A nonlinear principal component decomposition

Florian Gunsilius
Susanne Schennach

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP16/17

A nonlinear principal component decomposition[*]

Florian Gunsilius
Brown University

Susanne Schennach
Brown University

This version: March 28, 2017, First Version: June 25, 2016.
Preliminary and incomplete. Comments welcome.

Abstract

The idea of summarizing the information contained in a large number of variables by a small number of ”factors” or ”principal components” has been widely adopted in economics and statistics. This paper introduces a generalization of the widely used principal component analysis (PCA) to nonlinear settings, thus providing a new tool for dimension reduction and exploratory data analysis or representation. The distinguishing features of the method include (i) the ability to always deliver truly independent factors (as opposed to the merely uncorrelated factors of PCA); (ii) the reliance on the theory of optimal transport and Brenier maps to obtain a robust and efficient computational algorithm and (iii) the use of a new multivariate additive entropy decomposition to determine the principal nonlinear components that capture most of the information content of the data.

1 Introduction

The idea that the information contained in a large number of variables can be summarized by a small number of variables (the “factors” or “principal components”) has been widely adopted in economics and statistics. For example, asset returns are often modeled as a function of a small number of factors (e.g., Stock and Watson (1989), Ludvigson and Ng (2007), Bai and Ng (2002), Bai (2003), Bai and Ng (2012)). Cross-country variations are also found to have common components (e.g., Gregory and Head (1999)). Factor analysis is used for forecasting (Stock and Watson (1999)) and for Engel curves construction in demand analysis (Lewbel (1991)). More broadly, applications can be found in many fields of statistics (Loève (1978)) and include medical imaging (Sjöstrand, Stegmann, and Larsen (2006)), data compression (Wallace (1991)) and even search engines (Brin and Page (1998)).

Although Principal Component Analysis (PCA) has a long history as an effective device for dimension reduction (Jolliffe (1986)), it exhibits two main limitations. First, it is fundamentally a linear transformation of the data and is thus not the most appropriate representation to use if the different data dimensions exhibit some form of mutual nonlinear relationships. Second, the resulting principal components are merely uncorrelated, but not necessarily independent, thus suggesting

[*]This work is supported by the US National Science Foundation under grant SES-1659334.

that they do not capture truly unrelated effects, except, of course, in a simple linear Gaussian setting.

The importance of obtaining independent factors is perhaps best understood with a simple example: Consider two uncorrelated zero-mean variables X and Y that however exhibit statistical dependence because they are functionally related via $Y = X^2 - 1$ (with X satisfying $E\left[X^2\right] = 1$ and $E\left[X^3\right] = 0$). In a linear framework, two factors are needed (X and Y themselves) to fully describe the data, whereas one factor would be sufficient in a nonlinear framework, with a curvilinear coordinate system defined by:

$$(X, Y) = \left(F, F^2 - 1\right)$$

where F is the factor. This example is simple and low-dimensional — the savings in term of number of factors could be significantly greater in higher dimensions.

The aim of this paper is to introduce a practical nonlinear generalization of PCA that captures nonlinear forms of dependence and delivers truly independent factors. The output of the method is a low-dimensional curvilinear coordinate system that tracks the important features of the data. The key ingredients of our approach are (i) the reliance on the theory of Brenier maps (Brenier (1991)), which are a natural generalization of monotone functions in multivariate settings, (ii) the use of entropy (Kullback (1959), Csiszar (1991), Golan, Judge, and Miller (1996), Shore and Johnson (1980), Gray (2011), Shannon (1948)) to determine the principal nonlinear components that capture most of the information content of the data and (iii) the introduction of a novel multivariate additive decomposition of the entropy into one-dimensional contributions. The resulting method is computationally attractive, as it reduces to the well-studied problem of computing a Brenier map followed by a suitable matrix diagonalization step. These features distinguish our approach from the numerous other solutions that have been previously proposed in the very active literature seeking nonlinear generalizations of PCA (see, e.g., Lawrence (2012) and Lee and Verleysen (2007) for reviews). An appealing theoretical feature of our approach is the virtual absence of technical regularity conditions for the results to hold — all that is needed is that the data admits a density with respect to the Lebesgue measure.

This paper is organized as follows. In Section 2, we first informally outline and motivate our method before turning to more formal treatment of the approach and a description of its implementation. We then compare our approach with previously proposed nonlinear extensions of PCA in Section 3. We finally provide examples of applications, in Section 4, to both simulated and actual data. In particular, we focus on an application to the determination of a nonlinear version of the well-known Fama-French factors (Fama and French (1992), Fama and French (1993)).

2 Method

2.1 Outline

The proposed method relies on a powerful result of convex analysis, which characterizes the solution to the following optimization problem. Consider a random vector Y taking values in $\mathbb{R}^d$ with density $f(y)$ (with respect to the Lebesgue measure) and one wishes to find a (measurable) mapping $T : \mathbb{R}^d \mapsto \mathbb{R}^d$ such that the random variable $x = T(y)$ has a pre-specified density $\tilde{\Phi}(x)$.

As there are obviously an infinite number of possible T that satisfy this constraint, it is natural to select the simplest transformation in the sense that it minimizes:

$$\int \|y - T(y)\|^2 f(y)\, dy,$$

where $\|\cdot\|$ denotes the Euclidian norm. This minimization problem is known as the Monge-Kantorovich-Brenier optimal transportation problem, as it identifies the mapping that requires the least amount of probability mass movement in the mean square sense. (For introductions to this topic, we refer to Galichon (2016), Rachev and Rüschendorf (1998), Santambrogio (2015), Villani (2003), and Villani (2009).) The solution to this problem has desirable regularity properties, in particular, the so-called Brenier map T must take the form of the gradient of a convex function, which is often regarded as a natural generalization of the concept of monotonicity in multivariate settings (Brenier (1991), McCann (1995), Carlier, Chernozhukov, and Galichon (2016)). Remarkably, one can even show that $T(y)$ is the only transformation (subject to almost everywhere qualifications) mapping f to $\tilde{\Phi}$ that is the gradient of a convex function. This characterization of the Brenier map actually even relaxes any requirement of y having a finite variance. Numerous numerical methods to find $T(y)$ are available in the literature (e.g., Villani (2003), Villani (2009), Benamou and Brenier (2000), Chartrand, Wohlberg, Vixie, and Bollt (2009), Benamou, Froese, and Oberman (2014)).

We show that this optimal transportation problem is directly related to the determination of nonlinear independent components that best represent the data. By selecting a target density $\tilde{\Phi}(x)$ that factors as a product of univariate densities $\prod_{i=1}^{d} \tilde{\phi}_i(x_i)$, we obtain, by construction, independent components. These components define a curvilinear coordinate system in the space of the original variables via the inverse mapping $y = T^{-1}(x)$. Note that the factorization in terms of univariate marginals does not need to be along one specific Cartesian coordinate system. In general, one can have:

$$\tilde{\Phi}(x) = \prod_{i=1}^{d} \tilde{\phi}_i\left(u^i \cdot x\right)$$

where $\{u^i\}_{i=1}^{d}$ is a set of orthogonal unit vectors and $\tilde{\phi}_i(\cdot)$ are functions of one variable.

Obviously, there are many possible choices of u^i and $\tilde{\phi}_i(\cdot)$ and we need to be more specific to construct a well-defined procedure. First, we observe that, for a given choice of $\{u^i\}_{i=1}^{d}$, the choice of the $\tilde{\phi}_i(\cdot)$ is arbitrary, because different choices generate essentially equivalent curvilinear coordinate systems that only differ in the "speed" at which one travels along each axis. We exploit this arbitrariness by selecting the $\tilde{\phi}_i(x)$ to be of a particularly convenient form: A standard univariate normal, denoted $\phi(x)$. This choice is driven by the fact that a multivariate standard normal $\Phi(x) \equiv \prod_{i=1}^{d} \phi(x_i)$ is the only distribution which exhibits two properties: (i) it factors as a product of marginal and (ii) it is invariant under arbitrary rotations of the coordinate system. We can exploit the invariance under rotation to straightforwardly explore various possible choices of coordinate systems $\{u^i\}_{i=1}^{d}$ in search of an optimal one, in a sense to be made precise below.

Ultimately, our goal is to only keep the subset $\{u^i\}_{i=1}^{k}$ (with $k < d$) of the d dimensions that "explains" the most important features of the data. We show that, although the concept of variance is not very useful in nonlinear settings to identify the most important components, the concept of entropy proves extremely useful. Entropy is defined as

$$H = -\int f(y) \ln f(y) \, dy \tag{1}$$

for a given density $f(y)$ with respect to the Lebesgue measure and where the integral is over $\mathbb{R}^d$ and, by convention, $0 \ln 0 \equiv \lim_{t \to 0} t \ln t = 0$. The concept of entropy has a long history as a measure of the amount of information contained in a probability distribution (Kullback (1959), Csiszar (1991), Golan, Judge, and Miller (1996), Shore and Johnson (1980)). We seek the $\{u^i\}_{i=1}^k$ that accounts for the largest possible fraction of this entropy. We demonstrate that the k most important components $u_1, \ldots, u_k$ can be simply identified from the (normalized) eigenvectors associated with the k largest eigenvalues of the matrix $\bar{J} \equiv -\int f(y) \ln J(y) \, dy$ where $J(y) = \frac{\partial T(y)}{\partial y'}$ is the Jacobian of the transformation T (the previously obtained Brenier mapping f onto Φ) and the $\ln$ of a matrix M, diagonalizable as $M = P \operatorname{diag}(\lambda_1, \ldots, \lambda_d) P^{-1}$ is defined in the usual way (Gantmacher (1959)) as $\ln M \equiv P \operatorname{diag}(\ln \lambda_1, \ldots, \ln \lambda_d) P^{-1}$.

Our low-dimensional nonlinear representation of the data, denoted y^{k*} then takes the form:

$$y^{k*} = T^{-1} \left(\sum_{j=1}^{k} u^j x_j \right). \tag{2}$$

where $x_j \in \mathbb{R}$ for $j = 1, \ldots, k$ are arbitrary coordinates expressed in our curvilinear coordinate system. This expression clearly reduces to standard PCA if T^{-1} is a linear map, which would be the case if the data is Gaussian. In this work, we take the number of factor k as given and leave the question of its data-driven determination for future work.

2.2 Main results

The first step in the construction is to obtain the Brenier map $T : \mathbb{R}^d \to \mathbb{R}^d$ mapping a given density f (with respect to the Lebesgue measure) to the standard Normal Φ of the same dimension. Once the Brenier map T has been determined, the principal components (or factors) can be determined by rotating the coordinate system of the standard normal variables. The implied curvilinear coordinate system in the space of the original vector y provides the nonlinear factors. The principal components are determined by keeping the coordinates that contribute the most to the entropy of the distribution of y.

Our only regularity condition, which we assume throughout, is:

Assumption 1 *The random vector y admits a density $f(y)$ with respect to the Lebesgue measure.*

In order to be able to select which nonlinear factor contributes the most to the overall entropy of the observed distribution, we need to introduce a formal definition of the entropy contribution of each factor. The following lemma shows that the entropy of the distribution of y, denoted H, can be naturally expressed as a sum of factor-specific contributions.

Lemma 1 *Let T be the Brenier map transporting f onto Φ. Then, for any set of unit vectors $\{u^j\}_{j=1}^d$ forming an orthogonal basis, the entropy of $f(y)$ can we written as*

$$H = \sum_{j=1}^{d} H_{u^j}$$

where, for a given unit vector u, H_u is the the ***effective contribution of factor*** u ***to the entropy****, given by*

$$H_u = -\frac{1}{2}\ln(2\pi e) + u'\bar{J}u.$$

Here, $-\frac{1}{2}\ln(2\pi e)$ is the entropy of a univariate standard normal while

$$\bar{J} \equiv -\int \Phi(x)\ln\left(J\left(T^{-1}(x)\right)\right)dx = -\int f(y)\ln J(y)\,dy$$

where $J(y) = \frac{\partial T(y)}{\partial y'}$. (The $\ln$ of a matrix M, diagonalizable as $M = P\operatorname{diag}(\lambda_1,\ldots,\lambda_d)P^{-1}$ is defined as $\ln M \equiv P\operatorname{diag}(\ln\lambda_1,\ldots,\ln\lambda_d)P^{-1}$.)

An automatic consequence of this Lemma is that the most important factors (based on our entropy criterion) can be determined as follows.

Theorem 2 *For a given $k \leq d$, a solution to*

$$\left(u^1,\ldots,u^k\right) = \underset{\left(u^1,\ldots,u^k\right)\in\mathcal{U}_{k,d}}{\operatorname{argmax}} \sum_{j=1}^{k} H_{u^j}$$

where $\mathcal{U}_{k,d} = \left\{u_i \in \mathbb{R}^d : u_i \cdot u_j = \mathbf{1}\{i=j\} \text{ for } i,j \in \{1,\ldots,k\}\right\}$, is given by the k eigenvectors associated with the k largest eigenvalues of the matrix $\bar{J}$ (defined in Lemma 1).

Remark There may be multiple solutions, corresponding to trivial changes in the signs of u^i or permutations among them. Also, as in standard PCA, eigenvectors are not unique if some eigenvalues are degenerate.

The definition of the effective contribution of a factor to the entropy in Lemma 1 exhibits a number of desirable properties. First, in well-known special cases, it reduces to the standard properties relied upon in linear PCA.

Corollary 3 *(Special cases) (i) For independent random variables, the decomposition of Lemma 1 reduces to the usual fact that the entropy of independent random variables is additive. (ii) For normally distributed variables, picking the $k < d$ factors with largest variance is equivalent to picking the k factors with largest entropy.*

However, the advantage of our concept of additive entropy decomposition is that maintains the same natural form and interpretation in general nonlinear and non-Gaussian models. In contrast, the concept of variance does not generalize well to nonlinear factor setting (as it is not clear what is the meaning of comparing the variances of random variables that are nonlinearly related). The problem can best be seen by the following example. Consider two bivariate distributions, which could represent the projection of the same data along two directions. One is uniformly distributed on a "s"-shaped set and one is uniformly distributed on an "l"-shaped set. The longest linear dimension of the "s" could be shorter that the "l" and yet, the length of the "s" along the its curve could be longer than the "l". Variance would identify the distribution with support "l" as explaining more variation in the data, whereas, in fact, it is arguably the distribution with support "s" that does. Uniform distributions with a larger support have a larger entropy and thus, in our example, the distribution with "s"-shaped support would be correctly identified as more informative.

A second desirable property is the fact that the principal factors (that contribute the most to the entropy) can be easily determined by diagonalizing the matrix $\bar{J}$, which is no more involved than for linear PCA. The computation of the Brenier map is an additional preliminary step relative to the linear case, but it only needs to be performed once for one arbitrary choice of coordinate system. The optimization of the "orientation" of the principal factors can be done via simple linear algebra operations, despite the nonlinear nature of the original problem.

If k components are kept, then our low-dimensional nonlinear representation of the data, denoted y^{k*}, takes the form:

$$y^{k*} = T^{-1}\left(\sum_{j=1}^{k} u^j x_j\right)$$

where u^j for $j = 1, \ldots, k$ are the normalized eigenvectors of the $\bar{J}$ matrix associated with the k largest eigenvalues and $x_j \in \mathbb{R}$ for $j = 1, \ldots, k$ are an arbitrary coordinates expressed in our curvilinear coordinate system. This expression clearly reduces to standard PCA if T^{-1} is a linear map, which would be the case if the data is Gaussian.

2.3 Implementation

Our implementation is based on Chartrand, Wohlberg, Vixie, and Bollt (2009) and ideas from Benamou, Froese, and Oberman (2014) and uses the known fact that the Brenier map is the only mapping (i) that will transform the one given density into another given density and (ii) that can be written as the gradient of a convex function. The constraint that the original density $f(y)$ be mapped to $\Phi(x)$ by the map $x = T(y)$ can be expressed using the usual change of variables formula:

$$f(y) = \Phi(T(y)) \det\left(\frac{\partial T(y)}{\partial y'}\right).$$

We also know, more specifically, that the Brenier map can be written as a gradient $T(y) = \partial c(y) / \partial y$ of some convex function $c(y)$. This implies that the problem reduces to finding the convex function $c(y)$ solving the equation:

$$f(y) - \Phi\left(\frac{\partial c(y)}{\partial y}\right) \det\left(\frac{\partial^2 c(y)}{\partial y \partial y'}\right) = 0. \tag{3}$$

Chartrand, Wohlberg, Vixie, and Bollt (2009) further showed that $c(y)$ minimizes a functional whose gradient (or, more formally, whose functional derivative with respect to the function $c(y)$) is the left-hand side of Equation (3). Hence, one can simply update a trial $c(y)$ in the direction of this gradient to iteratively converge to the solution:

$$c_{n+1}(y) = c_n(y) + \tau\left(f(y) - \Phi\left(\frac{\partial c_n(y)}{\partial y}\right) \det\left(\frac{\partial^2 c_n(y)}{\partial y \partial y'}\right)\right)$$

where $c_n(y)$ represents a converging sequence of approximations to the solution and τ denotes a user-specified step size parameter.

A useful heuristic rule that improves the convergence of the iterative solution method is to multiply τ by factor $\theta_d \approx 0.2$ whenever the current value of τ would have lead $\left\| f(y) - \Phi\left(\frac{\partial c_n(y)}{\partial y}\right) \det\left(\frac{\partial^2 c_n(y)}{\partial y \partial y'}\right)\right\|$ to increase from one iteration to the next. This prevents the method from making steps that overshoot the solution. We also found, empirically, that convergence is sped up if τ is increased by a factor $\theta_u \approx 1.1$ whenever $\left\| f(y) - \Phi\left(\frac{\partial c_n(y)}{\partial y}\right) \det\left(\frac{\partial^2 c_n(y)}{\partial y \partial y'}\right)\right\|$ has been decreasing for $n_u \approx 20$ iterations.

Another useful fail-safe strategy is to enforce convexity of $c_n(y)$ at each step. This can be accomplished by checking if one of the eigenvalues of the Hessian of $c_n(y)$ is negative at some point y and, if so, by reducing $c_n(y)$ at y so that this eigenvalue becomes equal to a small user-specified positive number ε_c. This is iterated until all points of nonconvexity have been eliminated. The value ε_c is gradually reduced as iterations over n progress, so that, asymptotically, the positive curvature constraint is not binding at the solution.

The density $f(y)$ is first obtained by kernel smoothing and we implement Equation (3) via finite differences and by sampling the functions on a grid. We place a regular, fixed, grid on the original data, with grid points y_m, indexed by $m \in \{-M, \ldots, +M\}^d$. The corresponding curvilinear grid in the transformed space is $x_m = T(y_m)$ where the d elements $T_j(y_m)$ of $T(y_m)$ are approximated via centered finite differences[1] as

$$T_j(y_m) \approx \frac{c\left(y_{m+\Delta_j}\right) - c\left(y_{m-\Delta_j}\right)}{\left\| y_{m+\Delta_j} - y_{m-\Delta_j}\right\|}$$

where Δ_j is a d-dimensional vector containing 1 at the j-th element and zero elsewhere. The Jacobian is also approximated via centered finite differences:

$$\frac{\partial^2 c(y)}{\partial y_i \partial y_j} \approx \frac{c\left(y_{m+\Delta_j+\Delta_i}\right) - c\left(y_{m+\Delta_j-\Delta_i}\right) - c\left(y_{m-\Delta_j+\Delta_i}\right) + c\left(y_{m-\Delta_j-\Delta_i}\right)}{\left\| y_{m+\Delta_i} - y_{m-\Delta_i}\right\| \left\| y_{m+\Delta_j} - y_{m-\Delta_j}\right\|}.$$

Once $c(y)$ has been determined, the optimal rotation can be found as follows. The matrix $\bar{J}$ from Lemma 1 can approximated by

$$\bar{J} \approx -\sum_{m \in \{-M,\ldots,+M\}^d} \Phi(x_m) \ln(J(y_m)) \prod_{j=1}^{d} \left\| x_{m+\Delta_j} - x_{m-\Delta_j}\right\| /2.$$

Diagonalization of this matrix yields the (normalized) eigenvectors $u^1, \ldots, u^k$ associated with the k largest eigenvalues. The curvilinear coordinate system representing the k most important nonlinear factors is then given by Equation (2).

Our implementation is general in that it can handle data of any dimensions, although computational requirements do increase steeply with the dimension due to the grid-type representation of the factors. For very high-dimensional problems, it may not be practical to perform a full nonlinear PCA analysis. In such a case, one can exploit the fact that, by a Taylor expansion argument, the effect of the less important factors can often be linearized. This implies that a very effective approach is to initially perform a linear PCA step to first identify the very small components that can be linearized and only perform a nonlinear PCA on the remaining components that are large enough to even have nonlinear features.

[1] At boundary points, noncentered differences need to be used instead. We use noncentered differences that are second-order accurate (so that their accuracy is theoretically equivalent to the centered differences used for non boundary points). This remark applies to all finite differences throughout the paper.

3 Discussion

While the idea of extending PCA to nonlinear settings has apparently not been explored in the field of econometrics, this problem has received more attention in the field of machine learning. It is thus instructive to identify key distinguishing features of the proposed method that clearly differ from general features shared by many other existing methods.

Our approach guarantees, by construction, that the resulting factors are statistically independent, thus implying that they each truly represent distinct and unrelated features of the data. Many existing methods (e.g., Schölkopf, Smola, and Müller (1998), Gorban and Zinovyev (2010), Gashler, Ventura, and Martinez (2008), Tenenbaum, de Silva, and Langford (2000)) specifically target the goal of accurately representing the data by a manifold of a given dimension and thus perform very well in this respect. However, the goal of obtaining independent factors is largely overlooked. Even methods designed with independence in mind (e.g., Bell and Sejnowski (1995)), only achieve it approximately in general. The importance of independence can also be appreciated from a data compression perspective: Any remaining dependence in the factors implies that one could, in principle, obtain a more compact representation of the data by exploiting the statistical dependence to partially predict some of the factor from the values of others and thus reduce the amount of information that needs to be stored (the prediction error could have a smaller variance than the factors themselves, for instance). This is not possible under full independence of the factors, thus indicating that the data has already been optimally "compressed".

Our approach relies on the concept of entropy (e.g., Kullback (1959), Shore and Johnson (1980), Schennach (2005)) to gauge the importance of the factors, whereas most existing methods employ some concept of "distance" to identify the important factors. Unfortunately, the concept of distance becomes somewhat ambiguous in the context of curvilinear coordinate systems (e.g., is distance measured in, say, the Euclidian metric in terms of the data coordinates y or in the curvilinear coordinates x?). In contrast, the idea of entropy is directly tied to the information content of the data and can be defined independently of a choice of metric,[2] a key realization that has, so far, only been used in a few methods (e.g., Bell and Sejnowski (1995), although they use entropy in a very different way).

Our procedure has a well-defined unique global optimal solution thanks to a direct connection to the theory of optimal transport and Brenier maps (Brenier (1991), McCann (1995)). Some existing methods enjoy global optimization properties (e.g., Tenenbaum, de Silva, and Langford (2000)) but most do not. Many methods rely on an iterative refinement of a manifold based on some local rules that penalize complexity and reward accuracy. While these rules convey useful properties to the decomposition, their complexity and locality make it hard to ascertain convergence to a global optimum. Many methods (e.g., Demartines and Hérault (1997), Bell and Sejnowski (1995), Kramer (1991)) rely on neural networks for optimization, and convergence properties are typically verified by experimentation rather than by formal proof.

Our procedure reduces, without user input, to linear PCA in the classic linear Gaussian case. This apparently simple property is not guaranteed in most sophisticated nonlinear dimension reduction techniques, even those (e.g., Schölkopf, Smola, and Müller (1998)) that have a very direct connection to linear PCA. Yet, this property ensures that (i) the procedure is at least as good as linear PCA and that (ii) it can be freely combined with linear methods to reach the best compromise

[2] However, it does depend on the choice of reference probability measure, here taken to be the Lebesgue measure.

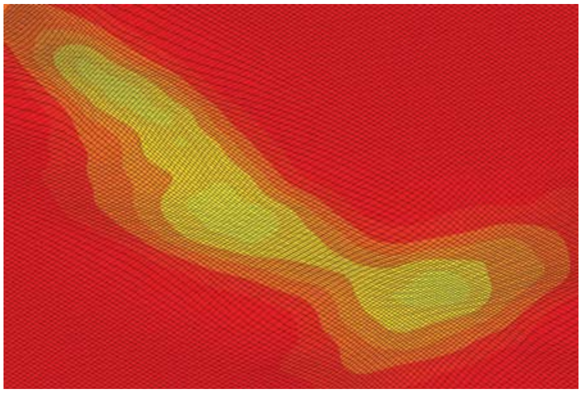

Figure 1: Nonlinear principal component analysis in a simple two-dimensional example using a mixture of 3 normals as an input density. The density is shown as a color map while the overlaid curvilinear grid represents the nonlinear factors.

between computational and statistical efficiency.

A large fraction of existing methods (e.g., Roweis and Saul (2000), Tenenbaum, de Silva, and Langford (2000)) only work directly with data points, rather than with a density of the input data. Our approach can work with both, which is extremely useful if the input data can be accurately modeled (in part or entirely) by a parametric model. Perhaps even more importantly, the ability to work with densities represents a major theoretical advantage to study the asymptotic properties of the method in the limit of large data sets.

4 Illustrative Examples

Our first example employs simulated data to clearly illustrate the method's ability to capture both the general "direction" and the nonlinear nature of the main features of the data. As an input density, we use a mixture of three normals:

$$N\left(\begin{bmatrix} 3 \\ -3 \end{bmatrix}, \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}\right),\ N\left(\begin{bmatrix} -3 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 & -3 \\ -3 & 4 \end{bmatrix}\right),\ N\left(\begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 4 & -2 \\ -2 & 2 \end{bmatrix}\right)$$

with equal weights. The resulting nonlinear principal component analysis, depicted in Figure 1, shows that the method correctly identifies the direction along which the data exhibits the most variation. The curvilinear coordinate system also roughly follows the clear ridge in the data despite its multimodal nature. Additionally, the grid lines are further apart in areas where the density is spread over a bigger area, indicating that they do "adapt" to the target distribution in a nontrivial nonlinear fashion.

We have also empirically verified that our approach recovers the usual PCA result in the special case of normally distributed data. The resulting coordinate system is indeed linear and the rotation matrix that yields the linear factors matches the known principal axes of the normal data within the expected numerical accuracy (typically two decimal places).

As an empirical illustration, we revisit the well-known Fama-French factors (Fama and French (1992), Fama and French (1993)) with a nonlinear perspective. We take their 3-factor data (French (2017)) as an input and check if they could be better reparametrized by a curvilinear coordinate

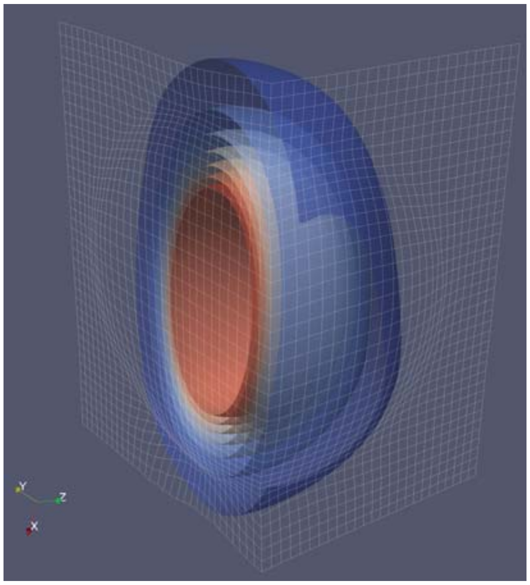

Figure 2: Nonlinear principal component analysis of the 3-factor Fama-French data. The color surfaces show contours of constant probability density, with a portion cut out to better show the geometry of the inner contours. The overlaid grid represents the curvilinear coordinate system identified with our method (for clarity, only the subset of the mesh lying along the boundary of the cutout is shown). The linear axes labelled x, y and z represent the original coordinate system.

system. The existence of nonlinear (rather than linear) factors is a necessary condition for the proposed method to deliver a more efficient representation of the data. As shown in Figure 2, we find that nonlinear factors are indeed necessary to obtain independent latent factors. We also find that our entropy-driven method to select the "directions" of the curvilinear coordinate system is effective at identifying the dominant orientation of the main features of the data.

Of course, the ultimate test of the usefulness of the method in this context would be to ascertain that predictions of asset returns made with the nonlinear factors are more accurate than with the same number of linear factors. This could be shown, for instance, by reducing the 5-factor Fama-French model to 3 nonlinear factor and comparing the performance the latter with the conventional linear 3-factor Fama-French model. This analysis is under way and will be included in future versions of this paper.

A Proofs

Proof of Lemma 1. The density of the observed data, $f(y)$, can be expressed in terms of the Brenier map $T(y)$ and the standard multivariate normal $\Phi(x)$:

$$f(y) = \Phi(T(y)) \det\left(\frac{\partial T(y)}{\partial y'}\right),$$

where the Jacobian matrix $\partial T(y)/\partial y'$ is almost everywhere well-defined as Brenier maps are differentiable almost everywhere, by a theorem from Aleksandrov (Aleksandrov (1939); see also Villani (2003)).

We can then find a simple expression for the entropy $H = -\int f(y) \ln f(y)\, dy$, via the change of variable: $x = T(y)$ (so that $dx = \det\left(\frac{\partial T(y)}{\partial y'}\right) dy$):

$$\begin{aligned}
H &= -\int \Phi(T(y)) \det\left(\frac{\partial T(y)}{\partial y'}\right) \ln\left(\Phi(T(y)) \det\left(\frac{\partial T(y)}{\partial y'}\right)\right) dy \\
&= -\int \Phi(x) \ln\left(\Phi(x)\left[\det\left(\frac{\partial T(y)}{\partial y'}\right)\right]_{y=T^{-1}(x)}\right) dx \\
&= -\int \Phi(x) \ln\left(\Phi(x) \det J\left(T^{-1}(x)\right)\right) dx
\end{aligned}$$

where $J(y) = \frac{\partial T(y)}{\partial y'}$.

$$\begin{aligned}
H &= -\int \Phi(x) \ln\left(\Phi(x) \det J\left(T^{-1}(x)\right)\right) dx \\
&= -\int \left(\prod_{i=1}^{d} \phi(x_i)\right) \ln\left(\left(\prod_{i=1}^{d} \phi(x_i)\right) \det J\left(T^{-1}(x)\right)\right) dx \\
&= A + B
\end{aligned}$$

where

$$\begin{aligned}
A &= -\int \left(\prod_{i=1}^{d} \phi(x_i)\right) \ln\left(\left(\prod_{i=1}^{d} \phi(x_i)\right)\right) dx \\
B &= -\int \Phi(x) \ln\left(\det J\left(T^{-1}(x)\right)\right) dx.
\end{aligned}$$

Each term can then be simplified:

$$\begin{aligned}
A &= -\sum_{j=1}^{d} \int \left(\prod_{i=1}^{d} \phi(x_i)\right) \ln(\phi(x_j))\, dx \\
&= -\sum_{j=1}^{d} \int \phi(x_j) \ln(\phi(x_j))\, dx_j \prod_{i \neq j} \left(\int \phi(x_i)\, dx_i\right) \\
&= -\sum_{j=1}^{d} \int \phi(x_j) \ln(\phi(x_j))\, dx_j = \sum_{j=1}^{d} -H_0
\end{aligned}$$

where $H_0 = -\frac{1}{2} \ln(2\pi e)$ is the entropy of a univariate normal.

To evaluate B, we use the equality:

$$\ln \det J\left(T^{-1}(x)\right) = \ln \prod_{i=1}^{d} \lambda_i(x)$$

where λ_i are the eigenvalues of $J\left(T^{-1}(x)\right)$. Note that since T is the gradient of a continuously differentiable convex function, $J(y) = J\left(T^{-1}(x)\right)$ is symmetric and therefore diagonalizable.

Also, a Brenier map between two Lebesgue densities is almost everywhere strictly convex (by Theorem 2.12 in Villani (2003)), which implies that $\lambda_i(x) > 0$ for $i = 1, \ldots, d$ almost everywhere. Next, we observe that

$$\ln\det J\left(T^{-1}(x)\right) = \sum_{j=1}^{d} \ln \lambda_j(x) = \operatorname{tr} \ln J\left(T^{-1}(x)\right) = \sum_{j=1}^{d} u^{j\prime}\left(\ln J\left(T^{-1}(x)\right)\right) u^j$$

where we introduced the logarithm of a matrix, have exploited the fact that the sum of eigenvalues is equal to the trace and that the trace of a matrix can be evaluated in any orthogonal coordinate system $\{u^j\}_{j=1}^{d}$. Then,

$$\begin{aligned} B &= -\int \Phi(x) \sum_{j=1}^{d} u^{j\prime}\left(\ln J\left(T^{-1}(x)\right)\right) u^j dx \\ &= \sum_{j=1}^{d} -u^{j\prime}\left(\int \Phi(x)\left(\ln J\left(T^{-1}(x)\right)\right) dx\right) u^j \\ &= \sum_{j=1}^{d} u^{j\prime} \bar{J} u^j \end{aligned}$$

where $\bar{J} = \int \Phi(x)\left(\ln J\left(T^{-1}(x)\right)\right) dx$, as defined in the statement of the Lemma. (Note that we also have $\bar{J} = -\int f(y) \ln J(y)\, dy$, by the simple change of variable $y = T^{-1}(x)$.) Collecting these results, we then have:

$$H = A + B = \sum_{j=1}^{d} -H_0 + \sum_{j=1}^{d} u^{j\prime}\bar{J}u^j = \sum_{j=1}^{d} H_{u^j}$$

for H_{u^j} defined in the statement of the theorem. ■

Proof of Theorem 2. Since matrix $\bar{J}$ is symmetric, it is diagonalizable with orthogonal eigenvectors. We can thus decompose it as $\bar{J} = P\Lambda P'$ where Λ is diagonal and its elements are ordered in decreasing order of magnitude and P is normalized so that $P'P = I$ (this also states that, in case of degenerate eigenvalues, we select an orthogonal set of eigenvectors among the infinite number of possibilities). We then have (observing that the additive constants $-\frac{1}{2}\ln(2\pi e)$ do not affect the optimization problem):

$$\begin{aligned} \left(u^1, \ldots, u^k\right) &= \underset{\left(u^1,\ldots,u^k\right)\in\mathcal{U}_{k,d}}{\operatorname{argmax}} \sum_{j=1}^{k} H_{u^j} = \underset{\left(u^1,\ldots,u^k\right)\in\mathcal{U}_{k,d}}{\operatorname{argmax}} \sum_{j=1}^{k} u^{j\prime}\bar{J}u^j \\ &= \underset{\left(u^1,\ldots,u^k\right)\in\mathcal{U}_{k,d}}{\operatorname{argmax}} \sum_{j=1}^{k} u^{j\prime}P\Lambda P' u^j = P \underset{\left(v^1,\ldots,v^k\right)\in\mathcal{U}_{k,d}}{\operatorname{argmax}} \sum_{j=1}^{k} v^{j\prime}\Lambda v^j \\ &= P\left[e^1, e^2, \ldots, e^k\right] = \left(P_{\cdot 1}, P_{\cdot 2}, \ldots P_{\cdot k}\right) \end{aligned}$$

where e^i is a d-dimensional column vector with 1 as its i entry and 0 elsewhere and $P_{\cdot i}$ is the i-th column of P, i.e., the i-th eigenvector. ■

Proof of Corollary 3. The special case (i) of independent factors corresponds to the case where $T(y)$ takes the element-by-element form $T_i(y) = g_i(y_i)$ for some strictly increasing function $g_i(y_i)$. Note that this mapping is a Brenier map because it is the gradient of the convex function $c(y) \equiv \sum_{i=1}^{d} G_i(y_i)$, where $G_i(y_i) = \int_{y^*}^{y_i} g_i(u)\, du$ for some $y^* \in \mathbb{R}$. Indeed, since $g_i(y_i)$ is strictly increasing, $G_i(y_i)$ is strictly convex, i.e. $G_i(\alpha y_i^1 + (1-\alpha) y_i^2) < \alpha G_i(y_i^1) + (1-\alpha) G_i(y_i^2)$ for any $y^1 \equiv (y_1^1, \ldots, y_d^1) \in \mathbb{R}^d$ and $y^2 \equiv (y_1^2, \ldots, y_d^2) \in \mathbb{R}^d$, which implies that

$$\begin{aligned} c\left(\alpha y^1 + (1-\alpha) y^2\right) &= \sum_{i=1}^{d} G_i\left(\alpha y_i^1 + (1-\alpha) y_i^2\right) < \sum_{i=1}^{d} \left(\alpha G_i\left(y_i^1\right) + (1-\alpha) G_i\left(y_i^2\right)\right) \\ &= \alpha \sum_{i=1}^{d} G_i\left(y_i^1\right) + (1-\alpha) \sum_{i=1}^{d} G_i\left(y_i^2\right) = \alpha c\left(y^1\right) + (1-\alpha) c\left(y^2\right), \end{aligned}$$

i.e., $c(y)$ is convex.

We also observe that $J(T^{-1}(x))$ is diagonal since $\partial T_i(y)/\partial y_j = 0$ for $j \neq i$. We then have, for an orthogonal basis u^j that is aligned with the independent factors, that $u^{j\prime} (\ln J(T^{-1}(x))) u^j = [\ln J(T^{-1}(x))]_{jj} = \ln J_{jj}(T^{-1}(x)) = \ln \left[\frac{\partial g_j(y_j)}{\partial y_j}\right]_{y_j = g_j^{-1}(x_j)}$ and

$$\begin{aligned} H_{u^j} &= -\frac{1}{2}\ln(2\pi e) - \int \left(\prod_{i=1}^{d} \phi(x_i)\right) \left(\ln \left[\frac{\partial g_j(y_j)}{\partial y_j}\right]_{y_j = g_j^{-1}(x_j)}\right) dx \\ &= -\frac{1}{2}\ln(2\pi e) - \int \phi(x_j) \ln \left[\frac{\partial g_j(y_j)}{\partial y_j}\right]_{y_j = g_j^{-1}(x_j)} dx_j \prod_{i \neq j} \int \phi(x_i)\, dx_i \\ &= -\frac{1}{2}\ln(2\pi e) - \int \phi(x_j) \ln \left[\frac{\partial g_j(y_j)}{\partial y_j}\right]_{y_j = g_j^{-1}(x_j)} dx_j \end{aligned}$$

Using the fact that $-\frac{1}{2}\ln(2\pi e) = -\int \phi(x_i) \ln \phi(x_i)\, dx_i$ and performing the change of variables $x_j = g_j(y_j)$, we have:

$$\begin{aligned} H_{u^j} &= -\int \phi(x_j) \ln \left(\phi(x_j) \left[\frac{\partial g_j(y_j)}{\partial y_j}\right]_{y_j = g_j^{-1}(x_j)}\right) dx_j \\ &= -\int \phi(x_j) \left[\frac{\partial g_j(y_j)}{\partial y_j}\right]_{y_j = g_j^{-1}(x_j)} \ln \left(\phi(x_j) \left[\frac{\partial g_j(y_j)}{\partial y_j}\right]_{y_j = g_j^{-1}(x_j)}\right) \left(\left[\frac{\partial g_j(y_j)}{\partial y_j}\right]_{y_j = g_j^{-1}(x_j)}\right)^{-1} dx_j \\ &= -\int \phi(g_j(y_j)) \frac{\partial g_j(y_j)}{\partial y_j} \ln \left(\phi(g_j(y_j)) \frac{\partial g_j(y_j)}{\partial y_j}\right) dy_j \\ &= -\int f_j(y_j) \ln f_j(y_j)\, dy_j \end{aligned}$$

where f_j is the marginal density of y_j with respect to Lebesgue measure. Thus, our definition generalizes this simple additive result to the case where the y_j are not independent (they are not, in general).

To show statement (ii), we observe that the entropy of a multivariate normal where each independent factor has variance σ_i^2 is given by $\sum_{j=1}^{d} H_{u^j}$ with

$$H_{u^j} = -\frac{1}{2}\ln(2\pi e) + \frac{1}{2}\ln\sigma_i^2$$

Hence, picking the $k < d$ factors with largest variance is equivalent to picking the k factors with largest entropy. ■

References

ALEKSANDROV, A. D. (1939): "Almost everywhere existence of the second differential of a convex function and some properties of convex functions," *Leningrad Univ. Ann.*, 37, 3–35.

BAI, J. (2003): "Inferential Theory for factor models of large dimensions," *Econometrica*, 71, 135–171.

BAI, J., AND S. NG (2002): "Determining the number of factors is approximate factor models," *Econometrica*, 70, 191–221.

BAI, J., AND S. NG (2012): "Determining the number of primitive shocks in factor models," *Journal of Business & Economic Statistics*, 25, 52–60.

BELL, A. J., AND T. J. SEJNOWSKI (1995): "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, 7, 1129–1159.

BENAMOU, J.-D., AND Y. BRENIER (2000): "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem," *Numerische Mathematik*, 84, 375–393.

BENAMOU, J.-D., B. D. FROESE, AND A. M. OBERMAN (2014): "Numerical solution of the Optimal Transportation problem using the Monge-Ampère equation," *Journal of Computational Physics archive*, 260, 107–126.

BRENIER, Y. (1991): "Polar factorization and monotone rearrangement of vector-valued functions," *Communications on pure and applied mathematics*, 44, 375–417.

BRIN, S., AND L. PAGE (1998): "The Anatomy of a Large-Scale Hypertextual Web Search Engine," in *Seventh International World-Wide Web Conference (WWW 1998)*.

CARLIER, G., V. CHERNOZHUKOV, AND A. GALICHON (2016): "Vector quantile regression: an optimal transport approach," *Annals of Statistics*, 44, 1165–1192.

CHARTRAND, R., B. WOHLBERG, K. R. VIXIE, AND E. M. BOLLT (2009): "A Gradient Descent Solution to the Monge-Kantorovich Problem," *Applied Mathematical Sciences*, 3, 1071–1080.

CSISZAR, I. (1991): "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems," *Annals of Statistics*, 19, 2032–2066.

Demartines, P., and J. Hérault (1997): "Curvilinear Component Analysis: A Self-Organizing Neural Network for Nonlinear Mapping of Data Sets," *IEEE Transactions on Neural Networks*, 8, 148–154.

Fama, E. F., and K. R. French (1992): "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47, 427–465.

Fama, E. F., and K. R. French (1993): "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33, 3–56.

French, K. R. (2017): "Data Library," http://mba.tuck.dartmouth.edu/ pages/ faculty/ ken.french/ data_library.html.

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

Gantmacher, F. R. (1959): *The Theory of Matrices*, vol. 1. Chelsea, New York.

Gashler, M., D. Ventura, and T. Martinez (2008): "Iterative Non-linear Dimensionality Reduction with Manifold Sculpting," in *Advances in Neural Information Processing Systems*, ed. by J. C. Platt, D. Koller, Y. Singer, and S. Roweis, vol. 20, pp. 513–520. NIPS.

Golan, A., G. Judge, and D. Miller (1996): *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. John Wiley and Sons, New York.

Gorban, A. N., and A. Zinovyev (2010): "Principal manifolds and graphs in practice: from molecular biology to dynamical systems," *International Journal of Neural Systems*, 20, 219–232.

Gray, R. M. (2011): *Entropy and Information Theory*. Springer.

Gregory, A., and A. Head (1999): "Common and Country-Specific Fluctuations in Productivity Investment, and the Current Account," *Journal of Monetary Economics*, 44, 423–452.

Jolliffe, I. T. (1986): *Principal component analysis*. Spinger-Verlag, New York.

Kramer, M. A. (1991): "Nonlinear principal component analysis using autoassociative neural networks," *AICHE Journal*, 37, 233–243.

Kullback, S. (1959): *Information Theory and Statistics*. Wiley, Newyork.

Lawrence, N. D. (2012): "A unifying probabilistic perspective for spectral dimensionality reduction: insights and new models," *Journal of Machine Learning Research*, 13, 1609–1638.

Lee, J. A., and M. Verleysen (2007): *Nonlinear Dimensionality Reduction*. Springer.

Lewbel, A. (1991): "The Rank of Demand Systems: Theory and Nonparametric Estimation," *Econometrica*, 59, 711–730.

Loève, M. (1978): *Probability Theory II*. New York: Springer.

LUDVIGSON, S. C., AND S. NG (2007): "The empirical riskreturn relation: A factor analysis approach," *Journal of Financial Economics*, 83, 171–222.

MCCANN, R. J. (1995): "Existence and uniqueness of monotone measure-preserving maps," *Duke Mathematical Journal*, 80, 309–324.

RACHEV, S., AND L. RÜSCHENDORF (1998): *Mass Transportation Problems: Volume I: Theory*. Springer, New York.

ROWEIS, S. T., AND L. K. SAUL (2000): "Nonlinear dimensionality reduction by locally linear embedding," *Science*, 290, 2323–2326.

SANTAMBROGIO, F. (2015): *Optimal Transport for Applied Mathematicians*. Springer, New York.

SCHENNACH, S. M. (2005): "Bayesian Exponentially Tilted Empirical Likelihood," *Biometrika*, 92, 31–46.

SCHÖLKOPF, B., A. SMOLA, AND K.-R. MÜLLER (1998): "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, 10, 1299–1319.

SHANNON, C. E. (1948): "A Mathematical Theory of Communication," *Bell Sys. Tech. J.*, 27, 379–423.

SHORE, J., AND R. JOHNSON (1980): "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," *IEEE Transactions on Information Theory*, 26, 26–37.

SJÖSTRAND, K., M. B. STEGMANN, AND R. LARSEN (2006): "Sparse Principal Component Analysis in Medical Shape Modeling," *International Symposium on Medical Imaging*, 6144.

STOCK, J. H., AND M. WATSON (1989): "New Indexes of Coincident and Leading Economic Indications," in *NBER Macroeconomics Annual 1989*, ed. by O. J. Blanchard, and S. Fischer. M.I.T. Press, Cambridge.

STOCK, J. H., AND M. WATSON (1999): "Forecasting Inflation," *Journal of Monetary Economics*, pp. 293–335.

TENENBAUM, J. B., V. DE SILVA, AND J. C. LANGFORD (2000): "A global geometric framework for nonlinear dimensionality reduction," *Science*, 290, 2319–2323.

VILLANI, C. (2003): *Topics in Optimal Transportation*. American Mathematical Society, Providence.

VILLANI, C. (2009): "Optimal transport: Old and New," in *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Heidelberg.

WALLACE, G. K. (1991): "The JPEG Still Picture Compression Standard," *Communication of the ACM*, 34, 30–44.