Directional Nonlinear Principal Components

A Measure Transportation Approach

Marc Hallin
Department of Mathematics
Université libre de Bruxelles, Belgium
and
Czech Academy of Sciences, Prague, Czech Republic

based on joint work with

Klaus Nordhausen and Una Radojicic

(University of Helsinki and Vienna University of Technology)

International Conference on Robust Statistics

Stresa, Italy, May 19–23, 2025

1. Introduction.

1.1. Traditional Principal Components.

One of the earliest techniques in multivariate analysis—it can be traced back to Pearson (1901 and Hotelling (1933)—Principal Component Analysis (PCA) probably remains the most popular and widespread tool in the area, with countless applications in all disciplines.

A major motivation for PCA is, of course, dimension reduction; but it also serves as an instrument in a variety of other data-analytical methods such as factor, cluster, and discriminant analysis, principal component regression, noise reduction, etc.

Principal Components are defined via the spectral decomposition of covariance matrices. They achieve an important Karhunen-Loève optimality property—projecting a d-dimensional variable $\mathbf{X}$ on the linear space spanned by its $k < d$ first principal components yields the best k-dimensional linear approximation of $\mathbf{X}$.

This, at first sight, provides a strong theoretical justification for the method.

Away from elliptical symmetry, PCA has a number of weaknesses, though:

- it requires finite moments of order 2 (hence is poorly robust and precludes the analysis of heavy-tailed observations);
- it is "centered" at the mean, which for distributions with asymmetric and non-convex shapes may not provide the best centering and lie outside the support of the distribution under study;
- the criterion used—maximization of the variance of projections—is intrinsically two-sided, which does not take into account the possible asymmetries of the data.

Above all,

- it is a highly linear technique, involving straight lines, linear combinations, linear projections, ... the L^2 geometry induced by covariance/correlation matrices.

Actually, PCA is a perfect tool for elliptically symmetric variables with finite second-order moments ...

... but elliptical symmetry is a **very** strong assumption!

Away from elliptical symmetry, PCA has a number of weaknesses, though:

- it requires finite moments of order 2 (hence is poorly robust and precludes the analysis of heavy-tailed observations);
- it is “centered” at the mean, which for distributions with asymmetric and non-convex shapes may not provide the best centering and lie outside the support of the distribution under study;
- the criterion used—maximization of the variance of projections—is intrinsically two-sided, which does not take into account the possible asymmetries of the data.

Above all,

- it is a highly linear technique, involving straight lines, linear combinations, linear projections, ... the L^2 geometry induced by covariance/correlation matrices.

Actually, PCA is a perfect tool for elliptically symmetric variables with finite second-order moments ...

... but elliptical symmetry is a **very** strong assumption!

Based on measure transportation ideas, we are proposing a new concept of "Principal Curves," which

- does not require finite second-order moments
- is fully nonlinear (with self-induced (via monotone transportation) nonlinearities—data-driven in the sample)
- is centered at a measure-transportation-based median region (in dimension $d > 4$, not necessarily a point)
- is "directional," that is, sequentially selects $2d$ oriented "halfcurves," taking asymmetries into account, rather than d full straight lines that don't.

1.2. Principal Curves versus Principal Components.

Below are some (simulation-based) illustrations of the differences between classical Principal Components and the proposed Principal Curves.

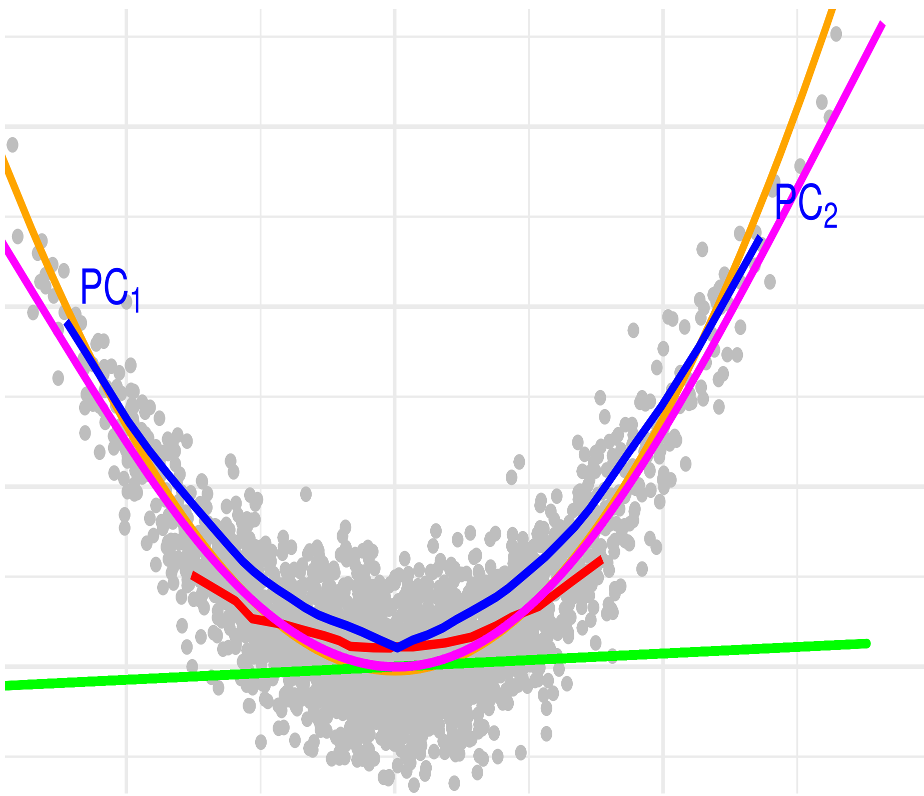
Based on measure transportation ideas, we are proposing a new concept of "Principal Curves," which

- does not require finite second-order moments
- is fully nonlinear (with self-induced (via monotone transportation) nonlinearities—data-driven in the sample)
- is centered at a measure-transportation-based median region (in dimension $d > 4$, not necessarily a point)
- is "directional," that is, sequentially selects $2d$ oriented "halfcurves," taking asymmetries into account, rather than d full straight lines that don't.

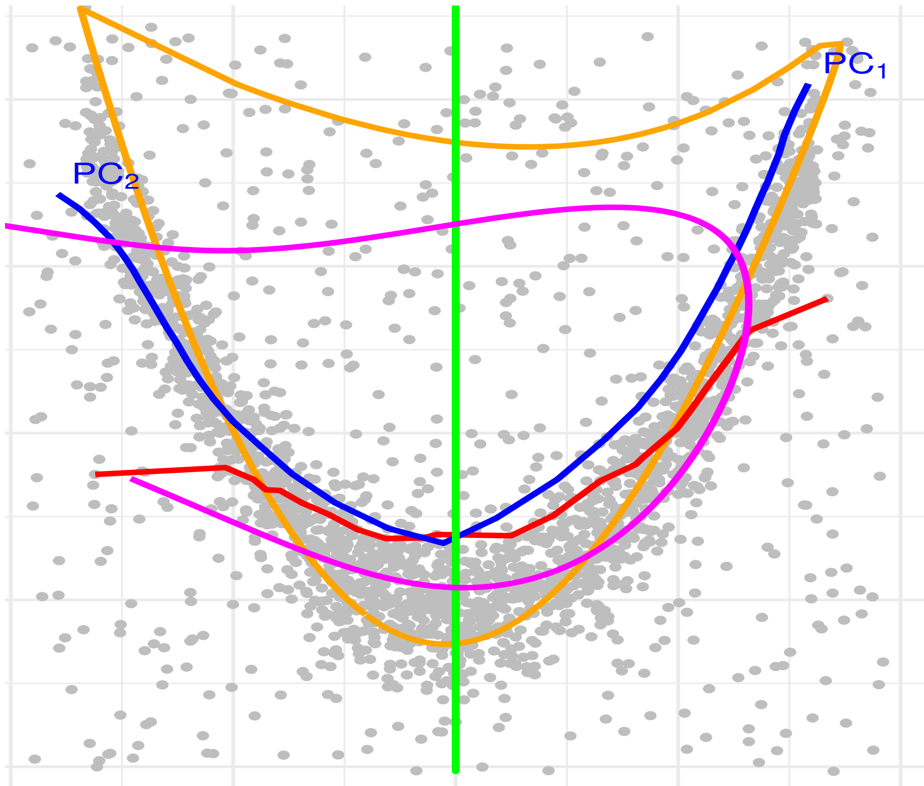
1.2. Principal Curves versus Principal Components.

Below are some (simulation-based) illustrations of the differences between classical Principal Components and the proposed Principal Curves.

Under non-elliptical and heavy-tailed distributions, traditional Principal Components, typically, do a poor job while nonlinear alternatives do much better.



Nonlinear versus linear. A banana-shaped distribution: traditional PC (green); Principal Curve (Hastie and Stuetzle (1989), magenta); neural network approach (Scholz et al. (2005), Hinton and Salakhutdinov (2006), orange); Gunsilius and Schennach (2023, red); measure-transportation-based PC (blue)



Robustness issues. A noisy version of the same banana-shaped distribution; Principal Curve (Hastie and Stuetzle (1989), magenta); neural network approach (Scholz et al. (2005), Hinton and Salakhutdinov (2006), orange); Gunsilius and Schennach (2023, red); measure-transportation-based PC (blue). The noise badly impacts the concepts that are not based on measure transportation

2. *Principal components: from bidirectional linear to directional nonlinear*

2.1. An elliptical reformulation of classical definitions

The most natural context for PCA is the family of elliptically symmetric densities, which suggests a presentation of the classical concept under the assumption of elliptical symmetry.

This approach, as we shall see, will naturally extend to a nonlinear, nonelliptical context.

Notation

- $\mathbf{X} := (X_1 \ldots, , X_d)'$ a d-dimensional elliptical random vector with location $\mathbf{0}$, full-rank scatter matrix $\mathbf{\Sigma}$, and *radial density* $\mathfrak{f}$—the density of the modulus $(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{1/2} =: \|\mathbf{\Sigma}^{-1/2}\mathbf{X}\|$

- $\mathrm{P}_{\mathbf{\Sigma},\mathfrak{f}}$ the elliptical distribution of X_i

- $F_{\mathfrak{f}} : t \mapsto F_{\mathfrak{f}}(t) := \int_0^t \mathfrak{f}(u)\mathrm{d}u = \mathrm{P}_{\mathbf{\Sigma},\mathfrak{f}}\left[\|\mathbf{\Sigma}^{-1/2}\mathbf{X}\| \leq t\right]$, $t \in \mathbb{R}_+$ the corresponding *radial distribution function,*

- $\mathbf{\Lambda} = \mathrm{Diag}(\lambda_1, \ldots, \lambda_d)$ the diagonal matrix of eigenvalues of $\mathbf{\Sigma}$ in decreasing order of magnitude (for simplicity, assume that they are all distinct),

- $\mathbf{P}_1, \ldots, \mathbf{P}_d$ the corresponding eigenvectors, and

- $\mathbf{P}$ the $d \times d$ orthogonal matrix with columns $\mathbf{P}_i$, $i = 1, \ldots, d$; let $\mathbf{\Sigma}^{1/2} := \mathbf{P}'\mathbf{\Lambda}^{1/2}\mathbf{P}$.

Then, $\mathbf{\Sigma}^{-1/2}\mathbf{X}$ is spherical with radial density $\mathfrak{f}$ and radial distribution function $F_{\mathfrak{f}}$

Since $\mathbf{\Sigma}^{-1/2}\mathbf{X}$ is spherical with radial distribution function $F_{\mathfrak{f}}$,

$$\mathbf{U} := F_{\mathfrak{f}}\left(\|\mathbf{\Sigma}^{-1/2}\mathbf{X}\|\right)\frac{\mathbf{\Sigma}^{-1/2}\mathbf{X}}{\|\mathbf{\Sigma}^{-1/2}\mathbf{X}\|}$$

is *spherical uniform* over the unit ball $\mathbb{S}_d$ —which we denote as $\mathbf{U} \sim \mathrm{U}_d$.

In the terminology of measure transportation, the transformation

$$\mathbf{T}_{\mathbf{\Sigma},\mathfrak{f}}:\ \mathbf{x} \mapsto \mathbf{T}_{\mathbf{\Sigma},\mathfrak{f}}(\mathbf{x}) := F_{\mathfrak{f}}\left(\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|\right)\frac{\mathbf{\Sigma}^{-1/2}\mathbf{x}}{\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|} =: F_{\mathfrak{f}}\left(\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|\right)\mathbf{T}_{\mathbf{\Sigma}}(\mathbf{x})$$

mapping $\mathbf{X}$ to $\mathbf{U}$ is a *transport map pushing* $\mathrm{P}_{\mathbf{\Sigma},\mathfrak{f}}$ *forward to* U_d—in general, not the gradient of a convex function, though, hence not an optimal transport in the sense of Monge and Kantorovich, nor a monotone transport in the sense of McCann.

Notation:

$$\mathbf{T}_{\mathbf{\Sigma},\mathfrak{f}}\#\mathrm{P} = \mathrm{U}_d.$$

Chapter 1. Gaspard Monge

Starting from very practical problems, Gaspard Monge, in 1781, with his *Mémoire sur la Théorie des Déblais et des Remblais*, initiated a profound mathematical theory anticipating different areas of differential geometry, linear programming, nonlinear partial differential equations, and probability

666· MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS ET DES REMBLAIS.

Par M. MONGE.

LORSQU'ON doit tranſporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit tranſporter, & le nom de *Remblai* à l'eſpace qu'elles doivent occuper après le tranſport.

In 1781, Gaspard Monge (1746-1818) was teaching mathematics at the Ecole Royale du Génie, a French military engineering school. During the French Revolution and the Empire, he developed quite an active political career: he went with Bonaparte in Italy then in Egypt; he served as a Minister (Navy), and was involved in the reform of the French educational system, the foundation of the Ecole Polytechnique, where he taught for many years, and the Ecole Normale Supérieure

Monge's 1781 *Mémoire* was motivated by a very practical problem: how do you best move a given pile of sand to fill up a given hole of the same total volume? The simplest and most intuitive abstract formulation of Monge's problem is as follows

Let P_1 and P_2 denote two probability measures over (for simplicity) $(\mathbb{R}^d, \mathcal{B}^d)$.

Let $L : \mathbb{R}^{2d} \to [0, \infty]$ be a Borel-measurable loss function: $L(\mathbf{x}_1, \mathbf{x}_2)$ represents the cost of transporting $\mathbf{x}_1$ to $\mathbf{x}_2$.

- find a measurable transport map $T_{\mathrm{P}_1;\mathrm{P}_2} : \mathbb{R}^d \to \mathbb{R}^d$ that achieves the infimum

$$\inf_T \int_{\mathbb{R}^d} L(\mathbf{x}, T(\mathbf{x}))\mathrm{d}\mathrm{P}_1 \qquad \text{subject to} \quad T\#\mathrm{P}_1 = \mathrm{P}_2$$

where $T\#\mathrm{P}_1$ denotes the "push forward of P_1 by T"—a more classical statistical notation for this would be $\mathrm{P}_1^{T\mathbf{X}} = \mathrm{P}_2$.

• A map $T_{\mathrm{P}_1;\mathrm{P}_2}$ that attains this infimum is called an "optimal transport map", in short, an "optimal transport", of P_1 into P_2.

• In the sequel, we restrict to the L^2 loss function $L(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2)\|_2^2$ (Monge was considering the more difficult loss $L(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2)\|_2$).

The problem looks simple, but it isn't (leads to the so-called Monge-Ampère equations, which are nonlinear PDEs); Monge actually could not solve it.

Chapter 2. Leonid Kantorovich

One century and a half later, Monge's problem was revisited in the 1940s by Leonid Vitalievitch Kantorovich (1912-1986; Nobel Prize in Economics in 1975) in relation to the economic problem of optimal allocation of resources.

The fundamental idea behind Kantorovich's approach (when he did it, Kantorovich was not aware of Monge's contribution) consists in relaxing Monge's problem into the more general one of constructing a distribution $\gamma_{\mathrm{P}_1\mathrm{P}_2}$ on $\mathbb{R}^d \times \mathbb{R}^d$ (Kantorovich considers abstract metric spaces) minimizing

$$\int \|\mathbf{x} - \mathbf{y}\|^2 d\gamma$$

(equivalently, maximizing $\int \langle \mathbf{x}, \mathbf{y} \rangle d\gamma$)

among the family $\Gamma(\mathrm{P}_1, \mathrm{P}_2)$ of all γ's having marginals P_1 and P_2, then showing that the solution is of the form

$$\gamma_{\mathrm{P}_1\mathrm{P}_2} = (\text{identity} \times T)\#\mathrm{P}_1 = (\mathrm{P}_1, T\#\mathrm{P}_1)$$

where $T\#\mathrm{P}_1 = \mathrm{P}_2$ for some mapping (some transport) T. This solution, thus, is the distribution of a variable

$$(\mathbf{X}, T(\mathbf{X})) \ \text{ where } \ \mathbf{X} \sim \mathrm{P}_1,$$

which is supported on the graph of $\mathbf{x} \mapsto T(\mathbf{x})$—so that T is indeed a solution of Monge's problem.

The huge advantage of this new formulation is that the class $\Gamma(\mathrm{P}_1, \mathrm{P}_2)$ of feasible solutions now is convex, so that the problem reduces to a linear optimization problem over a convex set for which Kantorovich develops a powerful duality approach.

The topic attracted a renewed surge of interest in the 1990s. Still for the quadratic loss function,

• Cuesta-Albertos and Matrán (1989) established (under continuity assumptions and the existence of finite second-order moments) the existence of solutions for Monge's problem;

• Rüschendorf, and Rachev (1990) characterized these solutions in terms of gradients of convex (potential) functions.

• Brenier (1991), with his celebrated *polar factorization theorem*, independently obtained the same results and, moreover, proved the (a.s.) uniqueness of the solution.

Brenier's Polar Factorization Theorem, in the present context, implies that, for L^2 loss, if P_1 and P_2 are absolutely continuous with finite second-order moments, the solution exists, is (a.e.) unique, and is the gradient $\nabla\psi$ of some convex (potential) function ψ—a form of multivariate monotonicity .

All this, however, is about Monge's optimization problem with L^2 transportation costs, and therefore only makes sense *under finite moments of order 2* (hence compactly supported distributions).

A completely different approach was taken, in 1995, by Robert J. McCann, offering a fresh approach to the problem.

Chapter 4. Robert McCann

McCann had the intuition that the problem is of a geometric rather than analytical nature. His main result (McCann 1995) implies that

(i) for any given (absolutely continuous—no second order moments needed) P_1 and P_2, there exists a P_1-essentially unique element $\nabla\psi$ in the class of gradients of convex functions mapping P_1 to P_2 (such that $\nabla\psi\#\mathrm{P}_1 = \mathrm{P}_2$);

(ii) under the existence finite moments of order two, that mapping moreover coincides with the L^2-optimal transport of P_1 to P_2.

This is the result opening the door to statistical applications of measure transportation.

(back to the elliptical presentation of classicla principal components)

In the terminology of measure transportation, the transformation

$$\mathbf{T}_{\boldsymbol{\Sigma},\mathfrak{f}} : \mathbf{x} \mapsto \mathbf{T}_{\boldsymbol{\Sigma},\mathfrak{f}}(\mathbf{x}) := F_{\mathfrak{f}}\left(\|\boldsymbol{\Sigma}^{-1/2}\mathbf{x}\|\right) \frac{\boldsymbol{\Sigma}^{-1/2}\mathbf{x}}{\|\boldsymbol{\Sigma}^{-1/2}\mathbf{x}\|} =: F_{\mathfrak{f}}\left(\|\boldsymbol{\Sigma}^{-1/2}\mathbf{x}\|\right)\mathbf{T}_{\boldsymbol{\Sigma}}(\mathbf{x})$$

mapping $\mathbf{X}$ to $\mathbf{U}$ is a *transport map pushing* $\mathrm{P}_{\boldsymbol{\Sigma},\mathfrak{f}}$ *forward to* U_d—in general, not the gradient of a convex function, though, hence not an optimal transport in the sense of Monge and Kantorovich, nor a monotone transport in the sense of McCann.

Notation:

$$\mathbf{T}_{\boldsymbol{\Sigma},\mathfrak{f}}\#\mathrm{P} = \mathrm{U}_d.$$

That transport map factorizes into the product

$$\mathbf{T}_{\mathbf{\Sigma},\mathfrak{f}} = \mathbf{F}_{\pm}^{\mathfrak{f}} \circ \mathbf{T}_{\mathbf{\Sigma}}$$

where $\mathbf{T}_{\mathbf{\Sigma}}$ is the linear transformation $\mathbf{x} \mapsto \mathbf{\Sigma}^{-1/2}\mathbf{x}$ and $\mathbf{F}_{\pm}^{\mathfrak{f}}$ the radial transformation $\mathbf{x} \mapsto F_{\mathfrak{f}}(\|\mathbf{x}\|)\,\mathbf{x}/\|\mathbf{x}\|$.

This suggests an alternative description of $\mathbf{X}$'s principal components, formally bypassing the recourse to the eigenvectors and eigenvalues of $\mathbf{\Sigma}$.

Denoting by $\mathcal{S}_{d-1}$ the unit hypersphere in $\mathbb{R}^d$ and assuming that $\mathfrak{f}$ is strictly positive over the positive real line $\mathbb{R}_+$ (this can be relaxed, e.g. to strictly positive on an interval $(0, T)$ for some $0 < T < \infty$) so that $\mathbf{T}_{\mathbf{\Sigma},\mathfrak{f}}$ is invertible, define

$$\pm\mathbf{U}_{1;\,\text{ell}} := \arg\max_{\mathbf{u}\in\mathcal{S}_{d-1}} \mathrm{E}\left[\left\|\mathbf{T}^{-1}_{\mathbf{\Sigma},\mathfrak{f}}(\varrho\mathbf{u})\right\|^2\right]$$

and

$$\pm\mathbf{U}_{i;\,\text{ell}} := \underset{\mathcal{S}_{d-1}\ni\mathbf{u}\perp\mathbf{U}_{j;\,\text{ell}},\ j=1,\ldots,i-1}{\arg\max} \mathrm{E}\left[\left\|\mathbf{T}^{-1}_{\mathbf{\Sigma},\mathfrak{f}}(\varrho\mathbf{u})\right\|^2\right] \quad i = 2, \ldots, d$$

where ϱ is uniform over $[0, 1]$.

Based on this, define the *ith principal halflines* of $\mathbf{X}$ (of P) as

$$\mathcal{L}_i^+ := \{\mathbf{T}_{\mathbf{\Sigma},\mathfrak{f}}^{-1}(r\mathbf{U}_{i;\,\text{ell}}) | r \in \mathbb{R}_+\} \quad \text{and} \quad \mathcal{L}_i^- := \{\mathbf{T}_{\mathbf{\Sigma},\mathfrak{f}}^{-1}(-r\mathbf{U}_{i;\,\text{ell}}) | r \in \mathbb{R}_+\}.$$

The ith principal component of $\mathbf{X} \sim \mathrm{P}_{\mathbf{\Sigma},\mathfrak{f}}$ then is

$$\mathbf{P}_i := \pm\langle \mathbf{U}_{i;\,\text{ell}}, \mathbf{X}\rangle \mathbf{U}_{i;\,\text{ell}},$$

the signed projection of $\mathbf{X}$ on $\pm\mathbf{U}_{i;\,\text{ell}}$ or the halflines $\pm\mathcal{L}_i^+$.

Considering $i = 1$, we have

$$\begin{aligned}\pm\mathbf{U}_{1;\,\text{ell}} &:= \arg\max_{\mathbf{u}\in\mathcal{S}_{d-1}} \mathrm{E}\left[\left\|\mathbf{T}_{\mathbf{\Sigma},\mathfrak{f}}^{-1}(\varrho\mathbf{u})\right\|^2\right] = \arg\max_{\mathbf{u}\in\mathcal{S}_{d-1}} \mathrm{E}\left[\left\|\mathbf{T}_{\mathbf{\Sigma}}^{-1}(F_{\mathfrak{f}}^{-1}(\varrho))\mathbf{u}\right\|^2\right] \\ &= \arg\max_{\mathbf{u}\in\mathcal{S}_{d-1}} \mathrm{E}\left[\left(F_{\mathfrak{f}}^{-1}(\varrho)\right)^2 \|\mathbf{\Sigma}^{1/2}\mathbf{u}\|^{\mathbf{2}}\right] = \arg\max_{\mathbf{u}\in\mathcal{S}_{d-1}} \left[\mathbf{u}'\mathbf{\Sigma}\mathbf{u}\right] = \pm\mathbf{P}_1,\end{aligned}$$

where $\mathbf{P}_1$ is P's traditional first principal direction ($\mathbf{\Sigma}$'s first eigenvector).

The case for $i = 2, \ldots, d$ is entirely similar.

As for the maximum in the definition of $\pm\mathbf{U}_{i;\,\text{ell}}$, it is easily seen to be $\sigma_{\mathfrak{f}}^2\lambda_i$, $i = 1, \ldots, d$, where λ_i is $\mathbf{\Sigma}$'s ith eigenvalue.

This, which, for an elliptical distribution P, constitutes an alternative definition of the classical concept of principal components, involves the transport

$$\mathbf{T}_{\mathbf{\Sigma},\mathfrak{f}} = \mathbf{F}_{\pm}^{\mathfrak{f}} \circ \mathbf{T}_{\mathbf{\Sigma}}$$

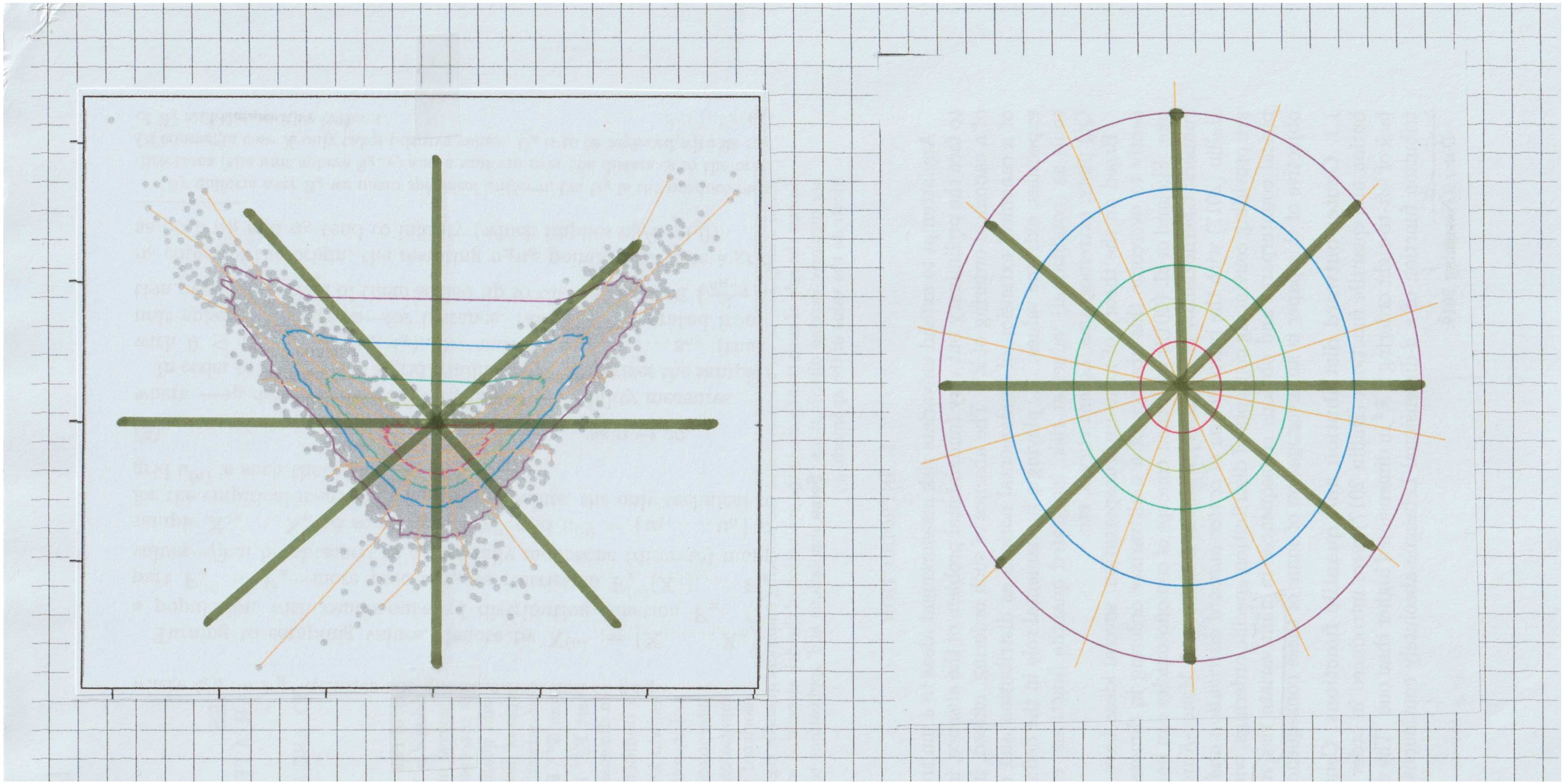
pushing the elliptical P to the uniform U_d over the unit ball.

That transport, however,

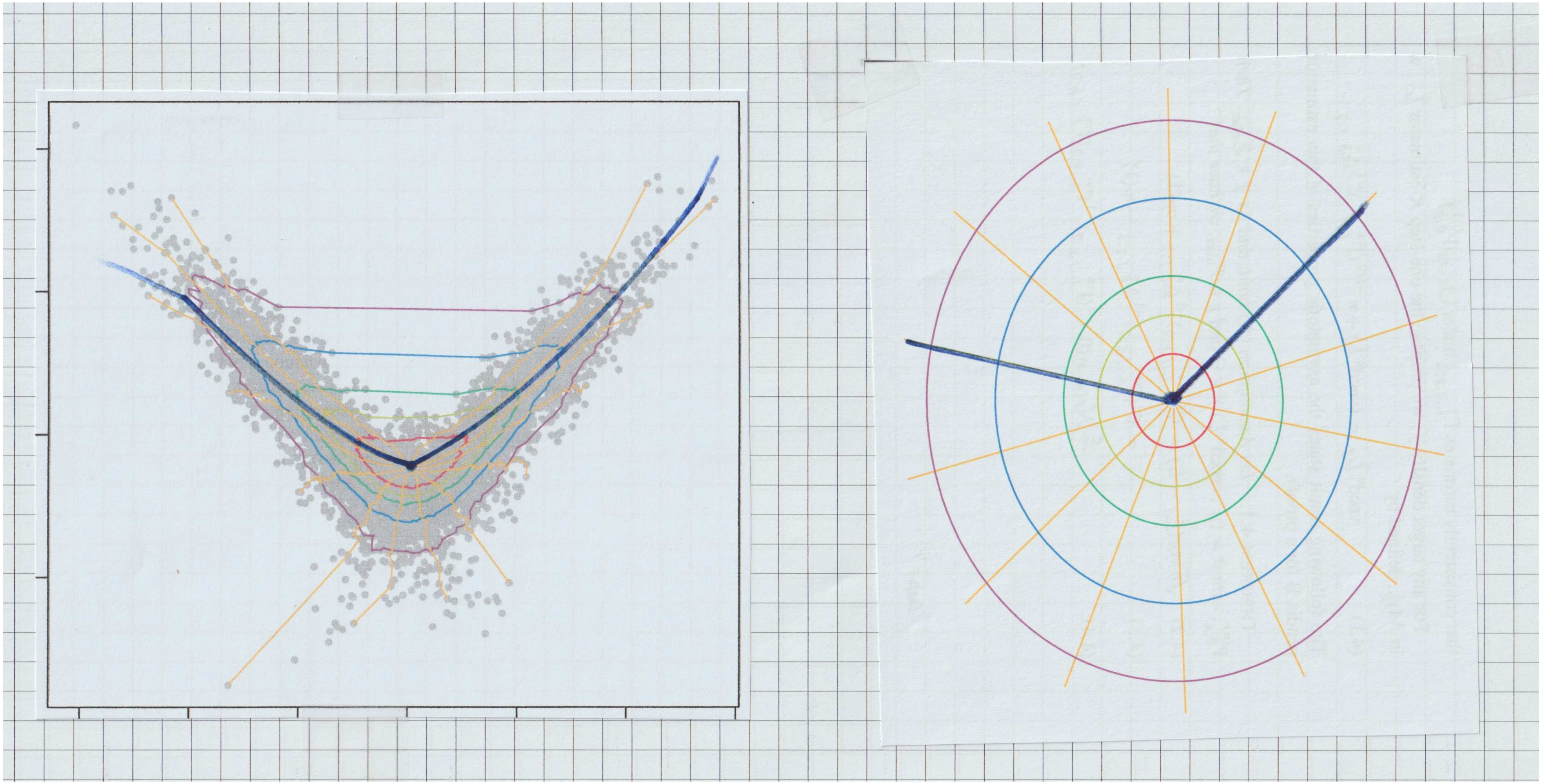
- makes no sense for a nonelliptical P (no radial distribution available)
- in general, is not the gradient of a convex function,[a] hence is neither optimal (in the sense of Monge and Kantorovich) nor monotone (in the sense of Mc Cann—that is, cyclically monotone).

[a] unless $\mathbf{\Sigma}$ is a multiple of $\boldsymbol{I}_d$

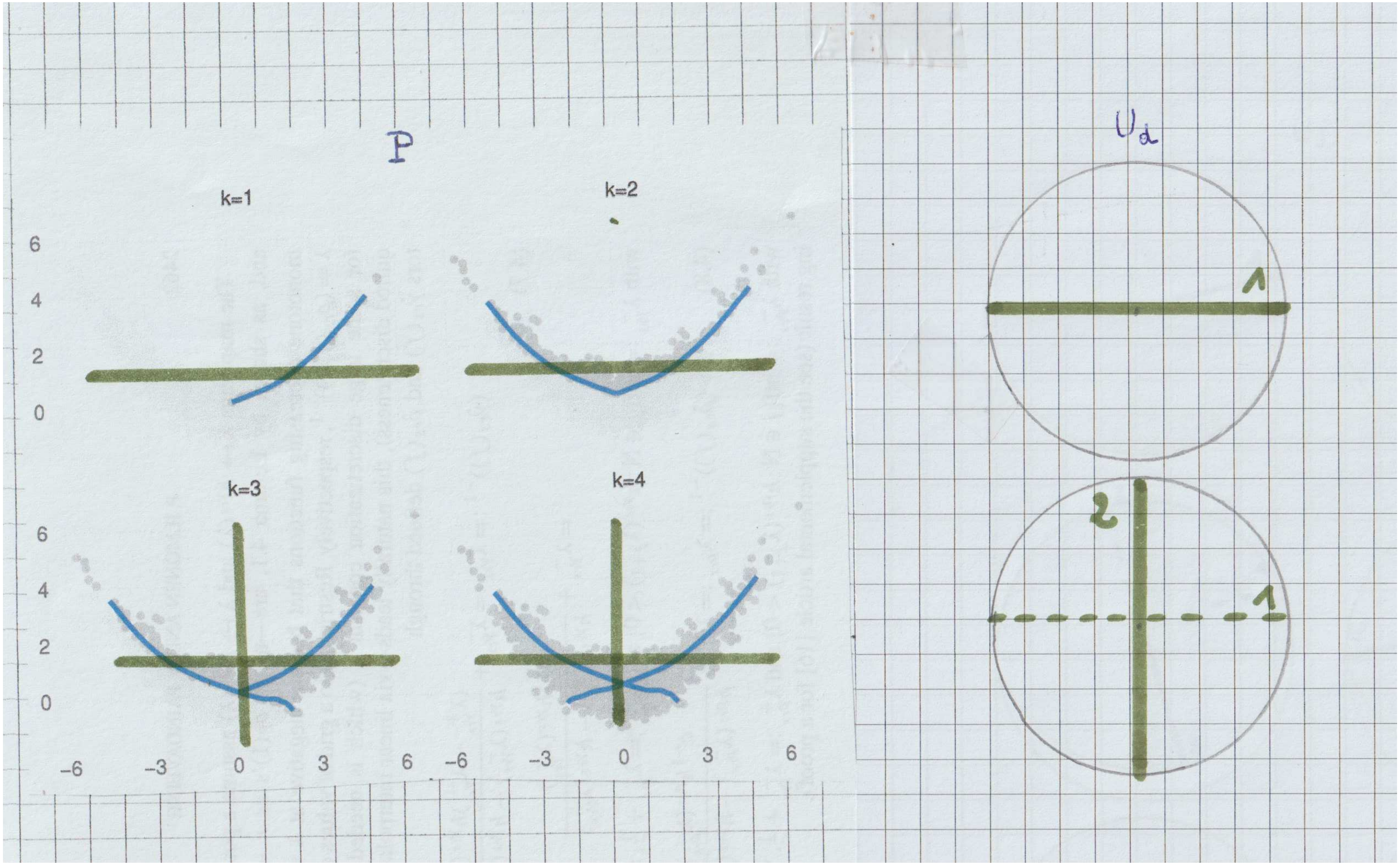
The transport map $\mathbf{T}_{\boldsymbol{\Sigma},\mathfrak{f}}$



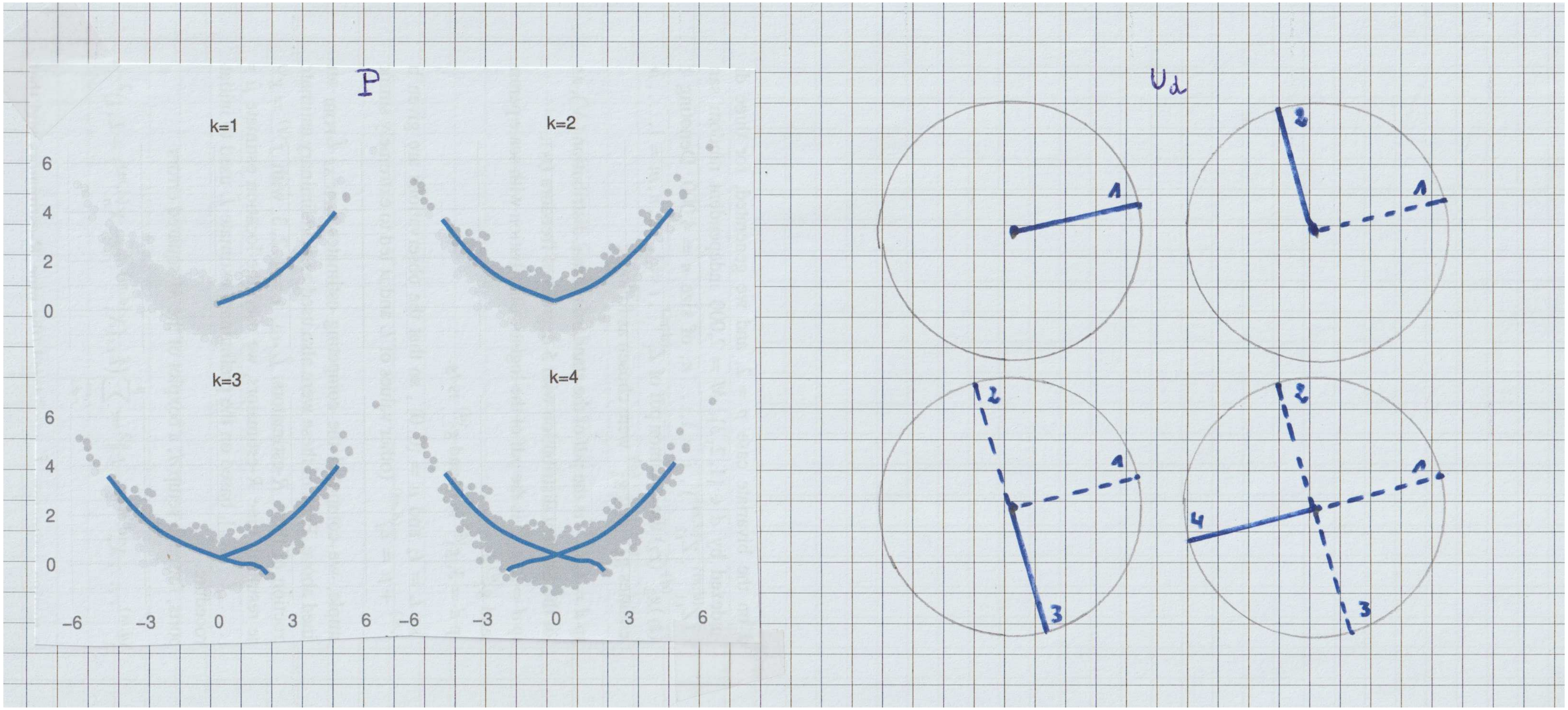
The monotone transport map $\mathbf{F}_{\pm}$



The transport map $\mathbf{T}_{\boldsymbol{\Sigma},\mathfrak{f}}$



The monotone transport map $\mathbf{F}_{\pm}$



2.2. Nonelliptical case: a measure-transportation-based definition

The idea is very simple: in the sequential choice of principal directions, replace the non-optimal linear transport

$$\mathbf{T}_{\mathbf{\Sigma},\mathfrak{f}} : \mathbf{x} \mapsto \mathbf{T}_{\mathbf{\Sigma},\mathfrak{f}}(\mathbf{x}) := F_{\mathfrak{f}}\left(\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|\right)\frac{\mathbf{\Sigma}^{-1/2}\mathbf{x}}{\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|} =: F_{\mathfrak{f}}\left(\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|\right)\mathbf{T}_{\mathbf{\Sigma}}(\mathbf{x})$$

pushing an elliptical P forward to U_d with the optimal transport $\mathbf{F}_{\pm}$ pushing an elliptical as well as nonelliptical P forward to U_d.

About pushing P forward to U_d, McCann (1985) tells us the following

Let $\mathbf{X} \sim \mathrm{P}$ be Lebesgue-absolutely continuous with (for simplicity of exposition: this can be relaxed) nonvanishing density.

(i) there exists a P*-almost unique gradient of convex function* $\mathbf{F}_{\mathrm{P}\pm}$ *such that*

$$\mathbf{F}_{\mathrm{P}\pm}\#\mathrm{P} = \mathrm{U_d}.$$

(ii) Moreover, if Var$(\mathbf{X}) < \infty$*, then* $\mathbf{F}_{\mathrm{P}\pm}$ *minimizes (the expected transport cost)*

$$\int_{\mathbb{R}^d} \|\mathbf{x} - \mathbf{F}(\mathbf{x})\|^2 \mathrm{dP}(\mathbf{x})$$

among all mappings $\mathbf{F}$ *from* $\mathbb{R}^d$ *to* $\mathbb{S}_d$ *such that* $\mathbf{F}\#\mathrm{P} = \mathrm{U}_d$*.*

Figalli (2018) moreover showed that

(iii) $\mathbf{F}_{\mathrm{P}\pm}$ *is a homeomorphism between* $\mathbb{S}_d \setminus \{\mathbf{0}\}$ *and* $\mathbb{R}^d \setminus \{\mathbf{F}_{\mathrm{P}\pm}(\mathbf{0})\}$*, hence admits an a.s. unique continuous inverse* $\mathbf{Q}_\pm$*—the a.s. unique gradient of convex function such that* $\mathbf{Q}_\pm\#\mathrm{U}_d = \mathrm{P}$*.*

Finally,

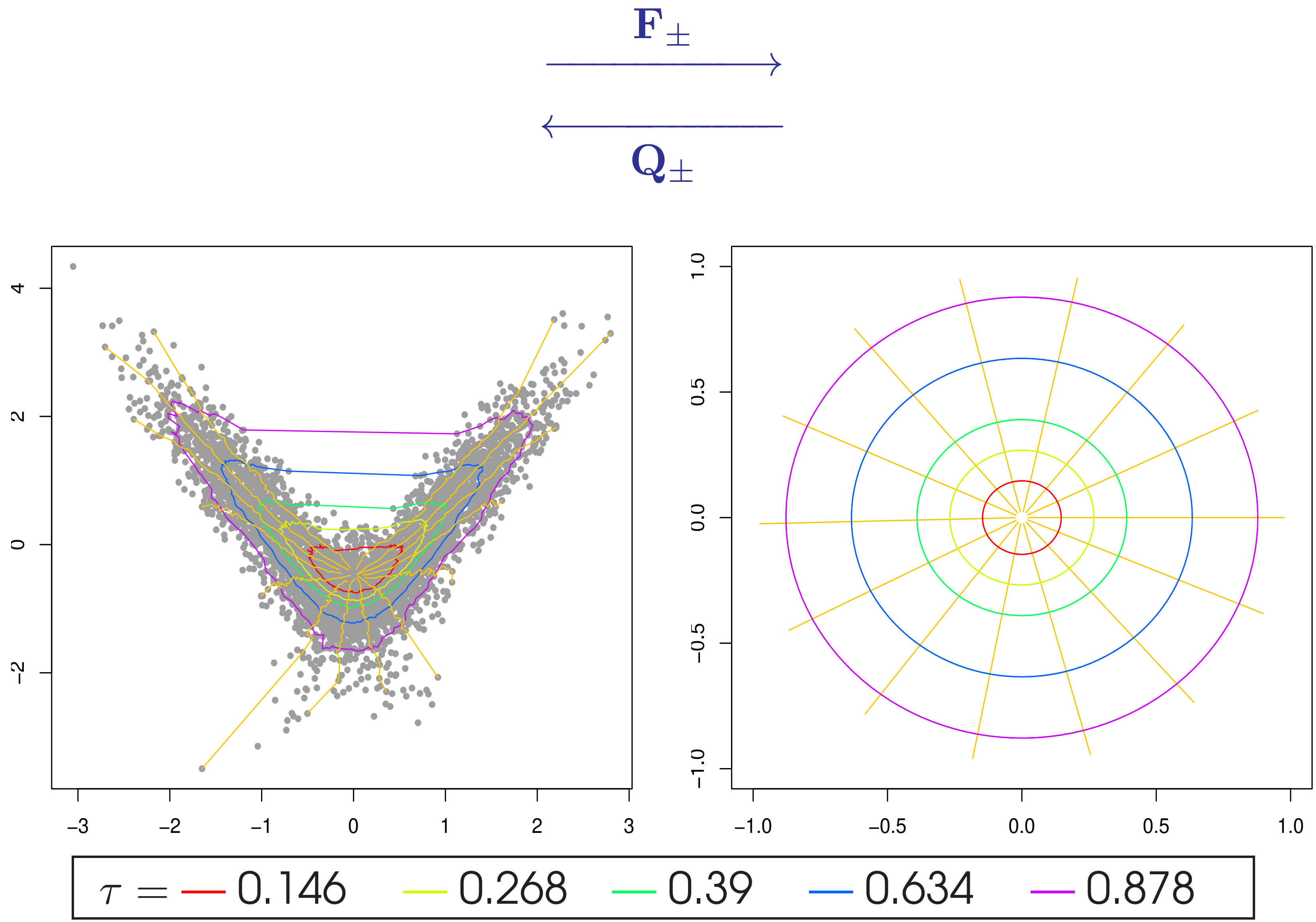
(iv) For $d = 1$*,* $\mathbf{F}_{\mathrm{P}\pm} = 2F_{\mathrm{P}} - 1$*.*

This suggests the following definitions (Hallin, del Barrio, Cuesta-Albertos, and Matrán, *Annals of Statistics* 2021) for a center-outward distribution function and a center-outward quantile function in dimension d.

Let $\mathbf{X} \sim \mathrm{P}^{\mathbf{X}}$ be absolutely continuous over $\mathbb{R}^d$. Call *center-outward distribution function of* $\mathrm{P}^{\mathbf{X}}$ the a.e. unique gradient of convex function $\mathbf{F}_{\mathrm{P}\pm}$ pushing $\mathrm{P}^{\mathbf{X}}$ forward to the spherical uniform U_d over $\mathbb{S}_d$.

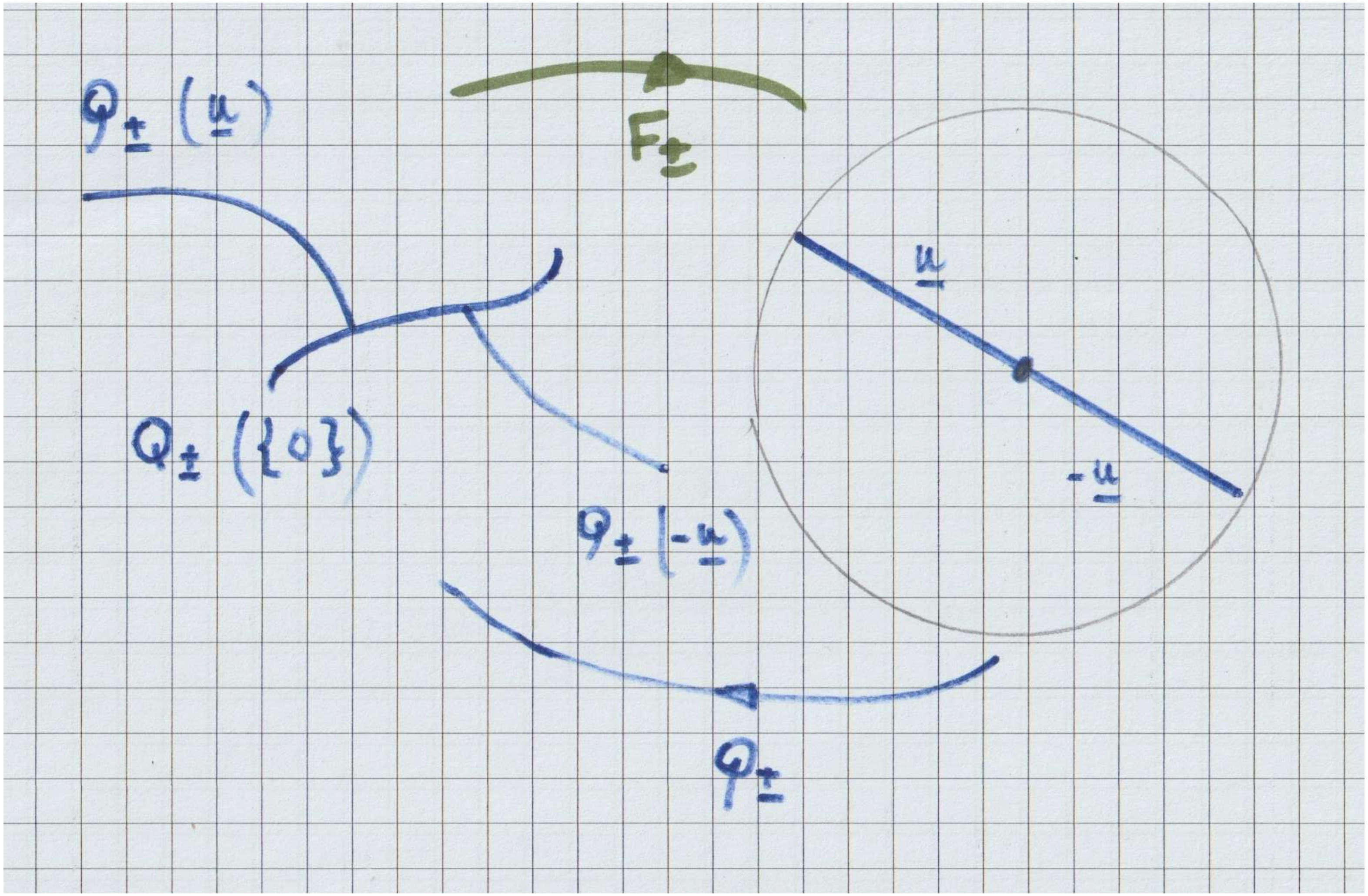
If $\mathrm{P}^{\mathbf{X}}$, e.g., has nonvanishing density over a convex support, it follows from Figalli (2018) that $\mathbf{F}_{\mathrm{P}\pm}$ is a homeomorphism between $\mathbb{S}_d \setminus \{\mathbf{0}\}$ and $\mathbb{R}^d \setminus \mathbf{F}^{-1}_{\mathrm{P}\pm}(\{\mathbf{0}\})$, hence has a continuous (on $\mathbb{S}_d \setminus \{\mathbf{0}\}$) inverse $\mathbf{Q}_{\mathrm{P}\pm}$ which is also the continuous (on $\mathbb{R}^d \setminus \mathbf{F}^{-1}_{\mathrm{P}\pm}(\{\mathbf{0}\})$) gradient of a convex function.

Call $\mathbf{Q}_{\mathrm{P}\pm}$ the *center-outward quantile function of* $\mathrm{P}^{\mathbf{X}}$.



$\mathbf{F}_\pm$
$\mathbf{Q}_\pm$
4
2
0
−2
−3
−2
−1
0
1
2
3
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
$\tau =$ 0.146 0.268 0.39 0.634 0.878

For $d > 4$, moreover, $\mathbf{Q}_{\mathrm{P}\pm}(\{\mathbf{0}\}) := \mathbf{F}_{\mathrm{P}\pm}^{-1}(\{\mathbf{0}\})$ could be a compact set with measure zero, with a discontinuity at $r = 0$ of $r \mapsto \mathbf{Q}_{\pm}(r\mathbf{u})$:



... which calls for disentangling $\mathbf{u}$ and $-\mathbf{u}$ and adopting a directional (oriented) definition of principal directions (curves).

Our definition of nonlinear directional principal curves mimics the previously given elliptical definition traditional principal components, but

- replaces the "linear" non-optimal transport $\mathbf{T}_{\boldsymbol{\Sigma},\mathfrak{f}}$ with the optimal one $\mathbf{F}_{\pm}$
- selects $2d$ half-curves instead of d full straight lines
- based on expected curve length
- replaces linear inner products and orthogonal projections with "curvilinear" ones

Our definition of nonlinear directional principal curves mimics the previously given elliptical definition traditional principal components, but

- replaces the "linear" non-optimal transport $\mathbf{T}_{\boldsymbol{\Sigma},\mathfrak{f}}$ with the optimal one $\mathbf{F}_{\pm}$
- selects $2d$ half-curves instead of d full straight lines
- based on expected curve length
- replaces linear inner products and orthogonal projections with "curvilinear" ones

Assume, for simplicity, that P belongs to the family $\mathcal{P}_{\pm}^{d}$ of Lebesgue-absolutely continuous distributions with support $\mathbb{R}^d$ and density f such that, for any compact subset C of $\mathbb{R}^d$, there exists constants $0 < b_C < B_C < \infty$ such that $b_C \leq f(\mathbf{x}) \leq B_C$ for all $\mathbf{x} \in C$ (these are the assumptions made in Figalli (2018) to obtain homeomorphisms; they can be relaxed).

Denote by $\mathbf{F}_{\pm}$ and $\mathbf{Q}_{\pm}$ the center-outward distribution and quantile functions of P, the existence of which does not require any assumption of elliptical symmetry nor finite variance:

$$\mathbf{F}_{\pm}\#\mathrm{P} = \mathrm{U}_d \qquad \mathbf{Q}_{\pm}\#\mathrm{U}_d = \mathrm{P}.$$

Define, parallel to what we did before based on $\mathbf{T}_{\mathbf{\Sigma},\mathfrak{f}}$, the *first nonlinear directional principal direction* of P (of $\mathbf{X} \sim \mathrm{P}$) as

$$\mathbf{U}_{\pm,1} := \underset{\mathbf{u}\in\mathcal{S}_{d-1}}{\arg\max}\ \mathrm{E}\left[L^2\left(\mathbf{Q}_{\pm}(\rho\mathbf{u})\right)\right]$$

where

$$L^2\left(\mathbf{Q}_{\pm}(r\mathbf{u})\right) := \left[\int_0^r \left\|\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Q}_{\pm}(t\mathbf{u})\right\| \mathrm{d}t\right]^2, \quad r \in (0,1).$$

with $\rho \sim \mathrm{U}_{[0,1]}$ and $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Q}_{\pm}(t\mathbf{u}) := \left(\frac{\mathrm{d}}{\mathrm{d}t}Q_{\pm,1}(t\mathbf{u}), \ldots, \frac{\mathrm{d}}{\mathrm{d}t}Q_{\pm,d}(t\mathbf{u})\right)'.$

Similarly define the *ith nonlinear directional principal direction* of P (of $\mathbf{X} \sim \mathrm{P}$) as

$$\mathbf{U}_{\pm,i} := \arg\max_{\substack{\mathcal{S}_{d-1} \ni \mathbf{u} \perp \mathcal{U}^{+}_{i-1} \\ \text{or } \mathbf{u} \in -\mathcal{U}^{+}_{i-1} \setminus \mathcal{U}^{+}_{i-1}}} \mathrm{E}\left[L^2\left(\mathbf{Q}_{\pm}(\rho\mathbf{u})\right)\right] \quad i = 2, \ldots, 2d$$

Note that $L^2\left(\mathbf{Q}_{\pm}(r\mathbf{u})\right)$ is the squared length of the image (a curve) by $\mathbf{Q}_{\pm}$ of the interval $(0, r\mathbf{u}]$. The criterion to be maximized is the expected squared curve length $\mathrm{E}[L^2(\mathbf{Q}_{\pm}(\rho\mathbf{u}))]$ instead of the expected squared modulus $\mathrm{E}[\|\mathbf{T}^{-1}_{\mathbf{\Sigma},\mathfrak{f}}(\varrho\mathbf{u})\|^2]$.

Associated with the ith principal *direction* $\mathbf{U}_{\pm,i}$, define the *ith principal curve* of P (of $\mathbf{X} \sim \mathrm{P}$) as the image $\mathcal{L}_{\pm,i} := \left\{\mathbf{Q}_{\pm}(r\mathbf{U}_i) | r \in (0,1)\right\}$ of the radius $\{r\mathbf{U}_i | r \in (0,1)\}$ by the quantile function $\mathbf{Q}_{\pm}$—the nonlinear counterpart of the principal halfline $\mathcal{L}^{+}_{i}$.

Finally, define the ith *directional non-linear principal component* of P (of $\mathbf{X} \sim \mathrm{P}$) as

$$P_{\pm,i} := \int_0^{\|\mathbf{F}_{\pm}(\mathbf{X})\|} \left\langle \frac{\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Q}_{\pm}(t\mathbf{U}_{\pm,i})}{\|\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Q}_{\pm}(t\mathbf{U}_{\pm,i})\|}, \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Q}_{\pm}\left(t\frac{\mathbf{F}_{\pm}(\mathbf{X})}{\|\mathbf{F}_{\pm}(\mathbf{X})\|}\right)\right\rangle \mathrm{d}t \quad i = 1, \ldots, 2d.$$

The integrand

$$\left\langle \frac{\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Q}_{\pm}(t\mathbf{U}_{\pm,i})}{\|\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Q}_{\pm}(t\mathbf{U}_{\pm,i})\|}, \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Q}_{\pm}\left(t\frac{\mathbf{F}_{\pm}(\mathbf{X})}{\|\mathbf{F}_{\pm}(\mathbf{X})\|}\right)\right\rangle$$

is the modulus of the projections of the derivative wrt t, computed at $t\mathbf{F}_{\pm}(\mathbf{X})/\|\mathbf{F}_{\pm}(\mathbf{X})\|$, of the image by $\mathbf{Q}_{\pm}$ of a point in the unit ball moving from $\mathbf{0}$ to $\mathbf{F}_{\pm}(\mathbf{X})$ along the radius pointing to $\mathbf{F}_{\pm}(\mathbf{X})$ onto the tangent to the image by $\mathbf{Q}_{\pm}$ of the ith *principal directions* $\mathbf{U}_{\pm,i}$.

That integral, therefore, can be interpreted as the length of a nonlinear projection of $\mathbf{X}$ onto $\mathcal{L}_{\pm,i}$.

2.3. Comparison with Gunsilius and Schennach (*JASA* 2023)..

Another measure-transportation-based approach has been taken by Gunsilius and Schennach (*JASA* 2023).

At population level, it consists of

• considering the optimal transport $\mathbf{T} : \mathbb{R}^d \to \mathbb{R}^d$ between $\mathbf{X} \sim \mathrm{P}$ with density f and $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
• selecting thedirections $\mathbf{u}_i$, $i = 1, \ldots, d$ as the eigenvectors of the average log-Jacobian matrix

$$\bar{\mathbf{J}} := \int f(\mathbf{x}) \ln(\mathbf{J}(\mathbf{x}))\mathrm{d}\mathbf{x}$$

of the inverse mapping $\mathbf{T}^{-1}$, where $\mathbf{J}(\mathbf{x}) := \nabla\mathbf{T}(\mathbf{x})$, and $\ln(\mathbf{J}(\mathbf{x}))$ is the matrix logarithm of the positive definite matrix $\mathbf{J}(\mathbf{x})$; these directions are maximizing, subject to the usual orthogonality constraints, their “contribution”

$$H_{\mathbf{u}_i} = c + \mathbf{u}_i ? \bar{\mathbf{J}}\mathbf{u}_i$$

to the entropy H_f of f (c is a constant corresponding to entropy of Gaussian distribution).

A comparison with our approach is not straightforward:

- their reference distribution is the spherical Gaussian, not the spherical uniform;
- they are selecting couples $\pm\mathbf{u}$ of directions, not directions;
- their criterion (for direction $\mathbf{u}$) involves the whole space $\mathbb{R}^d$, while we are using only the $\mathbf{Q}_\pm$ image of $\mathbf{u}$ (a curve), yielding the expected spread along that curve.

Our criterion, which selects principal directions $\mathbf{u}_i$ on the basis of the expected length of the sign curve running from $\mathbf{Q}_\pm(\mathbf{0}^+)$ to $\mathbf{Q}_\pm(\mathbf{X})$ is more closely in line with the traditional definition.

2.4. Empirical version.

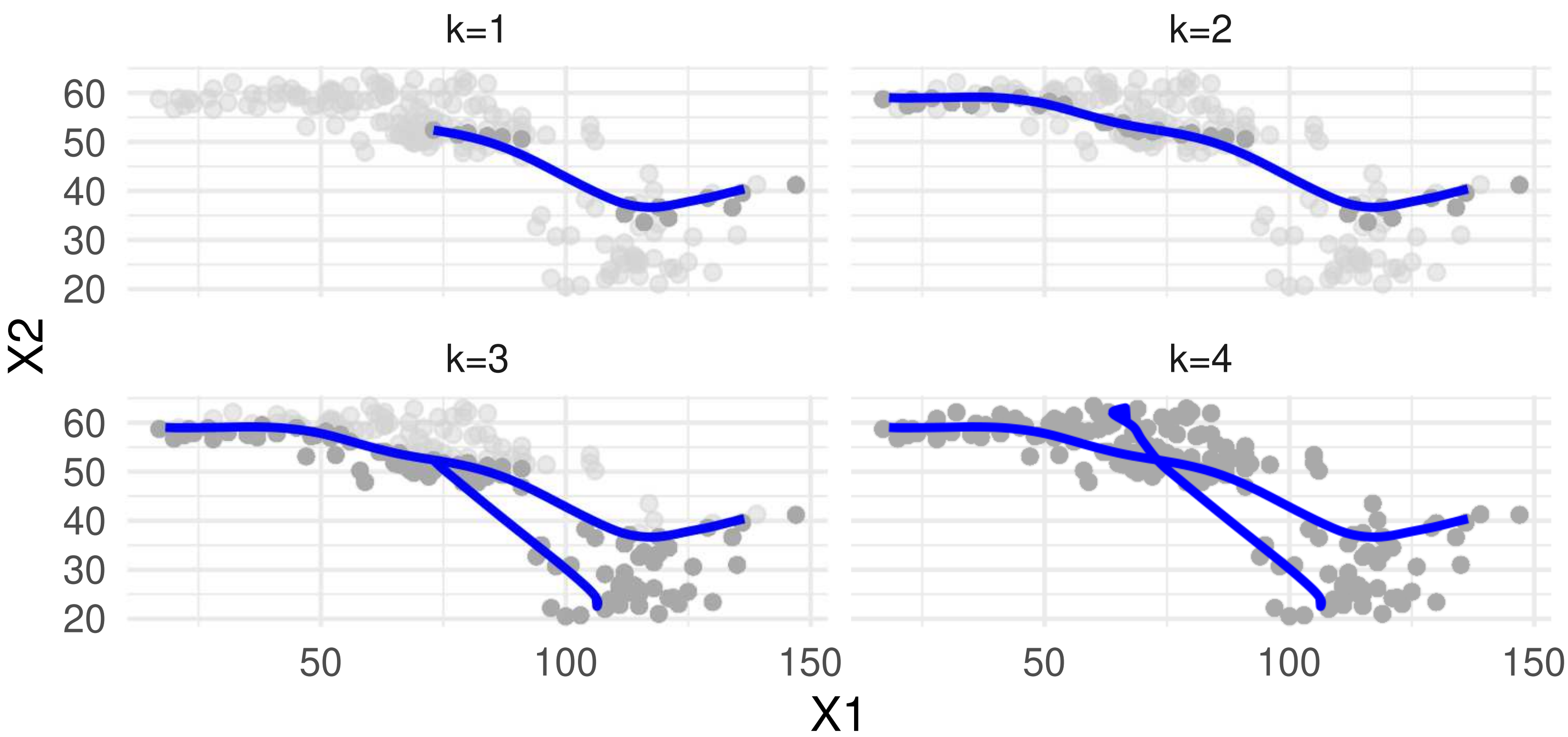
Based on empirical transport to a regular grid of the unit ball and discrete approximations of the integral criterion involved.

(in progress – details skipped)

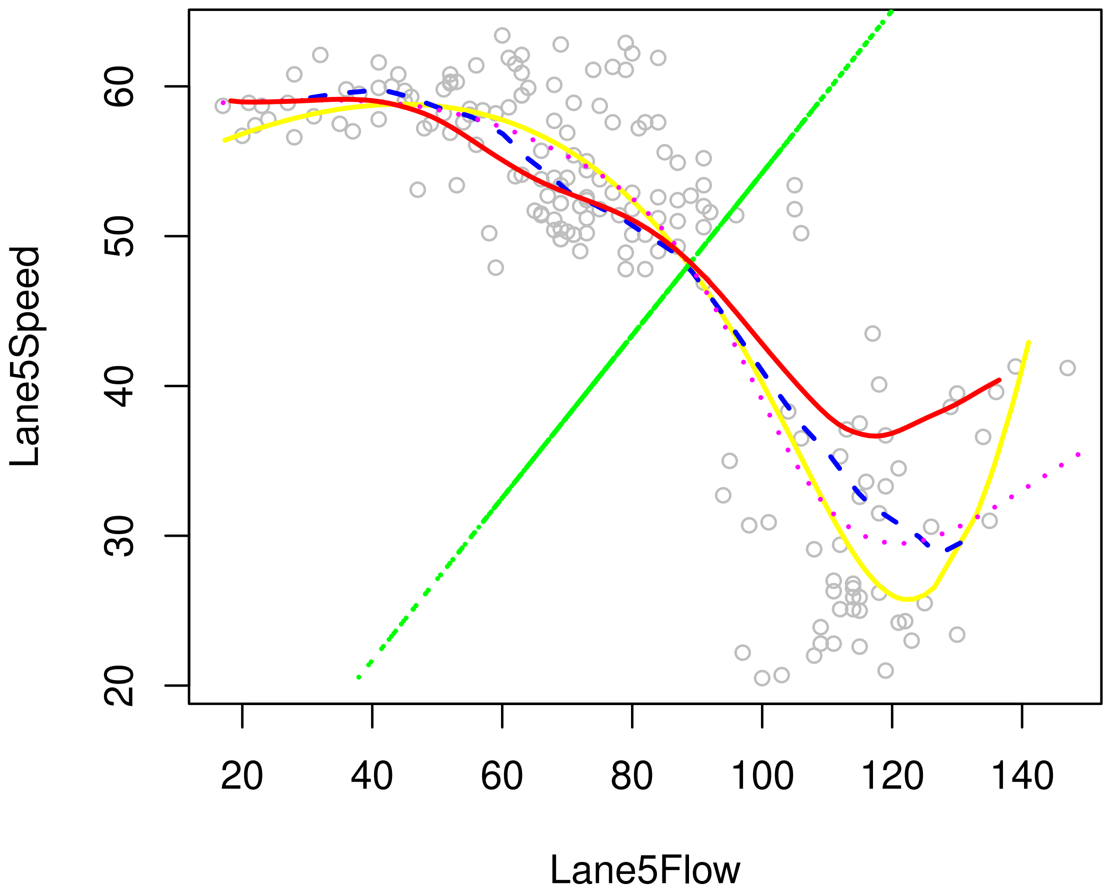
3. A real data application: drivers' profiles

Traffic data collected by a loop detector at a fixed vehicle detector station (VDS number 1202263) on Line 5 of the California Freeway SR57-N. The observations span from 9:00 AM on July 9, 2007, to 10:00 PM on July 10, 2007. Originally recorded at 30-second intervals, the data were aggregated into 5-minute intervals. Each point in the dataset represents a 5-minute interval and includes two variables: flow, the number of vehicles passing the detector, and speed, the average speed (in miles per hour) of those vehicles. This forms a "fundamental diagram" of traffic, showing the relationship between traffic flow and speed.

(1) Einbeck, J.; Evers, L. LPCM: Local Principal Curve Methods. R Package Version 0.47-4. Available online: https://CRAN.R-project.org/package=LPCM (accessed on 5 March 2024).



Principal directions for the California traffic flow data:
$X_1 = $ flow, X_2= average speed.



One-dimensional approximations of the California traffic flow data. Our first two one-sided principal curves (red), the Gunsilius and Schennach first optimal curve (blue), the principal curves of Hastie and Stuetzle (magenta) and the neural network principal curve of Scholz et al. (yellow), and the first traditional principal component (green).

4. *Conclusions*

- Traditional principal component concepts are deeply marked by the linearity and symmetry properties stemming from implicit Gaussian or elliptical assumptions;

- imposing such linear features is fine under Gaussian or elliptical distributions, but they are inappropriate under nonelliptical ones, for which they lead to misleading conclusions;

- our measure-transportation-based concepts are replacing the linear features imposed by the classical definitions with self-generated nonlinearities and proceed with one-sided selections of principal directions, which accounts for possible asymmetries;

- getting rid of all linearity features, we are losing an important feature of classical Principal Components: uncorrelatedness—a feature that makes no sense, though, in this nonlinear context.