \sqrt{p_j p_0} d\nu$ are positive, for some dominating measure ν and $p_j := dP_j/d\nu$. By (6), the local Gram matrix of the χ^2-divergence matrix at a distribution P_0 is $\mathbb{G}_{P_0} := \left[\int \frac{h_i h_j}{p_0} d\nu \right]_{1 \leq i, j \leq M}$, and the local Gram matrix of the Hellinger divergence matrix is $\mathbb{G}_{P_0}/4$.

Let $\Phi(X) := (\phi(dP_1/dP_0(X)), \ldots, \phi(dP_M/dP_0(X)))^\top$ denote the random vector containing the likelihood ratios of the M measures. Since $\text{Cov}(U, V) = E[UV] - E[U]E[V]$, we have

$$V_\phi(P_0|P_1, \ldots, P_M) = \text{Cov}_{P_0}\left(\Phi(X)\right), \tag{16}$$

where the covariance is computed with respect to the distribution P_0 as indicated by the subscript P_0. Moreover, we have

$$\mathbf{v}^\top V_\phi(P_0|P_1, \ldots, P_M)\mathbf{v} = \text{Var}_{P_0}\left(\mathbf{v}^\top \Phi(X)\right).$$

Applying Eq. (15) yields moreover

$$R_\phi(P_0|P_1, \ldots, P_M) = D \text{Cov}_{P_0}\left(\Phi(X)\right)D. \tag{17}$$

This shows that $V_\phi(P_0|P_1, \ldots, P_M)$ and $R_\phi(P_0|P_1, \ldots, P_M)$ can be interpreted as covariance matrices and are therefore symmetric and positive semi-definite. Applying the Taylor expansion to the likelihood ratios in the previous identities provides a direct way of recovering the local Gram matrix associated to the nonparametric Fisher information metric.

In a next step, we state a more specific identity for the χ^2-divergence matrix. To do so, we first extend the usual notion of the χ^2-divergence to the case where the first argument is a signed measure. Let μ be a finite signed measure and P be a probability measure defined on the same measurable space $(\Omega, \mathcal{A})$. We define the χ^2-divergence of μ and P by

$$\chi^2(\mu, P) := \begin{cases} \int \left(\frac{d\mu}{dP} - \mu(\Omega)\right)^2 dP, & \text{if } \mu \ll P, \\ +\infty & \text{else.} \end{cases} \tag{18}$$

Here, $d\mu/dP$ denotes the Radon–Nikodym derivative of the signed measured μ with respect to P (defined e.g. in Theorem 4.2.4 in [7]). This definition of $\chi^2(\mu, P)$ generalizes the case where μ is a probability measure and allows us to rewrite the χ^2-divergence matrix as

$$\mathbf{v}^\top \chi^2(P_0|P_1, \ldots, P_M)\mathbf{v} = \int \left(\sum_{j=1}^M \left(\frac{dP_j}{dP_0} - 1 \right) v_j \right)^2 dP_0 = \chi^2 \left(\sum_{j=1}^M v_j P_j, P_0 \right),$$
(19)

with $\sum_{j=1}^M v_j P_j$ the mixture (signed) measure of $P_1, \ldots, P_M$. Similarly, for the Hellinger divergence matrix it can be checked that

$$\mathbf{v}^\top \rho(P_0|P_1, \ldots, P_M)\mathbf{v} = \int \left(\sum_{j=1}^M \left(\frac{\sqrt{p_j}}{\int \sqrt{p_j p_0} dv} - \sqrt{p_0} \right) v_j \right)^2 dv.$$
(20)

Writing $\mathrm{Rank}(A)$ for the rank of a matrix A and $\mathrm{Rank}(x_1, \ldots, x_n)$ for the dimension of the linear span of n elements $x_1, \ldots, x_n$ in a vector space E, we will now derive an identity for the rank of divergence matrices.

Proposition 3.2 *Let $M \geq 1$, and let $P_0, P_1, \ldots, P_M$ be $(M + 1)$ probability distributions.*

(i) *Assume that $P_1, \ldots, P_M \ll P_0$. Then for any non-negative function $\phi : [0, \infty) \to [0, \infty)$ such that $\phi(1) = 1$, we have*

$$\mathrm{Rank}(R_\phi(P_0|P_1, \ldots, P_M)) = \mathrm{Rank}(V_\phi(P_0|P_1, \ldots, P_M))$$
$$= \mathrm{Rank}\left(1, \phi \circ \frac{dP_1}{dP_0}, \ldots, \phi \circ \frac{dP_M}{dP_0} \right) - 1,$$

where functions are considered as elements of the vector space $L^1(\mathcal{A}, \mathcal{B}, P_0)$, that is, linear independence is considered P_0-almost everywhere.

(ii) *Let v be a common dominating measure of $P_0, \ldots, P_M$. Assume that $\forall j = 1, \ldots, M, \int p_j p_0 dv > 0$ with $p_j := dP_j/dv$. Then we have*

$$\mathrm{Rank}(\rho(P_0|P_1, \ldots, P_M)) = \mathrm{Rank}(\sqrt{p_0}, \sqrt{p_1}, \ldots, \sqrt{p_M}) - 1,$$

where functions are considered as elements of the vector space $L^1(\mathcal{A}, \mathcal{B}, v)$.

Statement (ii) is not a consequence of (i) with $\phi(x) = x^{1/2}$. Indeed, (i) relies on likelihood ratios assuming that the measures $P_1, \ldots, P_M$ are dominated by P_0, while (ii) only requires that each of the probability measures $P_1, \ldots, P_M$ has a common support with P_0 of positive P_0-measure. The proof of (ii) exploits the specific property (20) of the Hellinger divergence.

Proposition [3.2] applied to $\phi(x) = x$ shows that whenever P_0 is a linear combination of $P_1, \ldots, P_M$, then $\text{Rank}(1, dP_1/dP_0, \ldots, dP_M/dP_0) < M + 1$ and $\text{Rank}(\chi^2(P_0|P_1, \ldots, P_M)) < M$, which means that the $\chi^2(P_0|P_1, \ldots, P_M)$ divergence matrix is singular. Similarly, whenever $\sqrt{p_0}$ is a linear combination of $\sqrt{p_1}, \ldots, \sqrt{p_M}$, the Hellinger divergence matrix is singular.

Proof of Proposition [3.2] We first prove (i). Since D is an invertible matrix, a direct consequence of Eq. (15) is $\text{Rank}(R_\phi(P_0|P_1, \ldots, P_M)) = \text{Rank}(V_\phi(P_0|P_1, \ldots, P_M))$. Applying Eq. (16) and then Lemma [6.3], we obtain that $r := \text{Rank}(V_\phi(P_0|P_1, \ldots, P_M)) = \text{Rank}(\text{Cov}_{P_0}(Z_1, \ldots, Z_M)) = \text{Rank}(Z_1 - E_{P_0}Z_1, \ldots, Z_M - E_{P_0}Z_M)$, where $Z_j := \phi(dP_j/dP_0(X))$ for $j = 1, \ldots, M$ and E_{P_0} denotes the expectation with respect to P_0. The random vectors $Z_1 - E_{P_0}Z_1, \ldots, Z_M - E_{P_0}Z_M$ are centered and therefore linearly independent of the (constant) random variable $Z_0 := 1 = \phi(dP_0/dP_0(X))$. Therefore,

$$\begin{aligned} r &= \text{Rank}(Z_1 - E_{P_0}Z_1, \ldots, Z_M - E_{P_0}Z_M) \\ &= \text{Rank}(1, Z_1 - E_{P_0}Z_1, \ldots, Z_M - E_{P_0}Z_M) - 1 \\ &= \text{Rank}(Z_0, Z_1, \ldots, Z_M) - 1. \end{aligned}$$

By Lemma [6.2], r is the highest integer such that there exists $i_1, \ldots, i_r \in \{0, \ldots, M\}$ with $(Z_{i_1}, \ldots, Z_{i_r})$ linearly independent random variables P_0-almost surely.

Using the definition of the Z_j and $X \sim P_0$, the random variables $\{Z_{i_1}, \ldots, Z_{i_r}\}$ are linearly independent P_0-almost surely if and only if $P_0\left(\sum_{j=1}^{r} a_j \phi(dP_{i_j}/dP_0(X)) = 0\right) = 1$ implies $a_0 = \ldots = a_r = 0$. This is the case if and only if the functions $\{\phi \circ dP_{i_1}/dP_0, \ldots, \phi \circ dP_{i_r}/dP_0\}$ are linearly independent P_0-almost everywhere, proving $\text{Rank}(Z_{i_1}, \ldots, Z_{i_r}) = \text{Rank}(\phi \circ dP_{i_1}/dP_0, \ldots, \phi \circ dP_{i_r}/dP_0)$.

Before proving (ii) in full generality, we first show that $\text{Rank}(\rho(P_0|P_1, \ldots, P_M)) = M$ if and only if all the $M + 1$ functions $\sqrt{p_0}, \ldots, \sqrt{p_M}$ are linearly independent ν-almost everywhere. The matrix is singular if and only if there exists a non-null vector v such that $\sum_{j=1}^{M} \frac{v_j \sqrt{p_j}}{\int \sqrt{p_j p_0} d\nu} = \sum_{j=1}^{M} v_j \sqrt{p_0}$ ν-almost everywhere. This is the case if and only if there are numbers $w_0, \ldots, w_M$, that are not all equal to zero, satisfying $\sum_{j=0}^{M} w_j \sqrt{p_j} = 0$, ν-almost everywhere. To verify the more difficult reverse direction of this equivalence, it is enough to observe that $\sum_{j=0}^{M} w_j \sqrt{p_j} = 0$ implies $w_0 = -\sum_{j=1}^{M} w_j \int \sqrt{p_j p_0} d\nu$ and thus, taking $v_j = w_j \int \sqrt{p_j p_0} d\nu$ yields $\sum_{j=1}^{M} \frac{v_j \sqrt{p_j}}{\int \sqrt{p_j p_0} d\nu} = \sum_{j=1}^{M} v_j \sqrt{p_0}$.

We now show the general case of (ii). For an $n \times n$ matrix A and index sets $I, J \subset \{1, \ldots, n\}$, the submatrix $A_{I,J}$ defines the submatrix consisting of the rows I and the columns J. If $I = J$, $A_{I,I}$ is called a principal submatrix of the matrix A. Let r be an integer in $\{1, \ldots, M\}$. By Lemma [6.4],

$$r = \text{Rank}(\rho(P_0|P_1, \ldots, P_M))$$

if and only if

$$\rho(P_0|P_1, \ldots, P_M) \text{ has an invertible principal submatrix of size } r$$
$$\text{and all principal submatrix of size } r + 1 \text{ of } \rho(P_0|P_1, \ldots, P_M) \text{ are singular}$$

if and only if (using the fact that the principal submatrices of $\rho(P_0|P_1, \ldots, P_M)$ of size k are exactly the matrices of the form $\rho(P_0|P_{i_1}, \ldots, P_{i_k})$ for some $i_1, \ldots, i_k \in \{1, \ldots M\}$)

$$r = \max\{k = 1, \ldots, M : \exists i_1, \ldots, i_k \in \{1, \ldots, M\}, \rho(P_0|P_{i_1}, \ldots, P_{i_k}) \text{ is invertible}\}$$

if and only if (using the case of full rank that was proved before)

$$r = \max\left\{k = 1, \ldots, M : \exists i_1, \ldots, i_k \in \{1, \ldots, M\},\right.$$
$$\left.\sqrt{p_0}, \sqrt{p_{i_1}}, \ldots, \sqrt{p_{i_r}} \text{ are linearly independent}\right\}$$

if and only if $r = \text{Rank}(\sqrt{p_0}, \sqrt{p_1}, \ldots, \sqrt{p_M}) - 1$. $\qquad\square$

4 Data processing inequality for the χ^2-divergence matrix

In a parametric statistical model $(Q_\theta)_{\theta \in \Theta}$, it is assumed that the statistician observes a random variable X following one of the distributions Q_θ for some $\theta \in \Theta$. If we transform X to obtain a new variable Y, then Y follows the distribution $P_\theta := KQ_\theta$ for some Markov kernel K. When θ is unknown but the Markov kernel K is known and independent of θ, this means that the new statistical model is $(P_\theta := KQ_\theta, \theta \in \Theta)$. As in the usual case for the χ^2-divergence, it is natural to think that such a transformation cannot increase the amount of information present in the model. In our more general framework, such an inequality still holds and is presented in the following data processing inequality.

Theorem 4.1 (Data processing/entropy contraction) *If K is a Markov kernel and $Q_0, \ldots, Q_M$ are probability measures such that Q_0 dominates $Q_1, \ldots, Q_M$, then,*

$$\chi^2(KQ_0|KQ_1, \ldots, KQ_M) \le \chi^2(Q_0|Q_1, \ldots, Q_M),$$

where $\le$ denotes the partial order on the set of positive semi-definite matrices.

In particular, the χ^2-divergence matrix is invariant under invertible transformations. The rest of this section is devoted to the proof of Theorem 4.1. First, we generalize the well-known data-processing inequality for the χ^2-divergence to the case (18), where one measure is a finite signed measure and use afterwards Eq. (19).

The $\chi^2(\mu, P)$-divergence with a signed measure can be computed from the usual χ^2-divergence between probability measures by the following relationship

Lemma 4.2 *Assume that $\mu \ll P$. Let $\mu = \alpha_+\mu_+ - \alpha_-\mu_-$ be the Jordan decomposition (3) of μ with $\alpha_+, \alpha_- \geq 0$ and μ_+, μ_- orthogonal probability measures. Then*

$$\chi^2(\mu, P) = \alpha_+^2 \chi^2(\mu_+, P) + \alpha_-^2 \chi^2(\mu_-, P) + 2\alpha_+\alpha_-.$$

Proof Observe that $\alpha_+^2 \chi^2(\mu_+, P) + \alpha_-^2 \chi^2(\mu_-, P) + 2\alpha_+\alpha_- = \int \left(\alpha_+\left(\dfrac{d\mu_+}{dP} - 1\right) - \alpha_-\left(\dfrac{d\mu_-}{dP} - 1\right)\right)^2 dP = \int \left(\dfrac{d\mu}{dP} - \mu(\Omega)\right)^2 dP = \chi^2(\mu, P)$. $\qquad\square$

Lemma 4.3 *If μ is a finite signed measure, P is a probability measure and both measures are defined on the same measurable space, then, for any Markov kernel K, the data-processing inequality*

$$\chi^2(K\mu, KP) \leq \chi^2(\mu, P)$$

holds.

Proof We can assume that $\mu \ll P$, since otherwise the right-hand side of the inequality is $+\infty$ and the result holds. In particular, $\mu \ll \nu$ for a positive measure ν implies that $K\mu \ll K\nu$. Indeed, if $K\nu(A) = 0$ for a given measurable set A, then, $\int K(A, x)d\nu(x) = 0$, implying $K(A, \cdot) = 0$ ν-almost everywhere. Since $\mu \ll \nu$, the equality also holds μ-almost everywhere and so $K\mu(A) = \int K(A, x)d\mu(x) = 0$, proving $K\mu \ll K\nu$. By the Jordan decomposition (3), there exist orthogonal probability measures μ_+, μ_- and non-negative real numbers α_+, α_-, such that $\mu = \alpha_+\mu_+ - \alpha_-\mu_-$ and $\mu(\Omega) = \alpha_+ - \alpha_-$. Thus, $K\mu = \alpha_+K\mu_+ - \alpha_-K\mu_-$. Observe that

$$\int \left(\frac{dK\mu_+}{dKP} - 1\right)\left(\frac{dK\mu_-}{dKP} - 1\right)dKP = \int \left(\frac{dK\mu_+}{dKP}\frac{dK\mu_-}{dKP} - \frac{dK\mu_-}{dKP} - \frac{dK\mu_+}{dKP} + 1\right)dKP$$

$$= \int \frac{dK\mu_+}{dKP}\frac{dK\mu_-}{dKP}dKP - 1 \geq -1.$$

Because μ_+ and μ_- are orthogonal, we similarly find that

$$\int \left(\frac{d\mu_+}{dP} - 1\right)\left(\frac{d\mu_-}{dP} - 1\right)dP = -1.$$

Using the data-processing inequality for the χ^2 divergence of probability measures twice, $K\mu = \alpha_+K\mu_+ - \alpha_-K\mu_-$ and $\mu(\Omega) = \alpha_+ - \alpha_-$, we get

$$\chi^2(K\mu, KP) = \int \left(\frac{dK\mu}{dKP} - \mu(\Omega)\right)^2 dKP$$

$$= \int \left(\alpha_+\left(\frac{dK\mu_+}{dKP} - 1\right) - \alpha_-\left(\frac{dK\mu_-}{dKP} - 1\right)\right)^2 dKP$$

$$= \alpha_+^2 \chi^2 \left(K\mu_+, KP \right) + \alpha_-^2 \chi^2 \left(K\mu_-, KP \right)$$

$$- 2\alpha_+ \alpha_- \int \left(\frac{dK\mu_+}{dKP} - 1 \right) \left(\frac{dK\mu_-}{dKP} - 1 \right) dKP$$

$$\leq \alpha_+^2 \chi^2 \left(\mu_+, P \right) + \alpha_-^2 \chi^2 (\mu_-, P) + 2\alpha_+ \alpha_- = \chi^2 (\mu, P),$$

by Lemma 4.2. □

We can now complete the proof of Theorem 4.1.

Proof of Theorem 4.1 Let $v = (v_1, \ldots, v_M)^\top \in \mathbb{R}^M$. Then, $\sum_{j=1}^M v_j Q_j$ is a finite signed measure dominated by Q_0. Using (18) and the previous lemma,

$$v^T \chi^2 (KQ_0|KQ_1, \ldots, KQ_M)v = \int \left(\sum_{j=1}^M v_j \left(\frac{dKQ_j}{dKQ_0} - 1 \right) \right)^2 dKQ_0$$

$$= \chi^2 \left(K \left(\sum_{j=1}^M v_j Q_j \right), KQ_0 \right)$$

$$\leq \chi^2 \left(\sum_{j=1}^M v_j Q_j, Q_0 \right)$$

$$= \int \left(\sum_{j=1}^M v_j \left(\frac{dQ_j}{dQ_0} - 1 \right) \right)^2 dQ_0$$

$$= v^T \chi^2 (Q_0|Q_1, \ldots, Q_M)v.$$

Since v was arbitrary, this completes the proof. □

A Markov kernel K implies by definition that for every fixed x, $A \mapsto K(A, x)$ is a probability measure. We now provide a simpler and more straightforward proof for Theorem 4.1 without using Lemma 4.3, under the additional common domination assumption:

$$\text{There exists a measure } \mu, \text{ such that } \forall x \in \Omega, K(x, \cdot) \ll \mu. \tag{21}$$

Simpler proof of Theorem 4.1 under the additional assumption (21) Because of the identity $v^\top \chi^2 (Q_0|Q_1, \ldots, Q_M)v = \int (\sum_{j=1}^M v_j (dQ_j/dQ_0 - 1))^2 dQ_0$, it is enough to prove that for any arbitrary vector $v = (v_1, \ldots, v_M)^\top$,

$$\int \left(\sum_{j=1}^M v_j \left(\frac{dKQ_j}{dKQ_0} - 1 \right) \right)^2 dKQ_0 \leq \int \left(\sum_{j=1}^M v_j \left(\frac{dQ_j}{dQ_0} - 1 \right) \right)^2 dQ_0. \tag{22}$$

Let ν be a dominating measure for $Q_0, \ldots, Q_M$ and recall that by the additional assumption (21), for any x, the measure μ is a dominating measure for the proba-bility measure $A \mapsto K(A, x)$. Write q_j for the ν-density of Q_j. Then, $dKQ_j(y) = \int_X k(y, x)q_j(x)d\nu(x)d\mu(y)$ for $j = 1, \ldots, M$ and a suitable non-negative kernel function k satisfying $\int k(y, x)d\mu(y) = 1$ for all x. Applying the Cauchy-Schwarz inequality, we obtain

$$
\left(\sum_{j=1}^{M} v_j \left(\frac{dKQ_j}{dKQ_0}(y) - 1 \right) \right)^2 = \left(\frac{\int k(y, x)[\sum_{j=1}^{M} v_j(q_j(x) - q_0(x))]d\nu(x)}{\int k(y, x')q_0(x')d\nu(x')} \right)^2
$$

$$
\leq \frac{\int k(y, x) \left(\sum_{j=1}^{M} v_j \frac{(q_j(x) - q_0(x))}{q_0(x)} \right)^2 q_0(x)d\nu(x)}{\int k(y, x')q_0(x')d\nu(x')}.
$$

Inserting this in (22), rewriting $dKQ_0(y) = \int_X k(y, x)q_0(x)d\nu(x)d\mu(y)$, interchang-ing the order of integration using Fubini's theorem, and applying $\int k(y, x)d\mu(y) = 1$, yields

$$
\int \left(\sum_{j=1}^{M} v_j \left(\frac{dKQ_j}{dKQ_0} - 1 \right) \right)^2 dKQ_0
$$

$$
\leq \iint k(y, x) \left(\sum_{j=1}^{M} v_j \frac{(q_j(x) - q_0(x))}{q_0(x)} \right)^2 q_0(x)d\nu(x)d\mu(y)
$$

$$
= \int \left(\sum_{j=1}^{M} v_j \left(\frac{q_j(x)}{q_0(x)} - 1 \right) \right)^2 q_0(x)d\nu(x)
$$

$$
= \int \left(\sum_{j=1}^{M} v_j \left(\frac{dQ_j}{dQ_0} - 1 \right) \right)^2 dQ_0.
$$

□

5 Derivations for explicit expressions for the R_α codivergence

In this section we derive closed-form expressions for the R_α codivergences in Table 1. We also obtain a closed-form formula for the case of Gamma distributions and discuss a first order approximation of it.

5.1 Multivariate normal distribution

Suppose $P_j = \mathcal{N}(\theta_j, \sigma^2 I_d)$ for $j = 0, 1, 2$. Here $\theta_j = (\theta_{j1}, \ldots, \theta_{jd})^\top$ are vectors in $\mathbb{R}^d$ and I_d denotes the $d \times d$ identity matrix. Then,

$$R_\alpha(P_0|P_1, P_2) = \exp\left(\alpha^2 \frac{\langle \theta_1 - \theta_0, \theta_2 - \theta_0 \rangle}{\sigma^2}\right) - 1. \tag{23}$$

Proof The Lebesgue density of P_j is

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|x - \theta_j\|^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) \exp\left(\frac{1}{\sigma^2}\theta_j^\top x - \frac{\|\theta_j\|^2}{2\sigma^2}\right),$$

with $\|\cdot\|$ the Euclidean norm. This is an exponential family $h(x)\exp(\langle \theta, T(x)\rangle - A(\theta))$ with $T(x) = \sigma^{-2}x$ and $A(\theta) = \|\theta\|^2/(2\sigma^2)$.

Applying Proposition 2.7 and quadratic expansion $\|\theta_0 + b\|^2 = \|\theta_0\|^2 + 2\langle\theta_0, b\rangle + \|b\|^2$ to all four terms yields

$$\begin{aligned}
R_\alpha(P_0|P_1, P_2) &= \exp\left(\frac{\|\theta_0 + \alpha(\theta_1 + \theta_2 - 2\theta_0)\|^2}{2\sigma^2} - \frac{\|\theta_0 + \alpha(\theta_1 - \theta_0)\|^2}{2\sigma^2}\right.\\
&\quad \left. - \frac{\|\theta_0 + \alpha(\theta_2 - \theta_0)\|^2}{2\sigma^2} + \frac{\|\theta_0\|^2}{2\sigma^2}\right) - 1\\
&= \exp\left(\frac{\|\alpha(\theta_1 + \theta_2 - 2\theta_0)\|^2 - \|\alpha(\theta_1 - \theta_0)\|^2 - \|\alpha(\theta_2 - \theta_0)\|^2}{2\sigma^2}\right) - 1\\
&= \exp\left(\alpha^2 \frac{\langle\theta_1 - \theta_0, \theta_2 - \theta_0\rangle}{\sigma^2}\right) - 1.
\end{aligned}$$

$\square$

5.2 Poisson distribution

If $\mathrm{Pois}(\lambda)$ denotes the Poisson distribution with intensity $\lambda > 0$, and $\lambda_0, \lambda_1, \lambda_2 > 0$, then,

$$R_\alpha\big(\mathrm{Pois}(\lambda_0) \,\big|\, \mathrm{Pois}(\lambda_1), \, \mathrm{Pois}(\lambda_2)\big) = \exp\left(\lambda_0^{1-2\alpha}\big(\lambda_1^\alpha - \lambda_0^\alpha\big)\big(\lambda_2^\alpha - \lambda_0^\alpha\big)\right) - 1.$$

Suppose $P_j = \otimes_{\ell=1}^d \mathrm{Pois}(\lambda_{j\ell})$ for $j = 0, \ldots, M$ and $\lambda_{j\ell} > 0$ for all j, ℓ. Then, as a consequence of Proposition 2.6,

$$R_\alpha(P_0|P_1, P_2) = \exp\left(\sum_{\ell=1}^d \lambda_{0\ell}^{1-2\alpha}\big(\lambda_{1\ell}^\alpha - \lambda_{0\ell}^\alpha\big)\big(\lambda_{2\ell}^\alpha - \lambda_{0\ell}^\alpha\big)\right) - 1,$$

with particular cases

$$\chi^2(P_0|P_1, P_2) = \exp\left(\sum_{\ell=1}^{d} \frac{(\lambda_{1\ell} - \lambda_{0\ell})(\lambda_{2\ell} - \lambda_{0\ell})}{\lambda_{0\ell}}\right) - 1, \tag{24}$$

and

$$\rho(P_0|P_1, P_2) = \exp\left(\sum_{\ell=1}^{d} \left(\sqrt{\lambda_{1\ell}} - \sqrt{\lambda_{0\ell}}\right)\left(\sqrt{\lambda_{2\ell}} - \sqrt{\lambda_{0\ell}}\right)\right) - 1. \tag{25}$$

Proof The density of the Poisson distribution with respect to the counting measure is

$$p_\lambda(x) = e^{-\lambda}\frac{\lambda^x}{x!} = \frac{1}{x!}e^{-\lambda + x\log(\lambda)} = h(x)\exp\left(\theta^\top T(x) - A(\theta)\right),$$

with $\theta = \log(\lambda)$, $T(x) = x$ and $A(\theta) = \exp(\theta)$. Applying Proposition 2.7 gives

$$R_\alpha(P_{\theta_0}|P_{\theta_1}, P_{\theta_2}) = \exp\left(A\left(\theta_0 + \alpha\left(\theta_1 + \theta_2 - 2\theta_0\right)\right) - A\left(\theta_0 + \alpha\left(\theta_1 - \theta_0\right)\right)\right.$$
$$\left. - A\left(\theta_0 + \alpha\left(\theta_2 - \theta_0\right)\right) + A\left(\theta_0\right)\right) - 1.$$

$$= \exp\left(\exp\left(\log(\lambda_0) + \alpha\left(\log(\lambda_1) + \log(\lambda_2) - 2\log(\lambda_0)\right)\right)\right.$$
$$- \exp\left(\log(\lambda_0) + \alpha\left(\log(\lambda_1) - \log(\lambda_0)\right)\right)$$
$$\left. - \exp\left(\log(\lambda_0) + \alpha\left(\log(\lambda_2) - \log(\lambda_0)\right)\right) + \lambda_0\right) - 1.$$

$$= \exp\left(\lambda_0^{1-2\alpha}\lambda_1^\alpha\lambda_2^\alpha - \lambda_0^{1-\alpha}\lambda_1^\alpha - \lambda_0^{1-\alpha}\lambda_2^\alpha + \lambda_0\right) - 1$$

$$= \exp\left(\lambda_0^{1-2\alpha}\left(\lambda_1^\alpha - \lambda_0^\alpha\right)\left(\lambda_2^\alpha - \lambda_0^\alpha\right)\right) - 1.$$

$$\square$$

5.3 Bernoulli distribution

If $\mathrm{Ber}(\theta)$ denotes the Poisson distribution with parameter $\theta \in (0, 1)$, and θ_0, θ_1, $\theta_2 \in (0, 1)$, then,

$$R_\alpha\left(\mathrm{Ber}(\theta_0) \mid \mathrm{Ber}(\theta_1), \mathrm{Ber}(\theta_2)\right)$$
$$= \frac{\theta_0^{1-2\alpha}\theta_1^\alpha\theta_2^\alpha + (1-\theta_0)^{1-2\alpha}(1-\theta_1)^\alpha(1-\theta_2)^\alpha}{\left(\theta_0^{1-\alpha}\theta_1^\alpha + (1-\theta_0)^{1-\alpha}(1-\theta_1)^\alpha\right)\left(\theta_0^{1-\alpha}\theta_2^\alpha + (1-\theta_0)^{1-\alpha}(1-\theta_2)^\alpha\right)} - 1,$$

Suppose $P_j = \otimes_{\ell=1}^d \mathrm{Ber}(\theta_{j\ell})$ for $j = 0, 1, 2$ and $\theta_{j\ell} \in (0, 1)$ for all j, ℓ. Then, as a consequence of Proposition 2.6,

$$R_\alpha(P_0|P_1, P_2) = \prod_{\ell=1}^d \frac{\theta_{0\ell}^{1-2\alpha}\theta_{1\ell}^\alpha\theta_{2\ell}^\alpha + (1 - \theta_{0\ell})^{1-2\alpha}(1 - \theta_{1\ell})^\alpha(1 - \theta_{2\ell})^\alpha}{r(\theta_{0\ell}, \theta_{1\ell})r(\theta_{0\ell}, \theta_{2\ell})} - 1,$$

where $r(\theta_0, \theta_1) := \theta_0^{1-\alpha}\theta_1^\alpha + (1 - \theta_0)^{1-\alpha}(1 - \theta_1)^\alpha$. In particular,

$$\chi^2(P_0|P_1, P_2) = \prod_{\ell=1}^d \left(\frac{(\theta_{1\ell} - \theta_{0\ell})(\theta_{2\ell} - \theta_{0\ell})}{\theta_{0\ell}(1 - \theta_{0\ell})} + 1\right) - 1, \tag{26}$$

and

$$\rho(P_0|P_1, P_2) = \prod_{\ell=1}^d \frac{\widetilde{r}(\theta_{1\ell}, \theta_{2\ell})}{\widetilde{r}(\theta_{1\ell}, \theta_{0\ell})\widetilde{r}(\theta_{2\ell}, \theta_{0\ell})} - 1, \tag{27}$$

with $\widetilde{r}(\theta, \theta') := \sqrt{\theta\theta'} + \sqrt{(1 - \theta)(1 - \theta')}$.

Proof The Bernoulli distributions $\mathrm{Ber}(\theta), \theta \in (0, 1)$ form an exponential family, dominated by the counting measure on $\{0, 1\}$ with density $P(\mathrm{Ber}(\theta) = k) = \theta^k(1 - \theta)^{1-k} = \exp(k\log(\theta) + (1 - k)\log(1 - \theta)) = \exp(k\beta - \log(1 + e^\beta))$, where $\beta = \log(\theta/(1 - \theta))$ is the natural parameter and $A(\beta) = \log(1 + e^\beta)$. Therefore, we can apply Proposition 2.7 and obtain

$$R_\alpha(P_{\beta_0}|P_{\beta_1}, P_{\beta_2}) = \exp\Big(A\big(\beta_0 + \alpha(\beta_1 + \beta_2 - 2\beta_0)\big) - A\big(\beta_0 + \alpha(\beta_1 - \beta_0)\big)$$
$$- A\big(\beta_0 + \alpha(\beta_2 - \beta_0)\big) + A\big(\beta_0\big)\Big) - 1.$$

Note that

$$\beta_0 + \alpha(\beta_1 - \beta_0) = \log\left(\frac{\theta_0}{1 - \theta_0}\right) + \alpha\left(\log\left(\frac{\theta_1}{1 - \theta_1}\right) - \log\left(\frac{\theta_0}{1 - \theta_0}\right)\right)$$
$$= \log\left(\frac{\theta_0^{1-\alpha}\theta_1^\alpha}{(1 - \theta_0)^{1-\alpha}(1 - \theta_1)^\alpha}\right),$$

so that

$$A\big(\beta_0 + \alpha(\beta_1 - \beta_0)\big) = \log\left(1 + \frac{\theta_0^{1-\alpha}\theta_1^\alpha}{(1 - \theta_0)^{1-\alpha}(1 - \theta_1)^\alpha}\right)$$
$$= \log\left(\frac{\theta_0^{1-\alpha}\theta_1^\alpha + (1 - \theta_0)^{1-\alpha}(1 - \theta_1)^\alpha}{(1 - \theta_0)^{1-\alpha}(1 - \theta_1)^\alpha}\right).$$

Similarly,

$$\beta_0 + \alpha\left(\beta_1 + \beta_2 - 2\beta_0\right)$$

$$= \log\left(\frac{\theta_0}{1-\theta_0}\right) + \alpha\left(\log\left(\frac{\theta_1}{1-\theta_1}\right) + \log\left(\frac{\theta_2}{1-\theta_2}\right) - 2\log\left(\frac{\theta_0}{1-\theta_0}\right)\right)$$

$$= \log\left(\frac{\theta_0^{1-2\alpha}\theta_1^\alpha\theta_2^\alpha}{(1-\theta_0)^{1-2\alpha}(1-\theta_1)^\alpha(1-\theta_2)^\alpha}\right),$$

so that

$$A\left(\beta_0 + \alpha\left(\beta_1 + \beta_2 - 2\beta_0\right)\right) = \log\left(\frac{\theta_0^{1-2\alpha}\theta_1^\alpha\theta_2^\alpha + (1-\theta_0)^{1-2\alpha}(1-\theta_1)^\alpha(1-\theta_2)^\alpha}{(1-\theta_0)^{1-2\alpha}(1-\theta_1)^\alpha(1-\theta_2)^\alpha}\right).$$

Combining all these results together yields

$$R_\alpha(P_{\beta_0} | P_{\beta_1}, P_{\beta_2}) = \exp\left(\log\left(\frac{\theta_0^{1-2\alpha}\theta_1^\alpha\theta_2^\alpha + (1-\theta_0)^{1-2\alpha}(1-\theta_1)^\alpha(1-\theta_2)^\alpha}{(1-\theta_0)^{1-2\alpha}(1-\theta_1)^\alpha(1-\theta_2)^\alpha}\right)\right.$$

$$- \log\left(\frac{\theta_0^{1-\alpha}\theta_1^\alpha + (1-\theta_0)^{1-\alpha}(1-\theta_1)^\alpha}{(1-\theta_0)^{1-\alpha}(1-\theta_1)^\alpha}\right)$$

$$\left.- \log\left(\frac{\theta_0^{1-\alpha}\theta_2^\alpha + (1-\theta_0)^{1-\alpha}(1-\theta_2)^\alpha}{(1-\theta_0)^{1-\alpha}(1-\theta_2)^\alpha}\right) + \log(1-\theta_0)\right) - 1$$

$$= \frac{\theta_0^{1-2\alpha}\theta_1^\alpha\theta_2^\alpha + (1-\theta_0)^{1-2\alpha}(1-\theta_1)^\alpha(1-\theta_2)^\alpha}{\left(\theta_0^{1-\alpha}\theta_1^\alpha + (1-\theta_0)^{1-\alpha}(1-\theta_1)^\alpha\right)\left(\theta_0^{1-\alpha}\theta_2^\alpha + (1-\theta_0)^{1-\alpha}(1-\theta_2)^\alpha\right)} - 1,$$

finishing the proof.

□

5.4 Gamma distribution

Let $P_\theta = \Gamma(\alpha, \beta)$ with $\theta = (\alpha - 1, -\beta)$ denote the Gamma distribution with shape $\alpha > 0$ and inverse scale $\beta > 0$. If $\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \alpha_0 + \alpha(\alpha_1 + \alpha_2 - 2\alpha_0), \alpha_0 + \alpha(\alpha_1 - \alpha_0), \alpha_0 + \alpha(\alpha_2 - \alpha_0), \beta_0 + \alpha(\beta_1 + \beta_2 - 2\beta_0), \beta_0 + \alpha(\beta_1 - \beta_0) > 0$, and $\beta_0 + \alpha(\beta_2 - \beta_0)$ are all positive, then we have

$$R_\alpha(P_{\theta_0} | P_{\theta_1}, P_{\theta_2}) = \frac{\Gamma(\alpha_0)\Gamma\left(\alpha_0 + \alpha(\alpha_1 + \alpha_2 - 2\alpha_0)\right)}{\Gamma\left(\alpha_0 + \alpha(\alpha_1 - \alpha_0)\right)\Gamma\left(\alpha_0 + \alpha(\alpha_2 - \alpha_0)\right)}$$

$$\times \frac{\left(\beta_0 + \alpha(\beta_1 - \beta_0)\right)^{\alpha_0 + \alpha(\alpha_1 - \alpha_0)}\left(\beta_0 + \alpha(\beta_2 - \beta_0)\right)^{\alpha_0 + \alpha(\alpha_2 - \alpha_0)}}{\beta_0^{\alpha_0}\left(\beta_0 + \alpha(\beta_1 + \beta_2 - 2\beta_0)\right)^{\alpha_0 + \alpha(\alpha_1 + \alpha_2 - 2\alpha_0)}} - 1,$$

otherwise $R_\alpha(P_{\theta_0}|P_{\theta_1}, P_{\theta_2}) = +\infty$. This can be checked by writing the explicit expression of the integrals that appear in the definition of R_α.

Proof The Gamma distributions $\Gamma(\alpha, \beta), \alpha > 0, \beta > 0$ form an exponential family, dominated by the Lebesgue measure with density

$$\beta^\alpha x^{\alpha-1} \exp(-\beta x)\Gamma(\alpha)^{-1} = \exp\left((\alpha - 1)\log(x) - \beta x + \alpha \log(\beta) - \log(\Gamma(\alpha))\right),$$

natural parameter

$$\theta = (\theta_a, \theta_b) = (\alpha - 1, -\beta),$$

and

$$A(\theta) = \log\left(\Gamma(\theta_a + 1)\right) - (\theta_a + 1)\log(-\theta_b).$$

Therefore, we can apply Proposition 2.7 with $\Theta = (-1, +\infty) \times (-\mathbb{R})$. Combining the assumed constraints on the parameters and the linearity of the mapping $(\alpha, \beta) \mapsto \theta$ ensures that $\theta_0 + \alpha(\theta_1 - \theta_0), \theta_0 + \alpha(\theta_1 - \theta_0), \theta_0 + \alpha(\theta_1 + \theta_2 - 2\theta_0) \in \Theta$. Therefore, we obtain

$$R_\alpha(P_{\theta_0}|P_{\theta_1}, P_{\theta_2}) = \exp\Big(A\big(\theta_0 + \alpha(\theta_1 + \theta_2 - 2\theta_0)\big) - A\big(\theta_0 + \alpha(\theta_1 - \theta_0)\big)$$
$$- A\big(\theta_0 + \alpha(\theta_2 - \theta_0)\big) + A(\theta_0)\Big) - 1,$$

where

$$A\Big(\theta_0 + \alpha(\theta_1 - \theta_0)\Big) = \log\left(\Gamma\Big(\theta_{0,a} + \alpha\Big(\theta_{1,a} - \theta_{0,a}\Big) + 1\Big)\right)$$
$$- \Big(\theta_{0,a} + \alpha\Big(\theta_{1,a} - \theta_{0,a}\Big) + 1\Big)\log\Big(\beta_0 + \alpha\Big(\beta_1 - \beta_0\Big)\Big)$$
$$= \log\left(\Gamma\Big(\alpha_0 + \alpha\Big(\alpha_1 - \alpha_0\Big)\Big)\right)$$
$$- \Big(\alpha_0 + \alpha\Big(\alpha_1 - \alpha_0\Big)\Big)\log\Big(\beta_0 + \alpha(\beta_1 - \beta_0)\Big)$$

and

$$
A\big(\theta_0 + \alpha(\theta_1 + \theta_2 - 2\theta_0)\big) = \log\Big(\Gamma\big(\alpha_0 + \alpha(\alpha_1 + \alpha_2 - 2\alpha_0)\big)\Big)
$$
$$
- \big(\alpha_0 + \alpha(\alpha_1 + \alpha_2 - 2\alpha_0)\big)\log\big(\beta_0 + \alpha(\beta_1 + \beta_2 - 2\beta_0)\big).
$$

Combining all these results, we obtain

$$
R_\alpha(P_{\theta_0}|P_{\theta_1}, P_{\theta_2}) = \frac{\Gamma(\alpha_0)\Gamma\big(\alpha_0 + \alpha(\alpha_1 + \alpha_2 - 2\alpha_0)\big)}{\Gamma\big(\alpha_0 + \alpha(\alpha_1 - \alpha_0)\big)\Gamma\big(\alpha_0 + \alpha(\alpha_2 - \alpha_0)\big)}
$$
$$
\times \frac{\big(\beta_0 + \alpha(\beta_1 - \beta_0)\big)^{\alpha_0 + \alpha(\alpha_1 - \alpha_0)}\big(\beta_0 + \alpha(\beta_2 - \beta_0)\big)^{\alpha_0 + \alpha(\alpha_2 - \alpha_0)}}{\beta_0^{\alpha_0}\big(\beta_0 + \alpha(\beta_1 + \beta_2 - 2\beta_0)\big)^{\alpha_0 + \alpha(\alpha_1 + \alpha_2 - 2\alpha_0)}} - 1.
$$

$\square$

A formula for the product of exponential distributions can be obtained as a special case by setting $\alpha_{j\ell} = 1$ for all j, ℓ and applying Proposition 2.6. For the families of distributions discussed above, the formulas for the correlation-type R_α codivergences encode an orthogonality relation on the parameter vectors. This is less visible in the expressions for the Gamma distribution but can be made more explicit using the first order approximation that we state next. It shows that even for the Gamma distribution these matrix entries can be written in leading order as a term involving a weighted inner product of $\beta_1 - \beta_0$ and $\beta_2 - \beta_0$, where β_r denotes the vector $(\beta_{r\ell})_{1\le\ell\le d}$.

Lemma 5.1 *Suppose* $P_j = \otimes_{\ell=1}^{d}\Gamma(\alpha_\ell, \beta_{j\ell})$ *for every* $j = 1, 2, 3$ *and for some* $\alpha_\ell, \beta_{j\ell} > 0$. *Let* $A := \sum_{\ell=1}^{d}\alpha_\ell$ *and* $\Delta := \max_{j=1,2}\max_{\ell=1,\ldots,d}|\beta_{j\ell} - \beta_{0\ell}|/\beta_{0\ell}$. *Denote by* Σ *the* $d \times d$ *diagonal matrix with entries* $\beta_{0\ell}^2/\alpha_\ell$. *Then,*

$$
R_\alpha(P_0|P_1, P_2) = \exp\Big(-\alpha^2(\beta_1 - \beta_0)^\top\Sigma^{-1}(\beta_2 - \beta_0) + o(A\Delta^2)\Big) - 1.
$$

Proof Using that α_ℓ does not depend on j, the expression simplifies and a second order Taylor expansion of the logarithm (the sum of the first order terms vanishes) yields

$$
R_\alpha(P_{\theta_0}|P_{\theta_1}, P_{\theta_2})
$$
$$
= \prod_{\ell=1}^{d}\frac{(\beta_{1\ell} + \beta_{0\ell})^{\alpha_\ell}(\beta_{2\ell} + \beta_{0\ell})^{\alpha_\ell}}{(2\beta_{0\ell})^{\alpha_\ell}(\beta_{1\ell} + \beta_{2\ell})^{\alpha_\ell}}
$$
$$
= \prod_{\ell=1}^{d}\exp\Big(\alpha_\ell\Big(\log\Big(1 + \alpha\frac{\beta_{1\ell} - \beta_{0\ell}}{\beta_{0\ell}}\Big) + \log\Big(1 + \alpha\frac{\beta_{2\ell} - \beta_{0\ell}}{\beta_{0\ell}}\Big)
$$
$$
- \log\Big(1 + \alpha\frac{\beta_{1\ell} - \beta_{0\ell} + \beta_{2\ell} - \beta_{0\ell}}{\beta_{0\ell}}\Big)\Big)\Big)
$$
$$
= \exp\Big(\sum_{\ell=1}^{d}\alpha_\ell\alpha^2\Big(-\frac{(\beta_{1\ell} - \beta_{0\ell})^2}{2\beta_{0\ell}^2} - \frac{(\beta_{2\ell} - \beta_{0\ell})^2}{2\beta_{0\ell}^2} + \frac{(\beta_{1\ell} - \beta_{0\ell} + \beta_{2\ell} - \beta_{0\ell})^2}{2\beta_{0\ell}^2} + o(\Delta^2)\Big)\Big)
$$

$$= \exp\left(\alpha^2 \sum_{\ell=1}^{d} \frac{\alpha_\ell(\beta_{1\ell} - \beta_{0\ell})(\beta_{2\ell} - \beta_{0\ell})}{\beta_{0\ell}^2} + o\left(A\Delta^2\right)\right).$$

□

6 Facts about ranks

Definition 6.1 Let $X_1, \ldots, X_n$ be n random variables defined on the same probability space $(\Omega, \mathcal{A}, P)$. We define the *rank* of $\{X_1, \ldots, X_n\}$, denoted by $\text{Rank}(X_1, \ldots, X_n)$ as the dimension of the vector space $\text{Vect}(X_1, \ldots, X_n)$ generated by linear combinations of $\{X_1, \ldots, X_n\}$, where the equality is to be understood P-almost surely. Moreover, we say that $(X_1, \ldots, X_n)$ are *linearly independent P-almost surely* if for any vector $(a_1, \ldots, a_n)$,

$$P\left(\sum_{i=1}^{n} a_i X_i = 0\right) = 1 \quad \text{implies} \quad a_1 = \cdots = a_n = 0.$$

Lemma 6.2 *Let $X_1, \ldots, X_n$ be n random variables defined on the same probability space $(\Omega, \mathcal{A}, P)$. $\text{Rank}(X_1, \ldots, X_n)$ is the largest integer such that there exists $i_1, \ldots, i_r \in \{1, \ldots, M\}$ with $(X_{i_1}, \ldots, X_{i_r})$ linearly independent random variables P-almost surely.*

Proof Let r be the largest integer such that there exists $i_1, \ldots, i_r \in \{1, \ldots, M\}$ with $(X_{i_1}, \ldots, X_{i_r})$ linearly independent random variables P-almost surely. Then the space generated by $X_1, \ldots, X_n$ is at least of dimension r, and therefore $\text{Rank}(X_1, \ldots, X_n) \geq r$. If $\text{Rank}(X_1, \ldots, X_n) > r$, then there exists $(r+1)$ linear combinations of the random variables that are linearly independent, contradicting the definition of r. Therefore $\text{Rank}(X_1, \ldots, X_n) \leq r$, completing the proof. □

Lemma 6.3 *Let $\mathbf{Z} = (Z_1, \ldots, Z_M)^\top$ be a M-dimensional random vector with mean zero and finite second moments. Then $\text{Rank}(\text{Cov}_P(\mathbf{Z})) = \text{Rank}(Z_1, \ldots, Z_M)$, where the covariance matrix is computed with respect to the distribution P and the rank of a set of random variables is to be understood in the sense of* Definition 6.1.

Proof Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M$ be the eigenvalues of $\text{Cov}_P(\mathbf{Z})$, sorted in decreasing order, and let $\mathbf{e}_1, \ldots, \mathbf{e}_M$ be a corresponding orthonormal basis of eigenvectors. Let r be the rank of $\text{Cov}_P(\mathbf{Z})$. We have $\lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_M = 0$ and $\lambda_r > 0$. Let us define $Y_i = \mathbf{e}_i^\top \mathbf{Z}$ for $i = 1, \ldots, M$. By usual results on principal components, e.g. [13, Result 8.1], $\text{Var}[Y_i] = \lambda_i$ and $\text{Cov}(Y_i, Y_j) = \lambda_i 1_{\{i=j\}}$. Therefore,

$$\text{Rank}(Z_1, \ldots, Z_M) = \dim(\text{Vect}(Z_1, \ldots, Z_M)) = \dim(\text{Vect}(Y_1, \ldots, Y_M))$$
$$= \dim(\text{Vect}(Y_1, \ldots, Y_r)) = r,$$

where the first equality is the definition of the rank, the second equality is a consequence of the fact that $(\mathbf{e}_1, \ldots, \mathbf{e}_M)$ is a basis of $\mathbb{R}^M$, the third equality results from the fact that $\text{Var}[Y_i] = 0$ and $E[Y_i] = 0$ for any $i > r$ and the last equality is a consequence of the orthogonality of the $(Y_1, \ldots, Y_r)$ as elements of the Hilbert space $L_2(\Omega, \mathcal{A}, P)$. The proof is completed since by definition $r = \text{Rank}(\text{Cov}_P(\mathbf{Z}))$. □

Lemma 6.4 *(see for example Exercise 3.3.11 in [22]) A symmetric and positive semi-definite $M \times M$ matrix A is of rank r if and only if A has an invertible principal submatrix of size r, and all principal submatrices of size $r + 1$ of A are singular.*

7 Conclusion

We introduced the concept of codivergence as a notion of "angle" between three probability distributions. Divergence matrices can be viewed as an analogue of the Gram matrix for a finite sequence of probability distributions that are compared relative to one distribution.

Locally around the reference probability measure P_0, codivergences are bilinear forms up to remainder terms. Two classes of codivergences emerge that resemble the structure of the covariance and the correlation.

Natural follow-up questions relate to the spectral behavior of a divergence matrix and the link between properties of the divergence matrix and properties of the underlying probability measures.

Appendix A Proofs

A.1 Proof of Proposition 2.5

Proof As mentioned already, the first and second part of Definition 2.2 are satisfied. To check the third part of Definition 2.2 for $\phi(P_0|P_1, P_2)$, let $\mu, \widetilde{\mu} \in \mathcal{M}_{P_0}$. Then

$$\frac{d(P_0 + t\mu)}{dP_0} = 1 + t\frac{d\mu}{dP_0} = 1 + th,$$

is square-integrable with respect to P_0 for any real number t. Using $\phi(1) = 1$, Taylor expansion $\phi(1 + y) = 1 + y\phi'(1) + \frac{1}{2}y^2\phi''(1) + o(y^2)$, that $\int h\, dP_0 = 0$ and that h is bounded P_0-a.e. by the definition of $\mathcal{M}_{P_0}$, we obtain that, for all sufficiently small t,

$$\int \phi\Big(\frac{d(P_0 + t\mu)}{dP_0}\Big)dP_0 = \int \phi(1 + th)dP_0 = 1 + \frac{t^2}{2}\phi''(1)\int h^2 dP_0 + o(t^2).$$

Similarly $\int \phi\big(\frac{d(P_0 + s\widetilde{\mu})}{dP_0}\big)dP_0 = 1 + \frac{1}{2}s^2\phi''(1)\int g^2 dP_0 + o(s^2)$ and

$$\int \phi\Big(\frac{d(P_0 + t\mu)}{dP_0}\Big)\phi\Big(\frac{d(P_0 + s\widetilde{\mu})}{dP_0}\Big)dP_0$$

$$= \int \phi(1 + th)\phi(1 + sg)dP_0$$

$$= 1 + \frac{\phi''(1)}{2}\int (t^2h^2 + s^2g^2)dP_0 + st\phi'(1)^2 \int gh\, dP_0 + o(t^2 + s^2)$$

$$= \Big(1 + \frac{t^2}{2}\phi''(1)\int h^2 dP_0\Big)\Big(1 + \frac{s^2}{2}\phi''(1)\int g^2 dP_0\Big) + st\phi'(1)^2 \int gh\, dP_0 + o(t^2 + s^2).$$

Taylor expansion yields $1/(1-y) = 1 + O(y)$ for all $|y| \leq 1/2$ and thus for $(s,t) \to (0,0)$,

$$R_\phi(P_0|P_0+t\mu, P_0+s\widetilde{\mu})$$
$$= \frac{st\phi'(1)^2 \int gh \, dP_0 + o(t^2+s^2)}{(1+\frac{t^2}{2}\phi''(1)\int h^2 dP_0)(1+\frac{s^2}{2}\phi''(1)\int g^2 dP_0) + o(t^2+s^2)}$$
$$= st\phi'(1)^2 \int gh \, dP_0 + o(t^2+s^2).$$

By following the same arguments and replacing the definition of $R_\phi(P_0|P_0+t\mu, P_0+s\widetilde{\mu})$ by $V_\phi(P_0|P_0+t\mu, P_0+s\widetilde{\mu})$ in the last step, we also obtain $V_\phi(P_0|P_0+t\mu, P_0+s\widetilde{\mu}) = st\phi'(1)^2 \int gh \, dP_0 + o(t^2+s^2)$. $\qquad\square$

A.2 Proof of Proposition 2.6

Proof By Fubini's theorem,

$$R_\alpha\left(\bigotimes_{\ell=1}^{d} P_{0\ell} \,\bigg|\, \bigotimes_{\ell=1}^{d} P_{1\ell}, \bigotimes_{\ell=1}^{d} P_{2\ell}\right) + 1$$

$$= \frac{\int \left(\frac{d\bigotimes_{\ell=1}^{d} P_{1\ell}}{d\bigotimes_{\ell=1}^{d} P_{0\ell}}\right)^\alpha \left(\frac{d\bigotimes_{\ell=1}^{d} P_{2\ell}}{d\bigotimes_{\ell=1}^{d} P_{0\ell}}\right)^\alpha d\left(\bigotimes_{\ell=1}^{d} P_{0\ell}\right)}{\int \left(\frac{d\bigotimes_{\ell=1}^{d} P_{1\ell}}{d\bigotimes_{\ell=1}^{d} P_{0\ell}}\right)^\alpha d\left(\bigotimes_{\ell=1}^{d} P_{0\ell}\right)\int \left(\frac{d\bigotimes_{\ell=1}^{d} P_{2\ell}}{d\bigotimes_{\ell=1}^{d} P_{0\ell}}\right)^\alpha d\left(\bigotimes_{\ell=1}^{d} P_{0\ell}\right)}$$

$$= \prod_{\ell=1}^{d} \frac{\int \left(\frac{dP_{1\ell}}{dP_{0\ell}}\right)^\alpha \left(\frac{dP_{2\ell}}{dP_{0\ell}}\right)^\alpha dP_{0\ell}}{\int \left(\frac{dP_{1\ell}}{dP_{0\ell}}\right)^\alpha dP_{0\ell} \int \left(\frac{dP_{2\ell}}{dP_{0\ell}}\right)^\alpha dP_{0\ell}}$$

$$= \prod_{\ell=1}^{d} R_\alpha(P_{0\ell}|P_{1\ell}, P_{2\ell}) + 1.$$

$\qquad\square$

A.3 Proof of Proposition 2.7

Proof Write $P_i := P_{\theta_i}$ and p_i for the corresponding ν-densities. By assumption, we have $\overline{\theta}_\alpha := \alpha(\theta_1+\theta_2) + (1-2\alpha)\theta_0 \in \Theta$ and

$$\int \left(\frac{dP_1}{dP_0}\right)^\alpha \left(\frac{dP_2}{dP_0}\right)^\alpha dP_0 = \int (p_1(x)p_2(x))^\alpha p_0(x)^{1-2\alpha} d\nu(x)$$

$$= \int h(x) \exp\left(\overline{\theta}_\alpha^\top T(x)\right) dv(x)$$
$$\cdot \exp\left(-\alpha A(\theta_1) - \alpha A(\theta_2) + (1 - 2\alpha)A(\theta_0)\right)$$
$$= \exp\left(A(\overline{\theta}_\alpha) - \alpha A(\theta_1) - \alpha A(\theta_2) + (1 - 2\alpha)A(\theta_0)\right).$$

Setting $P_2 = P_0$ (or equivalently, $\theta_2 = \theta_0$) in the previous identity gives

$$\int \left(\frac{dP_1}{dP_0}\right)^\alpha dP_0 = \exp\left(A\left(\alpha\theta_1 + (1 - \alpha)\theta_0\right) - \alpha A(\theta_1) + (1 - \alpha)A(\theta_0)\right).$$

Interchanging the role of θ_2 and θ_1 provides moreover a closed-form expression for $\int \left(\frac{dP_2}{dP_0}\right)^\alpha dP_0$. Combining everything yields

$$R_\alpha(P_0|P_1, P_2) = \frac{\int \left(\frac{dP_1}{dP_0}\right)^\alpha \left(\frac{dP_2}{dP_0}\right)^\alpha dP_0}{\int \left(\frac{dP_1}{dP_0}\right)^\alpha dP_0 \int \left(\frac{dP_2}{dP_0}\right)^\alpha dP_0} - 1$$

$$= \exp\left(A(\overline{\theta}_\alpha) - A\left(\theta_0 + \alpha\left(\theta_1 - \theta_0\right)\right) - A\left(\theta_0 + \alpha\left(\theta_2 - \theta_0\right)\right) + A(\theta_0)\right) - 1.$$

$\square$

Data availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

References

1. Amari, S.-I.: Information Geometry and Its Applications, vol. 194. Springer, Japan (2016)
2. Averbukh, V.I., Smolyanov, O.G.: The theory of differentiation in linear topological spaces. Russ. Math. Surv. **22**(6), 201 (1967)
3. Ay, N., Jost, J., Vân Lê, H., Schwachhöfer, L.: Information Geometry, vol. 64. Springer, Switzerland (2017)
4. Bourles, H.: Fundamentals of Advanced Mathematics 3—Differential Calculus, Tensor Calculus, Differential Geometry, Global Analysis. ISTE Press; Elsevier, Oxford (2019). https://doi.org/10.1016/C2017-0-00728-0
5. Cena, A., Pistone, G.: Exponential statistical manifold. Ann. Inst. Stat. Math. **59**, 27–56 (2007)
6. Chen, T., Streets, J., Shahbaba, B.: A geometric view of posterior approximation (2015). arXiv preprint arXiv:1510.00861
7. Cohn, D.L.: Measure Theory. Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks], 2nd edn. Birkhäuser/Springer, New York, p. 457 (2013). https://doi.org/10.1007/978-1-4614-6956-8
8. Dawid, A.P.: Further comments on some comments on a paper by Bradley Efron. Ann. Stat. **5**(6), 1249 (1977)
9. Derumigny, A., Schmidt-Hieber, J.: On lower bounds for the bias-variance trade-off. To appear in the Annals of Statistics (2023). arXiv:2006.00278
10. Diakonikolas, I., Kane, D.M., Stewart, A.: Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In: 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), pp. 73–84 (2017)
11. Feldman, V., Grigorescu, E., Reyzin, L., Vempala, S.S., Xiao, Y.: Statistical algorithms and a lower bound for detecting planted cliques. J. ACM **64**(2), 1–37 (2017)
12. Holbrook, A., Lan, S., Streets, J., Shahbaba, B.: The nonparametric Fisher geometry and the chi-square process density prior (2017). arXiv preprint arXiv:1707.03117
13. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis, 6th edn. Pearson Prentice Hall, New Jersey (2007)
14. Lang, S.: Differential and Riemannian Manifolds, vol. 160. Springer, New York (2012)
15. Lee, J.M.: Manifolds and Differential Geometry, vol. 107. American Mathematical Society, Providence (2022)
16. Newton, N.J.: An infinite-dimensional statistical manifold modelled on Hilbert space. J. Funct. Anal. **263**(6), 1661–1681 (2012)
17. Newton, N.J.: Infinite-dimensional statistical manifolds based on a balanced chart. Bernoulli **22**(2), 711–731 (2016)
18. Newton, N.J.: A class of non-parametric statistical manifolds modelled on Sobolev space. Inf. Geom. **2**(2), 283–312 (2019)
19. Nielsen, F.: An elementary introduction to information geometry. Entropy **22**(10), 1100 (2020)
20. Pistone, G., Sempi, C.: An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. Ann. Stat. 1543–1561 (1995)
21. Pistone, G.: Nonparametric information geometry. In: Geometric Science of Information: First International Conference, GSI 2013, Paris, France, August 28–30, 2013. Proceedings, pp. 5–36 (2013). Springer
22. Ramachandra Rao, A., Bhimasankaram, P.: Linear Algebra, vol. 19. Springer, New Delhi (2000). https://doi.org/10.1007/978-93-86279-01-9
23. Rényi, A.: On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, vol. 4, pp. 547–562 (1961). University of California Press
24. Srivastava, A., Jermyn, I., Joshi, S.: Riemannian analysis of probability density functions with applications in vision. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007). IEEE
25. Srivastava, A., Klassen, E.P.: Functional and Shape Data Analysis, vol. 1. Springer, New York (2016)