Summary. Mathematical modeling of many real world problems involves describing distributions via lower dimensional sets. Examples are found in several fields, such as urban planning (transport networks, utility grids and irrigation networks) and data analysis (one-dimensional representation of distributions of data points). A common trait in these systems is the trade-off between the "quality" of the network and its complexity. The mathematical modeling of such systems introduces an energy (objective functional in data analysis) which rewards quality (approximation error in data analysis), and penalizes complexity. The aim is to minimize the energy among a given family of admissible sets. The average-distance problem (introduced in [4]) is one such model. Despite the apparently simple formulation, the non-local nature of the problem makes its analysis challenging from both theoretical and numerical point of view. My contributions include addressing fundamental theoretical problems such as regularity of minimizers, overcoming modeling issues by introducing improved functionals suitable for particular problems, and analyzing discriminating features of different models.

1. Low dimensional representation of measures: the average-distance problem

One model for dimension reduction is the so-called "average-distance problem", introduced by Buttazzo, Oudet and Stepanov [4].

Description of the problem. Let μ be a nonnegative, compactly supported, finite measure, $n \geq 2$, $p \geq 1$, λ be a penalization coefficient, the average-distance problem involves minimizing

$$\int_{\mathbb{R}^n} d(x,\Sigma)^p d\mu(x) + \lambda \mathcal{H}^1(\Sigma), \qquad d(x,\Sigma) := \inf_{y\in\Sigma} |x-y| \tag{1}$$

where $\mathcal{H}^1$ denotes the 1-Hausdorff distance, and Σ varies among compact, path-wise connected sets with Hausdorff dimension equal to 1.

The average-distance problem finds application in several fields.

- Urban planning: denoting by μ a distribution of population, and by Σ a transit network (to be built), the term $\int_{\mathbb{R}^n} d(x,\Sigma)^p d\mu(x)$ represets the "average distance" between passengers and network, while $\lambda\mathcal{H}^1(\Sigma)$ represents the cost to build the network. The goal of minimizing (1) is to find the network which best serves the population, under cost considerations (Buttazzo, Pratelli and Stepanov [6], Buttazzo and Santambrogio [8]).
- Data analysis: the formulation in (1) is closely related to the notion of "principal curves", introduced by Hastie [20], and extensively studied and used in principal component analysis (cf. [16, 21, 39, 19, 1, 12, 24]). In this setting, denoting by μ a distribution of data points, the term $\int_{\mathbb{R}^n} d(x,\Sigma)^p d\mu(x)$ represets the approximation error induced associated to Σ, while $\lambda\mathcal{H}^1(\Sigma)$ is a cost associated to its complexity. Thus minimizing (1) corresponds to finding the best approximation, balancing approximation error and complexity. Choice $p = 2$ is very common in statistics: in this case, if μ is sum of finitely many Dirac measures (i.e. $\mu = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$), given a minimizer Σ it follows $\int_{\mathbb{R}^n} d(x,\Sigma)^2 d\mu(x) = \frac{1}{N}\sum_{i=1}^N d(x,\Sigma)^2$, which is the mean square error. Choice $p = 1$ is also relevant, since it is more robust to outliers, i.e. adding an extra data point further away from the original distribution does not change the representation drastically.

Existence of minimizers for (1) follows from classic compactness theorems. Uniqueness is rarely true. Many properties for minimizers were proven by Buttazzo, Oudet and Stepanov [4], Buttazzo and Stepanov [9, 10], Paolini and Stepanov [35]. Further results were obtained by Santambrogio and Tilli [37], Tilli [41], Lemenant [25] and Slepčev [38]. Main properties on minimizers include:

- minimizers are topological trees composed of finitely many Lipschitz curves, which meet at triple junctions,
- minimizers are not C^1 regular in general, and can have "corners". This holds even if the reference measure μ has smooth density. However $C^{1,1}$ regularity near triple junctions holds; moreover, for any minimizer Σ and corner $v \in \Sigma$, the set $\{z : |z-v| = d(z,\Sigma)\}$ (i.e. the set of points z for which v is closest among all points of Σ) has positive μ-measure.

Research results. My research results are mainly in the following areas.

Regularity of minimizers. First order regularity of minimizers has been extensively studied. It is known that minimizers are not C^1 regular in general, but little was known about the set of non differentiability. I have proven that such set is not closed in general ([29]). Thus minimizers can be highly irregular.

Second order regularity was poorly understood. In collaboration with Slepčev, I have proven a sharp quantitative estimate on the curvature (understood as measures, since minimizers are not C^1 regular in general): any minimizer is composed of at most $\lfloor \mu(\mathbb{R}^n)/\lambda \rfloor$ branches, and the sum of the total variation of the curvature measure of such branches does not exceed $\mu(\mathbb{R}^n)/\lambda$ ([33]). This is the *first* "global" second order result.

Moreover, in modeling problems, topological complexity of minimizers is relevant, and discrete approximation of continuous reference measures are often used (e.g. in quantization). However little was known about the relation between the topological complexity of minimizers at discrete level, and minimizers at continuum level. In the same work [33] we have shown that with such kind of approximation, the topological complexity is lower semi-continuous, i.e. the topological complexity of minimizers at continuum level does not exceed the limit inferior of topological complexity of minimizers at discrete level. This is highly desirable in modeling.

Approximating via parameterized curves. The average-distance problem is related to the classic notion of "principal curves", widely used in data analysis. In such setting the minimization is restricted to the family of embedded curves. However well-posedness was unclear since it was not known whether minimizers are injective. In collaboration with Slepčev, I proved that minimizers are injective in two-dimensional domains, provided that either the exponent $p \geq 2$ (from (1)), or $p \in [1,2)$ but the reference measure μ is absolutely continuous with respect to Lebesgue measure with essentially bounded density.

Improving the data representation. In statistics, projecting a distribution μ of points on a representation gives a significant simplification, since it allows to work with one-dimensional (instead higher dimensional) objects. However projecting distinct data points on a singleton is undesirable, since it leads to a loss of information. Thus additional terms penalizing high data concentration on the representation are required. A first possibility is to add a L^2 penalization on the density ([30]) of the form

$$\begin{cases} \int_0^L \left|\frac{d\nu}{d\mathcal{L}^1}\right|^2 ds & \text{if } \nu \ll \mathcal{L}^1, \\ +\infty & \text{otherwise.} \end{cases}$$

The measure on $\gamma([0,L])$ is thus introduced as variable of the minimization. To remove the excessive geometric rigidity given by $\int_{\mathbb{R}^n} d(x,\gamma([0,L]))^p d\mu$, which forced points to project on one of the closest points on the representation, we replaced it by

$$\int_{\mathbb{R}^n \times \gamma([0,L])} |x-y|^p d\pi(x,y),$$

with π varying among transport plans between μ and ν. This is a transport cost, widely studied in the context of optimal transportation (cf. [3, 46] and references therein). Thus it is allowed

for points to project on any point on $\gamma([0,L])$. Projecting on points further away increases the transport cost (and it results that points will be projected only slightly further), but can be beneficial if it decreases the L^2 penalization term. An additional term $\eta(\gamma)$ penalizing non-injectivity was introduced, and I proved ([30]) that if (γ,ν) minimizes

$$\int_{\mathbb{R}^n\times\gamma([0,L])}|x-y|^p d\pi(x,y)+\lambda L+\varepsilon\int_0^L\left|\frac{d\nu}{d\mathcal{L}^1}\right|^2 ds+\varepsilon'\eta(\gamma),\qquad \varepsilon,\varepsilon'>0,$$

then ν has Lipschitz density.

Approximating with curvature penalization. Other possible penalization terms include curvature dependent energies. These arise in many mechanical and biological models, such as energy density of bending elastic membranes, closed lipid bilayers and cell membranes (cf. [34, 44, 45, 22]). Geometric variants include Willmore type energies, extensively used (e.g. [13, 15, 40]) due their regularizing effect, and closely related to splines. These are piecewise polynomial curves used for data interpolation, and cubic splines minimize the integrated squared curvature among polynomial interpolating curves (cf. [18, 2]). Spline interpolations, by allowing to use even low degree polynomials to achieve low approximation error, are preferable over polynomial interpolation.

In collaboration with Slepčev, in [31] we considered the minimization problem

$$\min_\gamma\int_{\mathbb{R}^n} d(x,\gamma([0,L]))^p d\mu+\lambda L+\int_0^L|\kappa_\gamma|^2 ds,$$

where γ varies among arc-length parameterized curves, and κ_γ denotes the curvature of γ. In this case we proved that minimizers are $C^{1,1}$ regular, and if μ is absolutely continuous with respect to Lebesgue density, with L^r integrable Radon-Nikodym derivative, then the mass projecting on the image $\gamma([t_0,t_1])$ $(t_0,t_1\in(0,L))$ does not exceed $c|t_1-t_0|^{1-1/r}$, for some constant $c>0$. This is desirable, since it provides an upper bound on the mass projecting on $\gamma([t_0,t_1])$.

Evolution problems. Many real world problems are intrinsically evolutionary and monotonically growing: for instance most transport networks and utility grids are built in phases. Thus evolutionary models are required for such problems. The average-distance problem admits an evolutionary variant: given an open, bounded, connected domain Ω a (nonnegative, finite) measure μ, absolutely continuous with respect to the Lebesgue measure on Ω, $p\geq 1$, a time step $\varepsilon>0$, an initial datum X_0, consider the recursive sequence

$$\begin{cases} w_0:=X_0,\\ w_{n+1}\in\ \operatorname{argmin}_{\Sigma\supseteq w_n}\int_\Omega d(x,\Sigma)^p d\mu(x)+\frac{1}{2\varepsilon}\mathcal{H}^1(\Sigma\triangle w_n),\end{cases}\tag{2}$$

where $\triangle$ denotes the symmetric difference. A solution of (2) is a sequence $\{w_n\}$. I proved ([28, 27]) that solutions of (2) exhibit several properties satisfied by minimizers of (1), namely absence of loops and Ahlfors regularity. However the topological complexity of $\{w_n\}$ can be greater, since there exist explicit examples where w_0 contains a single point triple junction (with the rest being simple points), but for some n, the set w_n contains a quadruple junction ([28]).

References

[1] Abdi H. and Williams L.J.: *Principal component analysis*, WIREs Comput. Stat., vol. 2(4), pp. 433–459, 2010

[2] Ahlberg J.H., Nilson E.N. and Walsh J.L.: *The Theory of Splines and Their Applications*, Mathematics in Science and Engineering, Academic Press, first edition, New York, 1967

[3] Ambrosio L., Gigli N. and Savaré G.: *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics Zürich, Birkhäuser Verlag, Basel, second edition, 2008

[4] Buttazzo G., Oudet E. and Stepanov E.: *Optimal transportation problems with free Dirichlet regions*, Progr. Nonlinear Differential Equations Appl., vol. 51, pp. 41–65, 2002

[5] Buttazzo G., Pratelli A., Solimini S. and Stepanov E.: *Optimal urban networks via mass transportation*, Springer Lecture Notes in Mathematics, 2009

[6] Buttazzo G., Pratelli A. and Stepanov E.: *Optimal pricing policies for public transportation networks*, SIAM J. Optimiz., vol. 16(3), pp. 826–853, 2006

[7] Buttazzo G. and Santambrogio F.: *A Mass Transportation Model for the Optimal Planning of an Urban Region*, SIAM Rev., vol. 51(3), pp. 593–610, 2009

[8] Buttazzo G. and Santambrogio F.: *A Model for the Optimal Planning of an Urban Area*, SIAM J. Math. Anal., vol. 37(2), pp.514–530, 2005

[9] Buttazzo G. and Stepanov E.: *Minimization problems for average distance functionals*, Calculus of Variations: Topics from the Mathematical Heritage of Ennio De Giorgi, (ed. D. Pallara), Quaderni di Matematica, Seconda Università di Napoli, Caserta, vol. 14, pp. 47–83, 2004

[10] Buttazzo G. and Stepanov E.: *Optimal transportation networks as free Dirichlet regions for the Monge-Kantorovich problem*, Ann. Sc. Norm. Sup. Pisa Cl. Sci., vol. 2, pp. 631–678, 2003

[11] Dal Maso G., Fonseca I. and Leoni G.: *Analytical validation of a continuum model for epitaxial growth with elasticity on vicinal surfaces*, Arch. Rational Mech. Anal., vol. 212, pp. 1037-1064, 2014

[12] Ding C. and He X.: *K-means clustering via principal component analysis*, Proc. of Int'l Conf. Machine Learning (ICML 2004), pp. 225–232, July 2004

[13] Dondl P.W., Mugnai L. and Röger M.: *A phase field model for the optimization of the Willmore energy in the class of connected surfaces*, Preprint
(available at `http://arxiv.org/pdf/1305.5054v1`)

[14] Drineas P., Frieze A., Kannan R., Vempala S. and Vinay V.: *Clustering large graphs via the singular value decomposition*, Mach. Learn., vol. 56, pp. 9–33, 2004

[15] Du Q., Liu C., Ryham R. and Wang X.: *A phase field formulation of the Willmore problem*, Nonlinearity, vol. 18(3), pp. 1249–1267, 2005

[16] Duchamp T. and Stuetzle W.: *Geometric properties of principal curves in the plane*, Robust Statistics, Data Analysis, and Computer Intensive Methods In Honor of Peter Huber's 60th Birthday, H. Rieder ed., vol. 109, pp. 135–152, Springer-Verlag, 1996

[17] Duport C., Politi P., and Villain J.: *Growth instabilities induced by elasticity in a vicinal surface*, J. Phys., vol. 1(5), pp. 1317–1350, 1995

[18] Fan J. and Yao Q.: *Nonlinear time series: nonparametric and parametric methods*, Springer series in statistics, Springer, New York, 2005

[19] Jolliffe I.T.: *Principal Component Analysis*, Springer Series in Statistics, second edition, Springer-Verlag, New York, 2002

[20] Hastie T.: *Principal curves and surfaces*, PhD Thesis, Stanford Univ., 1984

[21] Hastie T. and Stuetzle W.: *Principal curves*, J. Amer. Statist. Assoc., vol. 84, pp. 502–516, 1989

[22] Helfrich W.: *Elastic properties of lipid bilayers–Theory and possible experiments*, Z. Naturforsch., vol. 28, pp. 693–703, 1973

[23] Landau L.D. and Lifshitz E.M.: *Fluid mechanics*, Course of Theoretical Physics, vol. 6, Pergamon Press, London, 1959

[24] Le Roux B. and Rouanet H,: *Geometric data analysis: from correspondence analysis to structured data analysis*, Kluwer Academic Publishers, Dordrecht, 2004

[25] Lemenant A.: *About the regularity of average distance minimizers in* $\mathbb{R}^2$, Preprint
(available at `http://cvgmt.sns.it/paper/1615/`)

[26] Lyche T. and Schumaker L.L.: *On the convergence of cubic interpolating splines*, A. Meir (ed.), A. Sharma (ed.) , Spline Functions and Approximation Theory , Birkhäuser, pp. 169Ð189, 1973

[27] Lu X.Y.: *Geometric and topological properties of exponent related dynamic evolution*, Commun. Math. Anal., vol. 12(2), pp. 106–123, 2012

[28] Lu X.Y.: *Branching time estimates in quasi-static evolution for the average distance functional*, Commun. Appl. Anal., vol. 16(2), pp. 229–248, 2012

[29] Lu X.Y.: *Example of minimizer of the average distance problem with non closed set of corners*, Rend. Sem. Mat. Univ. Padova, Accepted paper

[30] Lu X.Y.: *Regularity of densities in relaxed and penalized average distance problem*, Netw. Heterog. Media, Accepted paper

[31] Lu X.Y. and Slepčev D.: *Average-distance problem with curvature penalization for data parameterization: regularity of minimizers*, Preprint

[32] Lu X.Y. and Slepčev D.: *Average-distance problem for parameterized curves*, ESAIM: Control Optim. Calc. Var., Accepted paper

[33] Lu X.Y. and Slepčev D.: *Properties of minimizers of the average distance functional via discrete approximation of measures*, SIAM J. Math. Anal., vol 45(5), pp. 3114–3131, 2013
[34] Ou Yang Z.C., Liu J.X. and Xie Y.Z.: *Geometric methods in the elastic theory of membranes in liquid crystal phases*, World Scientific, Singapore, 1999
[35] Paolini E. and Stepanov E.: *Qualitative properties of maximum and average distance minimizers in* $\mathbb{R}^n$, J. Math. Sci. (N.Y.), vol. 122(3), pp. 3290–3309, 2004
[36] Prenter P.M.: *Splines and variational methods*, Dover, Mincola, N.Y., 2008
[37] Santambrogio, F. and Tilli, P.: *Blow-up of Optimal Sets in the Irrigation Problem*, J. Geom. Anal., vol. 15(2), pp. 343-362, 2005
[38] Slepčev D.: *Counterexample to regularity in average-distance problem*, Ann. Inst. H. Poincaré (C), vol. 31(1), pp. 169–184, 2014
[39] Smola A.J., Mika S., Schölkopf B. and Williamson R.C.: *Regularized principal manifolds*, J. Mach. Learn., vol. 1, pp. 179–209, 2001
[40] Tibshirani R.: *Principal curves revisited*, Stat. Comput., vol. 2, pp.183–190, 1992
[41] Tilli, P.: *Some explicit examples of minimizers for the irrigation problem*, J. Convex Anal., vol. 17, pp. 583-595, 2010
[42] T'oth, F.: *A stability criterion to the moment theorem*, Studia Sci. Math. Hungar. vol. 34, pp. 209-224, 2001
[43] Triebel H.: *Spline bases and spline representations in function spaces*, Arch. Math., vol. 36(1), pp. 348-359, 1981
[44] Tu Z.C. and Ou Yang Z.C.: *A geometric theory on the elasticity of bio-membranes*, J. Phys. A: Math. Gen., vol. 37, pp. 11407–11429, 2004
[45] Tu Z.C. and Ou Yang Z.C.: *Elastic theory of low-dimensional continua and its applications in bio- and nano-structures*, J. Comput. Theor. Nanos., vol. 5(4), pp. 422–448, 2007
[46] Villani C.: *Topics in optimal transportation*, vol. 58, Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2003