New Fréchet features for random distributions and associated sensitivity indices

Jean-Claude Fort[a] and Thierry Klein [b**]

April 22, 2016

[a]MAP5, Université Paris Descartes, SPC, 45 rue des Saints Pères, 75006 Paris, France.
[b]IMT, Université Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse Cedex 9, France.

Abstract

Using contrasts we define new Fréchet features for random cumulative distribution functions. These contrasts allow to construct Wasserstein costs and our new features minimize the average costs as the Fréchet mean minimizes the mean square Wasserstein$_2$ distance. An example of new features is the median, and more generally the quantiles. From these definitions, we are able to define sensitivity indices when the random distribution is the output of a stochastic code. Associated to the Fréchet mean we extend the Sobol indices, and in general the indices associated to a contrast that we previously proposed.

Introduction

Nowadays the output of many computer codes is not only a real multidimensional variable but frequently a function computed on so many points that it can be considered as a functional output. This function may be the cumulative distribution function ($c.d.f$) of a real random variable (phenomenon). Here we focused on the case of a $c.d.f$ output. To analyze such outputs one needs to choose a distance to compare various $c.d.f.$. Among the large possibilities offered by the literature we have chosen the Wasserstein distances (for more details we refer to [6]).

Thus we consider the problem of defining a generalized notion of barycenter of random probability measures on $\mathbb{R}$. It is a well known fact that the set of Radon probability measures endowed with the 2-Wasserstein distance is not an Euclidean space. Consequently, to define a notion of barycenter for random probability measures, it is natural to use the notion of Fréchet mean [4] that is an extension of the usual Euclidean barycenter. If $\mathbb{Y}$ denotes a random variable with distribution $\mathbb{P}$ taking its value in a metric space $(\mathcal{M}, d_{\mathcal{M}})$, then a Fréchet mean (not necessarily unique) of the distribution $\mathbb{P}$ is a point $m^* \in \mathcal{M}$ that is a global minimum (if any) of the functional $J(m) = \frac{1}{2}\int_{\mathcal{M}} d^2_{\mathcal{M}}(m, y) d\mathbb{P}(y)$ i.e. $m^* \in \arg\min_{m\in\mathcal{M}} J(m)$.

We present an attempt to use these tools and some extensions for analyzing computer codes outputs in a random environment, what is the subject of computer code experiments. At first we define new contrasts for random $c.d.f.$ by considering generalized "Wasserstein" costs. From this, in a second step we define new features in the way of the Fréchet mean that we call Fréchet features, that we illustrate through the quantiles example. Finally we propose a sensitivity analysis of random $c.d.f.$, first from a Sobol point of view that we generalized to a contrast point of view as in [2].

1 Wasserstein distances and Wasserstein costs for unidimensional distributions

For any $p \geq 1$ we may define a Wasserstein distance between two distribution of probability, denoted F and G (their cumulative distribution functions, $c.d.f.$) on $\mathbb{R}^d$ by:

$$W_p^p(F, G) = \min_{(X,Y)} \mathbb{E}\|X - Y\|^p,$$

where the random variables ($r.v.$'s) have $c.d.f.$ F and G ($X \sim F, Y \sim G$), assuming that X and Y have finite moments of order p. We call Wassertein$_p$ space the space of all $c.d.f.$ of $r.v.$'s with finite moments of order p.

As previously mentioned, in the unidimensional case where $d = 1$, it is well known that $W_p(F, G)$ is explicitly computed by:

$$W_p^p(F, G) = \int_0^1 |F^-(u) - G^-(u)|^p du = \mathbb{E}|F^-(U) - G^-(U)|^p.$$

**Corresponding author. Email: jean-claude.fort@parisdescartes.fr

Here F^- and G^- are the generalized inverses of F and G that are increasing with limits 0 and 1, and U is a *r.v.* uniform on $[0,1]$.

This result extends to more general contrast functions.

Definition 1.1 *We call contrast functions any application c from* $\mathbb{R}^2$ *to* $\mathbb{R}$ *satisfying the "measure property"* $\mathcal{P}$ *defined by*

$$\mathcal{P} : \forall x \leq x' \text{ and } \forall y \leq y', c(x',y') - c(x',y) - c(x,y') + c(x,y) \leq 0,$$

meaning that c defines a negative measure on $\mathbb{R}^2$.

Remark 1 *If C is a convex real function then c(x,y)=C(x-y) satisfies* $\mathcal{P}$. *This is the case of* $|x-y|^p$, $p \geq 1$.

Our technical framework is the Skorohod space $\mathcal{D} := \mathcal{D}(\mathbb{R}, [0,1])$ of all distribution functions, that is the space of all non decreasing function from $\mathbb{R}$ to $[0,1]$ that are càd-làg with limit 0 (resp. 1) in $-\infty$ (resp. $+\infty$) equiped with the supremum norm.

Definition 1.2 (The $c-$Wasserstein cost) *For any* $F \in \mathcal{D}$, *any* $G \in \mathcal{D}$ *and any non-negative contrast function c, we define the c− Wasserstein cost by*

$$W_c(F,G) = \min_{(X\sim F, Y\sim G)} \mathbb{E}\left(c(X,Y)\right) < +\infty$$

The following theorem can be found in ([1]).

Theorem 1.1 (Cambanis, Simon, Stout [1]) *Let c a function from* $\mathbb{R}^2$ *taking values in* $\mathbb{R}$. *Assume that it satisfies the "measure property"* $\mathcal{P}$. *Then*

$$W_c(F,G) = \int_0^1 c(F^-(u), G^-(u))du = \mathbb{E}\ c(F^-(U), G^-(U)),$$

where U is a random variable uniformly distributed on $[0,1]$.

At this point we may notice that in a statistical framework many features of probability distribution can be characterized via such a contrast function. For instance an interesting case is the quantiles. Applying the remark 1 we get:

Proposition 1.1 *For any* $\alpha \in (0,1)$ *the contrast function (pinball function) associated to the* α*-quantile* $c_\alpha(x,y) = (1-\alpha)(y-x)\mathbf{1}_{x-y<0} + \alpha(x-y)\mathbf{1}_{x-y\geq 0}$ *satisfies* $\mathcal{P}$.

2 Extension of the Fréchet mean to other features

A Fréchet mean $\mathcal{E}X$ of a *r.v.* X taking values in a metric space $(\mathcal{M}, d)$ is define as (whenever it exists):

$$\mathcal{E}X \in \text{argmin}_{\theta\in\mathcal{M}} \mathbb{E}\ d(X,\theta)^2.$$

Thus a Fréchet mean minimizes the contrast $\mathbb{E}\ d(X,\theta)^2$ which is an extension of the classical contrast $\mathbb{E}\|X-\theta\|^2$ in $\mathbb{R}^d$.

Following this idea, taking c a positive contrast satisfying property $\mathcal{P}$, we define the Fréchet feature associated to c:

Definition 2.1 *Assume that* $\mathbb{F}$ *is a random variable taking values in* $\mathcal{D}$. *Let c be a non negative contrast function satisfying the property* $\mathcal{P}$. *We define a c-contrasted feature* $\mathcal{E}_c\mathbb{F}$ *of* $\mathbb{F}$ *by:*

$$\mathcal{E}_c\mathbb{F} \in argmin_{G\in\mathcal{D}} \mathbb{E}\left(W_c(\mathbb{F}, G)\right).$$

This definition coincides with the Fréchet mean in the Wasserstein$_2$ space when using $c(F,G) = W_2^2(F,G)$.

Theorem 2.1 *If c is a positive cost function satisfying the property* $\mathcal{P}$, *if* $\mathcal{E}_c\mathbb{F}$ *exists and is unique we have:*

$$(\mathcal{E}_c\mathbb{F})^-(u) = argmin_{s\in\mathbb{R}} \mathbb{E}c(\mathbb{F}^-(u), s).$$

For instance the Fréchet mean in the Wasserstein$_2$ space is the inverse of the function $u \longrightarrow \mathbb{E}\ (\mathbb{F}^-(u))$.

Another example is the Fréchet median defines through $c = |x-y|$.Thus we consider the Wassertein$_1$ space and the "contrast function" $c(F,G) = W_1(F,G)$. We obtain the Fréchet median of a random *c.d.f.* as :

$$(\text{Med}(\mathbb{F}))^-(u) \in \text{Med}(\mathbb{F}^-(u)).$$

More generally we can define an α-quantile of a random *c.d.f.* via the contrast function $c_\alpha(x,y)$, and we obtain $q_\alpha(\mathbb{F})$ as:

$$(q_\alpha(\mathbb{F}))^-(u) \in q_\alpha(\mathbb{F}^-(u)),$$

where $q_\alpha(X)$ is the set of the α-quantiles of the *r.v.* X taking its values in $\mathbb{R}$.

2.1 Sobol index

The global Sobol index quantifies the influence of the *r.v.* X_i on the output Y. This index is based on the variance (see [5]). We can define a Sobol index for the Fréchet mean of a random *c.d.f.* $\mathbb{F} = h(X_1, \dots, X_d)$. Actually we define $\mathrm{Var}(\mathbb{F}) = \mathbb{E}W_2^2(\mathbb{F}, \mathcal{E}(\mathbb{F}))$, and

$$S_i(F) = \frac{\mathrm{Var}(\mathbb{F}) - \mathbb{E}(\mathrm{Var}[\mathbb{F}|X_i])}{\mathrm{Var}\mathbb{F}}.$$

From Theorem 2.1 we get:

$$\mathrm{Var}(\mathbb{F}) = \mathbb{E}\int_0^1 |\mathbb{F}^-(u) - \mathcal{E}(\mathbb{F})^-(u)|^2 du = \mathbb{E}\int_0^1 |\mathbb{F}^-(u) - \mathbb{E}\mathbb{F}^-(u)|^2 du = \int_0^1 \mathrm{Var}(\mathbb{F}^-(u))du.$$

And the Sobol index is now:

$$S_i(\mathbb{F}) = \frac{\int_0^1 \mathrm{Var}(\mathbb{F}^-(u))du - \int_0^1 \mathbb{E}\mathrm{Var}[\mathbb{F}^-(u)|X_i]du}{\int_0^1 \mathrm{Var}(\mathbb{F}^-(u))du} = \frac{\int_0^1 \mathrm{Var}(\mathbb{E}[\mathbb{F}^-(u)|X_i])du}{\int_0^1 \mathrm{Var}(\mathbb{F}^-(u))du}.$$

2.2 Sensitivity index associated to a contrast function

The Sobol index can be extended to more general contrast functions. For a feature of a real *r.v.* associated to a contrast function c we have defined a sensitivity index (see ([2])):

$$S_{i,c} = \frac{\min_{\theta\in\mathbb{R}} \mathbb{E}c(Y,\theta) - \mathbb{E}\min_{\theta\in\mathbb{R}} \mathbb{E}[c(Y,\theta)|X_i]}{\min_{\theta\in\mathbb{R}} \mathbb{E}c(Y,\theta)}.$$

Along the same line, we now define a sensitivity index for a c-contrasted feature of a random *c.d.f.* by:

$$S_{i,c} = \frac{\min_{G\in\mathbb{W}} \mathbb{E}W_c(\mathbb{F},G) - \mathbb{E}\min_{G\in\mathbb{W}} \mathbb{E}[W_c(\mathbb{F},G)|X_i]}{\min_{G\in\mathbb{W}} \mathbb{E}W_c(\mathbb{F},G)}.$$

For instance if $c = |x-y|$, $(\mathcal{E}_c\mathbb{F})^-(u)$ is the median (assumed to be unique) of the random variable $\mathbb{F}^-(u)$ and:

$$S_{i,\mathrm{Med}} = \frac{\mathbb{E}\int_0^1 |\mathbb{F}^-(u) - Med(\mathbb{F}^-(u))|du - \mathbb{E}[\int_0^1 |\mathbb{F}^-(u) - Med[\mathbb{F}^-(u)|X_i]|du]}{\mathbb{E}\int_0^1 |\mathbb{F}^-(u) - Med(\mathbb{F}^-(u))|du}.$$

3 Conclusion

We have defined new features for a random *c.d.f.*, together with its sensitivity analysis. This theory is based on contrast functions that allow to compute Wasserstein costs. We intend to apply our methodology to an industrial problem: the PoD (Probability of Detection of a defect) in a random environment. In particular we hope that our α-quantiles will provide a relevantt tool to analyze that type of data.

References

[1] Stamatis Cambanis, Gordon Simons, and William Stout. Inequalities for $Ek(X,Y)$ when the marginals are fixed. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 36(4):285–294, 1976.

[2] J.-C. Fort, T. Klein, and N. Rachdi. New sensitivity analysis subordinated to a contrast. *Communications in Statistics-Theory and methods*, 2016.

[3] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H.Poincaré, Sect. B, Prob. et Stat.*, 10:235–310, 1948.

[4] I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Math. Modeling Comput. Experiment*, 1(4):407–414 (1995), 1993.

[5] C. Villani. *Topics in Optimal Transportation.* American Mathematical Society, 2003.