cemmap
centre for microdata methods and practice

Independent nonlinear component analysis

Florian Gunsilius
Susanne Schennach

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP46/19

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Independent Nonlinear Component Analysis*

Florian Gunsilius
MIT

Susanne Schennach
Brown University

This version: September 18, 2019, First Version: June 25, 2016.

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 broadly 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; (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 and (iv) formally nesting PCA as a special case, for linear Gaussian factor models. We illustrate the method's effectiveness in an application to the prediction of excess bond returns from a large number of macro factors.

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., Ludvigson and Ng (2009), 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)). Dimension reduction methods are also related to machine learning, which has been receiving increasing attention in economics (see Athey and Imbens (2019) for a recent review aimed at economists).

*This work is supported by the US National Science Foundation under grant SES-1659334. Earlier versions of this work were circulated under the title "A Nonlinear Principal Component Analysis". We thank the participants of the "Celebrating Whitney Newey's Contributions to Econometrics" conference for helpful comments. Computational resources were provided by the Center for Computation and Visualization at Brown University.

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

The importance of obtaining nonlinear 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 $E[X^2] = 1$ and $E[X^3] = 0$). 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:

$$(X, Y) = \left(F_1, F_1^2 - 1\right)$$

where F_1 is the only factor needed and the second factor $F_2 = 0$ is simply constant. Hence, replacing two dependent variables X, Y by two independent ones (F_1 and F_2) reveals that only one factor is actually needed. This example is simple and low-dimensional — the savings in terms of number of factors can 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), Schennach (2014)) 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. Computationally, the resulting method combines the well-studied problem of computing a Brenier map with a simple matrix diagonalization step. It yields independent (rather than merely uncorrelated) factors and, in the special case of Gaussian data, it reduces to conventional linear PCA. 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). In particular, the fact that our method is closely related to linear PCA enables a convenient hybrid approach that naturally combines linear and nonlinear PCA to yield a method that can efficiently handle high dimensional data.

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 prediction of excess bond returns from macroeconomic factors, in the spirit of Ludvigson and Ng (2009). This application illustrates that our approach can detect nonlinearities in the factors that are of economic relevance while avoiding a complex model selection procedure.

All proofs are collected in an Appendix.

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) where 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), Ekeland, Galichon, and Henry (2011)). 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., 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 marginals 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.[1] 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. The entropy of a density $f(y)$ is defined as

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

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), Schennach (2005)). 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. When the data is Gaussian, T^{-1} is a linear map and Equation (2) reduces to standard PCA.

2.2 Main results

2.2.1 Optimal transport-based entropy decomposition

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.

We assume the following regularity condition throughout:

[1] While the selection of the number of factors is discussed in Section 4, the formal derivation of a data-driven selector of the number of factors in fully nonlinear settings is left for future research.

Assumption 1 *The random vector y admits a uniformly bounded 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, as shown in the Appendix.

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 be written as*

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

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

$$H_u = \frac{1}{2}\ln(2\pi e) + u'\bar{J}u. \tag{3}$$

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

$$\bar{J} \equiv -\int \Phi(x) \ln\left(J\left(T^{-1}(x)\right)\right) dx = -\int f(y) \ln J(y)\, dy \tag{4}$$

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

Intuitively, the two terms in the expression for H_u (Equation (3)) arise from (i) the entropy of a standard univariate normal (since the standard normal is used as a target density) and (ii) a correction term that quantifies the total deviations from a standard normal along the direction u. An automatic consequence of this Lemma is that the most important factors (based on our entropy criterion) can be determined as follows.

Theorem 1 *For a given $k \leq d$, a solution to*

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

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 Equation (4)).

Remark As in standard PCA, there may be multiple solutions, corresponding to trivial changes in the signs of u^i, permutations or linear combinations among them. Also, eigenvectors are not unique if some eigenvalues are degenerate. These ambiguities (except for signs changes and the possibility of degenerate eigenvalues) can be circumvented by adopting the convention of iteratively defining the u_j for $j = 1, 2, \ldots, k$ as:

$$u_j = \underset{u^j\in\mathcal{U}^j}{\mathrm{argmax}}\, H_{u^j}$$

where $\mathcal{U}^j = \{u^j : u^i \cdot u^j = \mathbf{1}\{i = j\} \text{ for } i = 1, \ldots, j\}$.

The definition of the effective contribution of a factor to the entropy in Lemma 1 exhibits a number of desirable properties. First, it generalizes well-known special cases, as shown in the Appendix:

Corollary 1 *(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, the Brenier map $T(y)$ is linear and 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 it 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 than 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 analogous to linear PCA. 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. 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.

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

$$y^{k*} = T^{-1}\left(\sum_{j=1}^{k} u^j x_j\right) \tag{5}$$

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.

2.2.2 Relationship to linear PCA

Equation (5) clearly reduces to standard PCA if T^{-1} is a linear map, which is the case when the data is Gaussian. The fact that our approach nests PCA as a special case is not only conceptually appealing but also opens the way to a very efficient hybrid method. The idea is to exploit the fact that, by a Taylor expansion argument, the effect of the less important factors can often be linearized. This suggests that one could constrain T to be linear along the dimensions in which the data has the smallest spread while allowing for nonlinearity along the other dimensions. This

hybrid scheme can be efficiently implemented by first performing a linear PCA step to 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. This suggestion raises the question of whether PCA might misidentify some factors as being unimportant when, in fact, a nonlinear analysis would have revealed that they have a large entropy. The following inequality formally guards against this possibility.

Theorem 2 *Let the random vector y be partitioned as $\left(y^{k^*}, y^o\right)$, then*

$$H\left[f_{\left(y^{k^*},y^o\right)}\right] - H\left[f_{y^{k^*}}\right] \leq \operatorname{tr}\ln\left(\sqrt{2\pi e \operatorname{Var}\left[y^o\right]}\right). \tag{6}$$

This result can be used to convert output from PCA into statements regarding entropy: If, after a PCA step, one decides to keep the k^* factors with the largest variance while omitting the remaining ones, we let y^{k^*} denote the factors kept while letting y^o denote those omitted. The left-hand side of (6) then represents the entropy neglected by eliminating y^o, while the right-hand side bounds this entropy in terms of the PCA variance of the omitted factors, with a small variance implying a small entropy loss. More generally, this result shows that the total variance of a group of factors obtained via PCA provides an upper bound on the total entropy of those factors.

2.2.3 Asymptotics

We now establish consistency of our procedure, taking as given known results regarding the consistency and regularity of the building blocks entering our estimator.

Theorem 3 *Let $\hat{T}(y)$ and $\hat{f}(y)$ be uniformly (over y) consistent estimators of $T(y)$ and $f(y)$, respectively. If $\int \left|\hat{f}(y)\right| dy \leq \bar{f} < \infty$ and $J(y) \equiv \partial T(y)/\partial y'$ is uniformly continuous and its eigenvalues are bounded away from zero and infinity, then, there exist sequences ε_n and $\bar{y}_n$ such that*

$$\widehat{\bar{J}} \equiv -\int_{\|y\|\leq \bar{y}_n} \hat{f}(y) \ln \hat{J}(y)\, dy \xrightarrow{p} \bar{J}$$

where the elements of $\hat{J}(y)$ are given, for $i, j = 1, \ldots, d$, by

$$\hat{J}_{ij}(y) = \frac{1}{(4\varepsilon_n)}\left(\hat{T}_i(y + \varepsilon_n e_j) - \hat{T}_i(y - \varepsilon_n e_j) + \hat{T}_j(y + \varepsilon_n e_i) - \hat{T}_j(y - \varepsilon_n e_i)\right),$$

where e_j denotes the j-th unit vector. Furthermore, the eigenvalues and eigenvectors of $\widehat{\bar{J}}$ converge to those of $\bar{J}$, if the eigenvalues of $\bar{J}$ are distinct.

This establishes consistency of all the quantities defining our curvilinear coordinate system of Equation (5). The needed assumptions are stated in high-level form because they can be directly verified using existing results: Standard results on the uniform consistency of $\hat{f}(y)$ can be found in Andrews (1995), while the uniform consistency of the estimated Brenier map $\hat{T}(y)$ were obtained in Chernozhukov, Galichon, Hallin, and Henry (2017). Assumptions regarding the regularity of the Jacobian $J(y)$ follow from Caffarelli's Regularity theory (see, e.g. Villani (2003) Theorem 4.14 and de Philippis and Figalli (2014)).

Given the current state of development of the theory of estimated optimal transport maps, it is beyond the scope of this paper to provide complete distributional results. We however note that ongoing related works seek to establish the validity of subsampling for obtaining the asymptotic distribution of $\hat{T}(y)$ (Gunsilius and Schennach (2019)) and the convergence rates of Brenier maps and related quantities (Gunsilius (2019), Hütter and Rigollet (2019)).

2.3 Implementation

Our implementation is based on ideas from Villani (2003) and Chartrand, Wohlberg, Vixie, and Bollt (2009) and uses the known fact that the Brenier map $T(y)$ is the only mapping (i) that will transform one given density $f(y)$ into another given density $\Phi(x)$ and (ii) that can be written as the gradient of a convex function $c(y)$, called the "potential" of the Brenier map. The function $c(y)$ is the minimizer of the functional:

$$M[c] = \int c(y) f(y) dy + \int \left(\max_{y} (x \cdot y - c(y)) \right) \Phi(x) dx. \tag{7}$$

This functional admits a functional derivative $\delta M[c]/\delta c$ satisfying the condition

$$M(c + \delta c) = \int \frac{\delta M[c](y)}{\delta c} \delta c(y) dy + o(\|\delta c\|),$$

Remarkably, this functional derivative admits a simple closed-form expression:

$$\frac{\delta M[c](y)}{\delta c} = f(y) - \Phi\left(\frac{\partial c(y)}{\partial y} \right) \det\left(\frac{\partial^2 c(y)}{\partial y \partial y'} \right). \tag{8}$$

At the optimum of this unconstrained optimization problem, this derivative must be zero, which implies that $f(y) = \Phi(T(y)) \det(\partial T(y)/\partial y')$, i.e., that the original density $f(y)$ is mapped to $\Phi(x)$ by the map $x = T(y)$ using the usual change of variables formula.

The determination of the Brenier map thus reduces to finding a convex function c such that $\delta M[c]/\delta c = 0$. To facilitate the search for the solution, we include a penalty term that disfavors nonconvex functions. This curvature penalty is asymptotically not active when the solution is approached, but it helps steer away from nonconvex functions during the numerical convergence.

We propose a numerical solution methods based on approximating $c(y)$ by a discrete mesh and using finite differences to approximate gradients and Hessians of $c(y)$. The density $f(y)$ is first obtained by kernel smoothing and we implement Equation (8) via finite differences and by sampling the functions on a grid. We place a regular, fixed, grid on the original data, with grid points $\dot{y}_m$, indexed by $m \in \{-M, \ldots, +M\}^d$. The corresponding curvilinear grid in the transformed space is $\dot{x}_m = T(\dot{y}_m)$ where the d elements $T_j(\dot{y}_m)$ of $T(\dot{y}_m)$ are approximated via centered finite differences[2] as

$$T_j(\dot{y}_m) \approx \frac{c(\dot{y}_{m+\Delta_j}) - c(\dot{y}_{m-\Delta_j})}{\|\dot{y}_{m+\Delta_j} - \dot{y}_{m-\Delta_j}\|}$$

[2]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.

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:

$$J_{ij}\left(\dot{y}_m\right) = \left[\frac{\partial^2 c\left(y\right)}{\partial y_i \partial y_j}\right]_{y=\dot{y}_m} \approx \frac{c\left(\dot{y}_{m+\Delta_j+\Delta_i}\right) - c\left(\dot{y}_{m+\Delta_j-\Delta_i}\right) - c\left(\dot{y}_{m-\Delta_j+\Delta_i}\right) + c\left(\dot{y}_{m-\Delta_j-\Delta_i}\right)}{\left\|\dot{y}_{m+\Delta_i} - \dot{y}_{m-\Delta_i}\right\| \left\|\dot{y}_{m+\Delta_j} - \dot{y}_{m-\Delta_j}\right\|}.$$

During the optimization of $c\left(y\right)$, it is helpful to enforce convexity of $c_n\left(y\right)$ at each step. This can be accomplished by checking if one of the eigenvalues of the Hessian of $c\left(y\right)$ is negative at some grid point $\dot{y}_m$ and, if so, by reducing $c\left(y\right)$ at $\dot{y}_m$ so that this eigenvalue becomes equal to a small user-specified positive number ε. This is iterated until all points of nonconvexity have been eliminated. The value ε is gradually reduced as iterations progress, so that, asymptotically, the curvature penalty is not binding at the solution.

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

$$\bar{J} \approx -\sum_{m\in\{-M,\ldots,+M\}^d} f\left(\dot{y}_m\right) \ln\left(J\left(\dot{y}_m\right)\right) \prod_{j=1}^{d} \left\|\dot{y}_{m+\Delta_j} - \dot{y}_{m-\Delta_j}\right\| /2.$$

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).

An alternative numerical approach that may improve the scalability of the method to higher dimensions would be to use a series approximation to $c\left(y\right)$ instead of a grid representation. In this approach, the optimization problem associated with finding $c\left(y\right)$ can also be made more efficient by exploiting the closed-form expression (8) for the gradient of $M\left[c\right]$. The use of a curvature penalty term is also helpful in this context to prevent convergence to a nonconvex $c\left(y\right)$. Replacing summation over grid points by Monte Carlo sampling is another independnt way to improve tractability of the method in higher dimensions. The use of series approximation in the context of Brenier maps has been proposed before (Lee (2018)), but our suggestion to use (i) a curvature penalty, (ii) an explicit expression (8) for the gradient and (iii) integral evaluations through Monte Carlo sampling goes beyond that earlier contribution and opens the way to handle higher dimensional situations.[3]

Once the Brenier map $T\left(x\right)$ has been determined (from either methods above), obtaining the corresponding curvilinear coordinate system (5) involves computing its inverse $T^{-1}\left(x\right)$. We provide an efficient approach for its practical computation. In particular, we use the equivalence $T^{-1}\left(x\right) = \partial c^*\left(x\right)/\partial x$, where $c^*\left(x\right) \equiv \max_y\left(x \cdot y - c\left(y\right)\right)$ is the Legendre-Fenchel transform of c. This representation is convenient, as the Legendre-Fenchel transform can be computed in linear time using the algorithm proposed in Lucet (1997), which we adapt to a higher dimensional setting.

Our implementation is general in that it can handle data of any dimensions, although computational requirements do increase with the dimension. For very high-dimensional problems, it may not be practical to perform a full nonlinear PCA analysis due to computational requirements and

[3]Of course, in high dimensions one might also encounter a curse of dimensionality. This could be mitigated by using either parametric (yet nonlinear) functional forms for $T\left(y\right)$ or semiparametric modeling (in which only some of the factors are treated fully nonparametrically). This may be easier to accomplish in the sieve approach.

the potential for a curse of dimensionality. Our proposed hybrid linear-nonlinear PCA approach thus proves helpful in that context.

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 considerable attention in the field of machine learning. Instead of comparing our approach with a long list of existing methods, it is more 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,[4] 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), albeit after specifying a "neighborhood size") 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 to a unique solution are typically assessed by experimentation

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

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 that have a very direct connection to linear PCA (e.g., Schölkopf, Smola, and Müller (1998), which simply performs PCA on a set of nonlinear functions of the data). 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 improve the method's computational 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

4.1 Simulations

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.

4.2 Application to bond excess returns prediction

Linear factor models have found important applications in forecasting (e.g. Stock and Watson (2002)). Here, we revisit an influential example of this line of work (Ludvigson and Ng (2009)) by relaxing the constraint of linearity of the factor model. This exercise not only corroborates Ludvigson and Ng's findings under more general conditions, but also provides an avenue to simplify the implementation of the prediction process by avoiding an extensive model selection step.

Ludvigson and Ng (2009) seek to assess whether macroeconomic variables could improve the predictability of excess returns on bonds, beyond known financial factors, such as the well-known Cochrane and Piazzesi (2005) factor (hereafter, CP). Answering such questions tests the core of the basic expectations hypothesis, namely that deviations from expected future prices should be unpredictable conditional on current information. Ludvigson & Ng's approach is to first use linear PCA to extract the most important factors out of a large set of macro indicators proposed by Stock

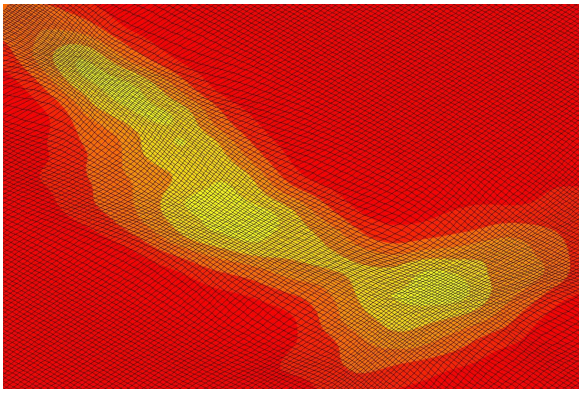

Figure 1: Nonlinear principal component analysis in a simple two-dimensional example using a mixture of 3 normals as an input density. The density is shown as a color map while the overlaid curvilinear grid represents the nonlinear factors.

and Watson (2002). In a second step, they use these factors as regressors in polynomial predictive regressions of future excess returns and see if they add predictive power relative to the CP variable alone. They rely on a model selection procedure to determine the form of the polynomial used in the predictions.

We propose to replace their use of PCA with our nonlinear PCA in the first step, which offers two advantages. First, we obtain truly independent factors by construction, which helps motivate that these factors truly represent different aspects of the economy. Second, and perhaps more importantly, in this case, our approach is found to eliminate the need for an extensive model selection procedure: In this application, the relevant nonlinearity is already accounted for by the factor model and does not need to be added in the predictive regression through a model selection step.

The input data consists of 132 macro series (provided in the supplementary information of Jurado, Ludvigson, and Ng (2015)) from which we extract monthly data for the time period extending from 1964 through 2003 (the same as in Ludvigson and Ng (2009)).[5] To improve the tractability of the calculations, we make use of a hybrid linear and nonlinear PCA. Following Ludvigson and Ng (2009), we use the formal PCA-based number of factors selection procedure of Bai and Ng (2002) to extract 8 relevant factors. Thanks to Theorem 2, we know that PCA cannot understate the importance of factors even in the presence of nonlinearity. Hence, the 124 factors eliminated in this step would also have been eliminated if we had carried out a nonlinear PCA-based procedure on all the variables instead. The end result of this step is an 8-dimensional dataset (over 480 time periods) obtained by projecting the original data onto the 8 PCA eigenvectors associated with the largest eigenvalues (sorted in decreasing order of eigenvalue).

Next, we apply our nonlinear PCA method to the resulting dataset on dimension 1 through k^* while treating factors k^*+1 through 8 linearly. To check that the value of k^* used is such that nonlinearity is negligible in factors k^*+1 through 8, we performed the analysis for various values of k^* (here, $2, 3$ and 4) and obtained essentially the same results. We also observed that the vectors u^3 and u^4 obtained via our nonlinear procedure (for $k^*=4$) match the corresponding eigenvectors of linear PCA (within 2 decimal places), thus further supporting the fact that linear PCA would be

[5]This data in Jurado, Ludvigson, and Ng (2015) is a slightly updated version of the data originally used in Ludvigson and Ng (2009). The differences are minor and do not affect our conclusions.

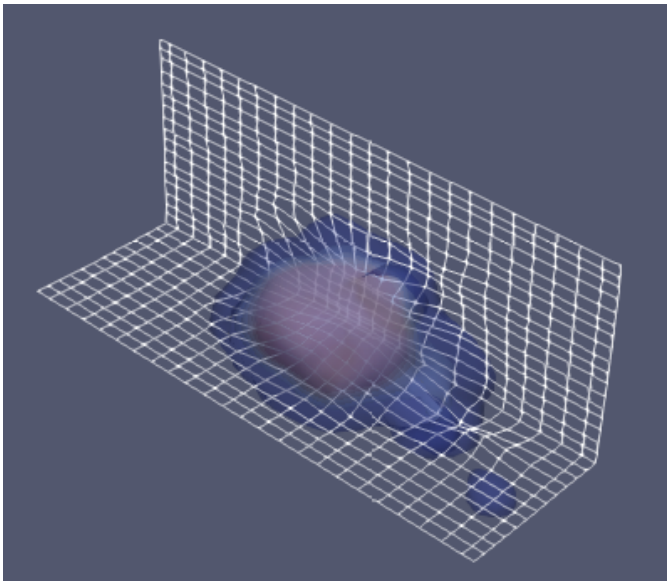

Figure 2: Graphical representation of the nonlinear factors. Colored surfaces show contours of constant estimated density of the data generating process while the grids represent the 3 most important nonlinear factors as a curvilinear coordinate system (L-shaped cross-section shown, for clarity). Both objects are three dimentional cross sections of four dimensional quantities, for plotting purposes. The underlying cartesian coordinate system used is defined by the four most important linear PCA components (with the 4th component not shown here).

adequate for these factors. We conservatively report here the analysis performed with $k^* = 4$, even though accounting for the nonlinearity in the first two factors would have been sufficient for this application. We use a grid-based implementation of the method with grid size of $31 \times 31 \times 31 \times 31$ and a kernel density estimator whose bandwidth is chosen by cross-validation.[6] The resulting 4 nonlinear factors are illustrated in Figure 2.

The process of checking if enough factors are nonlinearly included does not significantly add to the computational requirements, since the results of earlier optimizations (with few nonlinear factors) can be used as the starting point for further optimization with more nonlinear factors. This process should be familiar to practitioners, as it is entirely analogous to the selection of an adequate number of terms in nonparametric series or sieve estimation.

We use the factors resulting from the above analysis in a predictive regression of excess bond returns, $r_{t+1}^{(a)}$, which is defined as the difference between (i) the log holding period return from buying an a-year bond at time t and selling it as an $a-1$ year bond at time $t+1$ and (ii) the log yield on the one-year bond). Following Ludvigson and Ng (2009), the model takes the general form:

$$r_{t+1}^{(a)} = \alpha' F_t + \beta Z_t + \varepsilon_{t+1}$$

where ε_{t+1} is an error term, Z_t is the well-known CP financial factor and F_t denotes a vector of various macroeconomic factors. We use the following notation for these factors: $f_1, \ldots, f_8$ denote

[6] We use "leave-k consecutive observations out" cross-validation to allow for serial dependence. The bandwidth was found to not be sensitive to the specific value of k for $k \geq 18$ (i.e., beyond the standard 18-month lag traditionally used for time series exhibiting seasonality).

the most important factor found via PCA while $\tilde{f}_1, \ldots, \tilde{f}_4$ denote the nonlinear factors found via nonlinear PCA applied to $f_1, \ldots, f_4$. Our hybrid approach uses factors $\tilde{f}_1, \ldots, \tilde{f}_4, f_5, \ldots, f_8$.[7]

In our analysis, we consider models that either include or exclude Z_t and compare various choices of F_t (among $\tilde{f}_1, \ldots, \tilde{f}_4, f_1, \ldots, f_8$). To illustrate the advantages provided by our approach, we consider three broad classes of predictive models for $r_{t+1}^{(2)}$:

1. As a benchmark case (hereafter labelled "PCA+poly"), we use the model that Ludvigson and Ng (2009) select via the Bayesian Information Criterion (BIC), which consists of the first factor f_1 and its cube f_1^3, while the 7 other factors $f_2, \ldots, f_8$ enter linearly. We consider different possible numbers of factors, denoted κ, which means that we include factors $f_1, f_1^3, f_2, f_3, \ldots, f_\kappa$ in F_t. We also consider the Ludvigson and Ng's "preferred" model, found via an extensive model selection procedure,[8] which only includes a subset $f_1, f_1^3, f_3, f_4, f_8$ of the factors.

2. As another benchmark (labelled "PCA+lin"), we consider only including κ PCA factors linearly: $f_1, \ldots, f_\kappa$ (hence omitting f_1^3 from the previous case).

3. Our proposed approach (labelled "INPCA+lin") is to use κ nonlinear PCA factors in a linear predictive regression, in which case F_t contains $\tilde{f}_1, \ldots, \tilde{f}_\kappa$ (for $\kappa \leq 4$) or $\tilde{f}_1, \ldots, \tilde{f}_4, f_5, \ldots, f_\kappa$ (for $\kappa \geq 5$). This approach avoids the selection of which powers of each factor to use in the predictive regression.

We study the predictive power of each these approaches as a function of κ using an out-of-sample procedure. In addition to its robustness, this approach circumvents the need to perform a detailed analysis of the asymptotic properties of our method. We use the early half of the sample (years 1964–1983) to determine (i) the factors (either with PCA alone or with PCA followed by our nonlinear extension) and (ii) the regression coefficients. We then use the later half of the sample (years 1984–2003) to estimate the magnitude of the one-year ahead prediction errors, without re-optimizing either the regression coefficients or the choice of regressors.

First, our analysis broadly corroborates Ludvigson & Ng's findings, namely that macro factors have significant predictive power beyond the CP factor. Our results thus show that their analysis is robust to the use of more general nonlinear factor models. Our subsequent analysis thus focuses on the question of whether their analysis can be accomplished without an extensive model selection step.

Figure 3 compares the three approaches and reveals that our proposed approach (INPCA+lin) is as predictive as the best existing alternative (PCA+poly), both comparing across the same value of κ and overall, as described below. While there is clearly a benefit in going from a linear regression (in blue) to a polynomial regression (in orange) when using linear factors, we obtain essentially the

[7] This choice implicitly relies on the assumption that, if nonlinear PCA had been applied to factors $f_1, \ldots, f_8$, it would given the same result as linear PCA for factors $f_5, \ldots, f_8$. The fact that nonlinear PCA on the first 4 factors already gave the same result for factor f_3, f_4 as linear PCA is highly suggestive that this property holds for the less important factors beyond the fourth.

[8] They arrive at this model by first separately regressing returns linearly on each factor. The factors associated with better BIC scores are then further considered in univariate polynomial regressions. The regressions exhibiting the best BIC then point to the terms that are considered for inclusion in a range of possible multivariate regressions on the factors, which are then ranked using the BIC.

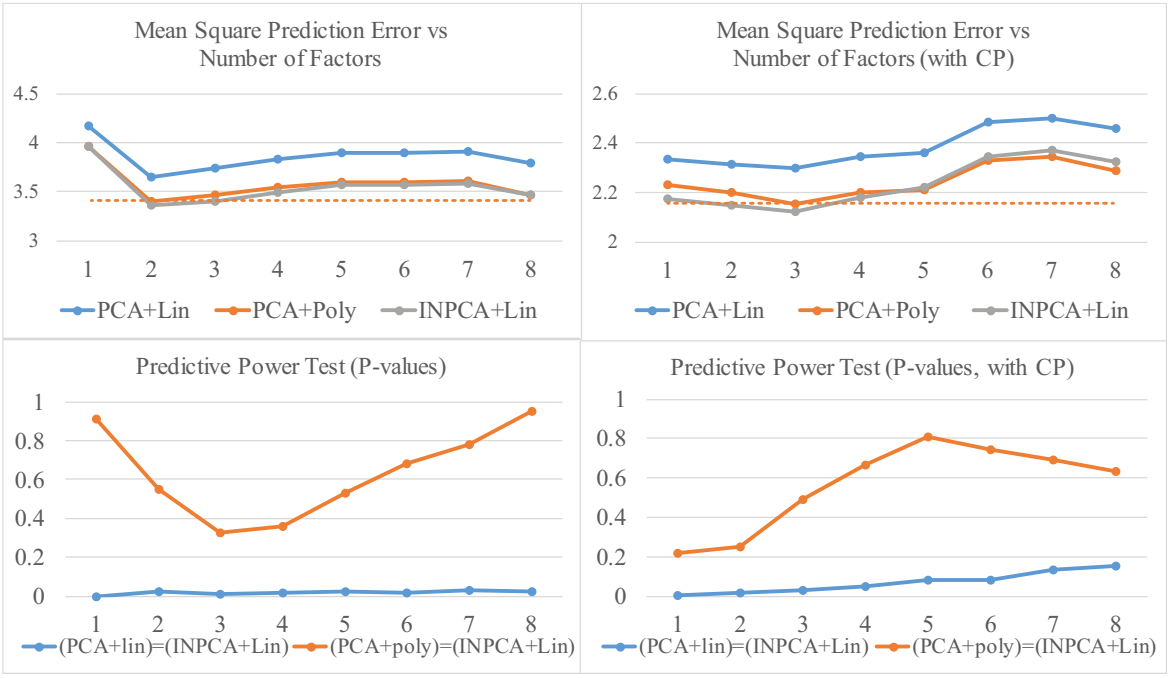

Figure 3: Top panels: Excess bond returns prediction errors for the different models considered in the text as a function of the number of factors included. Bottom panels: Corresponding p-values of tests that the expected squared prediction errors of the various models are equal, as a function of the number of factors included. The right panels report models that include the CP factor while the left panels report those that exclude it.

same benefit by using nonlinear factors in a linear predictive regression of returns instead and thus avoid the need to decide on the order of the polynomial to use. Furthermore, our $\kappa = 2$ and $\kappa = 3$ models also perform just as well as Ludvigson & Ng's best predictor (shown by a dotted line) based on factors $f_1, f_1^3, f_3, f_4, f_8$ that were obtained by an extensive model selection procedure.

Figure 3 also shows the p-values of a test that these predictions are statistically significantly different. This test is based on comparing the mean square prediction errors, using Newey-West standard errors with a maximum lag of 18 months.[9] The blue line shows that nonlinearity does yield statistically significantly better predictions than a fully linear model, while the orange line shows that the two nonlinear approaches are not statistically significantly different, as expected. A similar exercise, now including the financial CP factor, supports the same conclusion: The nonlinear PCA approach avoids an extensive model selection step and performs no worse than the best approach based on linear PCA in all cases.

It should be noted that the model selection procedure in Ludvigson and Ng (2009) that identified the importance of the f_1^3 regressor was performed using the entire sample, whereas our nonlinear PCA step only uses the earlier half of the sample — hence, if anything, this comparison puts our method at a disadvantage. The similarity of their results to ours also indicates that their results are not driven by their use of future data to find the functional form of the best model.

[9]The use of an 18 month maximum lag is standard in this literature: A minimum of 12-month lag is needed to account for seasonal effects, which is increased by 50% to circumvent the fact that the Newey-West estimator downweights the larger lags significantly.

Finally, it is noteworthy to observe that the nonlinearity of the factors in our method is determined before using the returns data (which is the dependent variable in the regression). This suggests that our nonlinear PCA procedure actually detects a nonlinearity in the data that not only provides just a better fit but also apparently detects genuine economic features, since it improves the factors' predictive abilities, before the bond data to be predicted are even included in the analysis. This observation also indicates that the nonlinear terms in the polynomial regression of Ludvigson & Ng was mostly needed to overcome the limitations of the linear PCA factors.

5 Conclusion

We have introduced a novel nonlinear generalization of principal component analysis (PCA) that offers a number of unique advantages. It generates truly independent (as opposed to the merely uncorrelated) factors, thus maximizing the amount of "information compression". Thanks to the use of the theory of optimal transport and Brenier maps, a unique optimal solution can be established and efficient computational algorithms can be devised. The method makes conceptual connections with entropy maximization as it relies on a new multivariate additive entropy decomposition to determine the principal nonlinear components that capture most of the information content of the data. An application to the prediction of excess bond returns from a large number of macro factors reveals that the method is able to naturally capture economically relevant nonlinearities and, as a by-product, reduces the reliance on extensive model selection procedures.

A Proofs

Proof of Lemma 1. We first observe that the contribution to the entropy $H = -\int f(y) \ln f(y)\, dy$ coming from values of y in a set $\mathcal{N}$ of null Lebesgue measure is zero since the density $f(y)$ is uniformly bounded (by a constant $\bar{f} > 0$): $\int_{y\in\mathcal{N}} f(y) |\ln f(y)|\, dy \leq \max\left\{e^{-1}, \bar{f}\left|\ln \bar{f}\right|\right\} \int_{y\in\mathcal{N}} dy = 0$. This implies that we can ignore a set of null measure Lebesgue from the entropy integral in our subsequent analysis.

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 $J(y) \equiv \partial T(y)/\partial y'$ is almost everywhere well-defined as Brenier maps between bounded Lebesgue densities 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, via the change of variable: $x = T(y)$ (so

that $dx = \det\left(\frac{\partial T(y)}{\partial y'}\right) dy$):

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

where $J\left(y\right) = \frac{\partial T(y)}{\partial y'}$ and where the inverse $T^{-1}\left(x\right)$ is also defined almost everywhere. Next, we have

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

where

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

Each term can then be simplified:

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

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

To evaluate B, we use the equality:

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

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

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

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 1. 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, for Gaussian random variables, one can always find a linear coordinate system that makes each coordinate statistically independent. Then, a monotone

mapping $x_i = g_i(y_i)$ that maps one univariate Gaussian y_i variable onto another (x_i) must be linear. It follows that $T(y)$ must be linear. Also, 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. ■

Proof of Theorem 2. By definition of the entropy $H[f]$ and of conditional densities, we have:

$$\begin{aligned} H\left[f_{y^{k^*},y^o}\right] - H\left[f_{y^{k^*}}\right] &= -\int\int f\left(y^{k^*},y^o\right)\ln f\left(y^{k^*},y^o\right)dy^{k^*}dy^o + \int\int f\left(y^{k^*},y^o\right)dy^o \ln f\left(y^{k^*}\right)dy^{k^*} \\ &= -\int\int f\left(y^{k^*},y^o\right)\ln\frac{f\left(y^{k^*},y^o\right)}{f\left(y^{k^*}\right)}dy^{k^*}dy^o \\ &= -\int f\left(y^{k^*}\right)\left(\int f\left(y^o|y^{k^*}\right)\ln f\left(y^o|y^{k^*}\right)dy^o\right)dy^{k^*}. \end{aligned}$$

Next, we use the known fact that, for a given variance, the Gaussian is the density that maximizes entropy,[10] i.e., for any random vector v with density $f_V(v)$ and finite variance $\mathrm{Var}[v]$, we have

$$\int f_V(v)\ln f_V(v)\,dv \leq \int \Phi_{\mathrm{Var}[v]}(v)\ln\Phi_{\mathrm{Var}[v]}(v)\,dv, \tag{9}$$

where $\Phi_{\mathrm{Var}[v]}$ denotes the a Gaussian density with variance $\mathrm{Var}[v]$ and zero mean. Next, if the elements of v are selected to be along the principal axes of its variance ellipsoid (so that v_i and v_j are uncorrelated), we have

$$\int \Phi_{\mathrm{Var}[v]}(v)\ln\Phi_{\mathrm{Var}[v]}(v)\,dv = \sum_{i=1}^{\dim V}\ln\sqrt{2\pi e\,\mathrm{Var}[v_i]}. \tag{10}$$

Now, letting q_i denote the normalized eigenvectors of $\mathrm{Var}\left[y^o|y^{k^*}\right]$, we can use Equations (9) and (10) with $v_i = q_i' y^o$ to write:

$$\begin{aligned} H\left[f_{y^{k^*},y^o}\right] - H\left[f_{y^{k^*}}\right] &\leq \int f\left(y^{k^*}\right)\left(\sum_{i=1}^{\dim y^o}\ln\sqrt{2\pi e\,\mathrm{Var}\left[y_i^o|y^{k^*}\right]}\right)dy^{k^*} \\ &= \sum_i \frac{1}{2}E\left[\ln\left(2\pi e\,\mathrm{Var}\left[y_i^o|y^{k^*}\right]\right)\right]. \end{aligned} \tag{11}$$

Next, we observe that, by Jensen's inequality,

$$E\left[\ln\left(2\pi e\,\mathrm{Var}\left[y_i^o|y^{k^*}\right]\right)\right] \leq \ln\left(2\pi e E\left[\mathrm{Var}\left[y_i^o|y^{k^*}\right]\right]\right) \tag{12}$$

[10] This can be shown by solving the Lagrangian of the constrained optimization problem: $\int f(v)\ln f(v)\,dv - \lambda_1\int f(v)\,dv - \lambda_2\int v f(v)\,dv - \lambda_3\int v^2 f(v)\,dv$, with Lagrange multipliers $\lambda_1, \lambda_2, \lambda_3$.

and that, by the law of total variance,

$$E\left[\operatorname{Var}\left[y_i^o|y^{k^*}\right]\right] = \operatorname{Var}\left[y_i^o\right] - \operatorname{Var}\left[E\left[y^o|y^{k^*}\right]\right] \leq \operatorname{Var}\left[y_i^o\right]. \tag{13}$$

Combining Equations (11), (12) and (13) yields:

$$H\left[f_{y^{k^*},y^o}\right] - H\left[f_{y^{k^*}}\right] \leq \sum_i \ln\left(\sqrt{2\pi e \operatorname{Var}\left[y_i^o\right]}\right) = \operatorname{tr}\ln\left(\sqrt{2\pi e \operatorname{Var}\left[y^o\right]}\right)$$

where the last equality holds since $\operatorname{Var}[y^o]$ is diagonal by assumption. Since the trace is invariant to orthogonal coordinate transformations, the equality also holds if the elements of y^o are not necessarily chosen to be along the principal axes of its covariance matrix. ■

Proof of Theorem 3. We will show, in turn, consistency of $\hat{J}(y)$, $\ln\hat{J}(y)$, $\widehat{\bar{J}}$ and $\hat{u}^j$.

We first note that the definition of $\hat{J}_{ij}(y)$ is simply a symmetrized ($\hat{J}_{ij}(y) = \hat{J}_{ji}(y)$) version of the simpler definition $\hat{J}_{ij}(y) = \left(\hat{T}_i(y+\varepsilon_n e_j) - \hat{T}_i(y-\varepsilon_n e_j)\right)/2\varepsilon_n$. We show consistency of the latter (which trivially implies convergence of its symmetrized version). Let $r_n = \sup_{y\in\mathbb{R}^d}\left|\hat{T}(y) - T(y)\right|$, which satisfies $r_n \xrightarrow{p} 0$ by the assumed uniform consistency of $\hat{T}(y)$ and select ε_n such that $\varepsilon_n \longrightarrow 0$ and $r_n/\varepsilon_n \xrightarrow{p} 0$. By the triangle inequality, we have

$$\left|\hat{J}_{ij}(y) - J_{ij}(y)\right| \leq \left|\frac{\hat{T}_i(y+\varepsilon_n e_j) - \hat{T}_i(y-\varepsilon_n e_j)}{2\varepsilon_n} - \frac{T_i(y+\varepsilon_n e_j) - T_i(y-\varepsilon_n e_j)}{2\varepsilon_n}\right| + R_n(y)$$

where $R_n \equiv \sup_{y\in\mathbb{R}^d}\left|\frac{T_i(y+\varepsilon_n e_j) - T_i(y-\varepsilon_n e_j)}{2\varepsilon_n} - J_{ij}(y)\right|$. Next, rearranging and using the triangle inequality again, we have

$$\begin{aligned}\left|\hat{J}_{ij}(y) - J_{ij}(y)\right| &\leq \frac{\left|\hat{T}_i(y+\varepsilon_n e_j) - T_i(y+\varepsilon_n e_j)\right|}{2\varepsilon_n} + \frac{\left|\hat{T}_i(y-\varepsilon_n e_j) - T_i(y-\varepsilon_n e_j)\right|}{2\varepsilon_n} + R_n \\ &\leq r_n/\varepsilon_n + R_n \xrightarrow{p} 0\end{aligned}$$

where we have used (i) the definition of r_n and the fact that $r_n/\varepsilon_n \xrightarrow{p} 0$ by construction and (ii) the fact that $R_n \longrightarrow 0$ by the uniform continuity of $J(y)$ and the definition of the partial derivative. Hence $\hat{J}(y)$ is uniformly consistent and so is its symmetrized version (hereafter also denoted $\hat{J}(y)$ in a slight above of notation), which is also guaranteed to be diagonalizable.

Next, the consistency of $\ln\hat{J}(y)$ is shown using the fact that a function of a matrix is continuous by Theorem 6.2.37 of Horn and Johnson (1991), provided the scalar version of this function is continuous on an open set that includes all the eigenvalues of the diagonalizable matrix $\hat{J}(y)$. This the case here since the eigenvalues of $J(y)$ are bounded away from zero, as are those of $\hat{J}(y)$ for n sufficiently large and the $\ln$ function is continuous on a set that excludes a neighborhood of the origin. We can similarly establish uniform continuity by observing that the lower bound on the eigenvalue is uniform so that the modulus on continuity of the $\ln$ on a common open set that includes all the eigenvalues (for different y) is uniform as well. This shows uniform consistency of $\ln\hat{J}(y)$.

To show consistency of $\widehat{\bar{J}}$, we write, for one element ij of the $\widehat{\bar{J}}$ matrix:

$$\begin{aligned}\left|\widehat{\bar{J}}_{ij}-\bar{J}_{ij}\right| &= \left|\int_{\|y\|\leq\bar{y}_n}\hat{f}\left(y\right)\left(\ln\hat{J}\left(y\right)\right)_{ij}dy-\int f\left(y\right)\left(\ln J\left(y\right)\right)_{ij}dy\right| \\ &\leq \int\left|\hat{f}\left(y\right)\right|\left|\left(\ln\hat{J}\left(y\right)\right)_{ij}-\left(\ln J\left(y\right)\right)_{ij}\right|dy \\ &\quad+\int\left|\hat{f}\left(y\right)1\left\{\|y\|\leq\bar{y}_n\right\}-f\left(y\right)\right|\left|\left(\ln J\left(y\right)\right)_{ij}\right|dy\end{aligned} \tag{14}$$

where the first term converges since we just showed $\left|\left(\ln\hat{J}\left(y\right)\right)_{ij}-\left(\ln J\left(y\right)\right)_{ij}\right| \xrightarrow{p} 0$ uniformly and since $\hat{f}\left(y\right)$ is absolutely integrable by assumption. To bound the second term of Equation (14), we define $s_n = \sup_{y\in\mathbb{R}^d}\left|\hat{f}\left(y\right)-f\left(y\right)\right|$ and note that $s_n \xrightarrow{p} 0$ by assumption. We select a truncation sequence $\bar{y}_n$ such that $\bar{y}_n \longrightarrow \infty$ and $\bar{y}_n^d s_n \xrightarrow{p} 0$. We also introduce $\bar{\Lambda} = \left|\max\left\{\ln\lambda_{\min},\ln\lambda_{\max}\right\}\right|$ where $\lambda_{\min}$ and $\lambda_{\max}$ are, respectively, the uniform lower and upper bound on the eigenvalues of $J\left(y\right)$, assumed finite and nonzero. We can then write:

$$\begin{aligned}&\int\left|\hat{f}\left(y\right)1\left\{\|y\|\leq\bar{y}_n\right\}-f\left(y\right)\right|\left|\left(\ln J\left(y\right)\right)_{ij}\right|dy \\ \leq\ &\bar{\Lambda}\int\left|\hat{f}\left(y\right)1\left\{\|y\|\leq\bar{y}_n\right\}-f\left(y\right)\right|dy \\ =\ &\bar{\Lambda}\int_{\|y\|\leq\bar{y}_n}\left|f\left(y\right)-\hat{f}\left(y\right)\right|dy+\bar{\Lambda}\int_{\|y\|\geq\bar{y}_n}f\left(y\right)dy \\ \leq\ &\bar{\Lambda}S^d\bar{y}_n^d s_n+\bar{\Lambda}\int_{\|y\|\geq\bar{y}_n}f\left(y\right)dy\end{aligned}$$

where S^d is the volume of a d-dimensional unit sphere. Both terms converge to zero since $\bar{y}_n^d s_n \xrightarrow{p} 0$ and $\bar{y}_n \longrightarrow \infty$ with $f\left(y\right)$ being absolutely integrable. This shows consistency of $\widehat{\bar{J}}$.

Finally, the fact that the eigenvalues and eigenvectors of $\widehat{\bar{J}}$ converge to those of $\bar{J}$ follows from standard first-order matrix perturbation theory (Stewart and Sun (1990)), under the assumption that eigenvalues of $\bar{J}$ are nondegenerate. ■

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.

ANDREWS, D. W. K. (1995): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–596.

ATHEY, S., AND G. W. IMBENS (2019): "Machine Learning Methods Economists Should Know About," Discussion Paper 1903.10075, ArXiv preprint.

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.

CHERNOZHUKOV, V., A. GALICHON, M. HALLIN, AND M. HENRY (2017): "Monge–Kantorovich depth, quantiles, ranks and signs," *Annals of Statistics*, 45, 223–256.

COCHRANE, J. H., AND M. PIAZZESI (2005): "Bond Risk Premia," *American Economic Review*, 95, 138–160.

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

DE PHILIPPIS, G., AND A. FIGALLI (2014): "The Monge-Ampère equation and its link to optimal transportation," *Bulletin of American Mathematical Society*, 51, 527–580.

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.

EKELAND, I., A. GALICHON, AND M. HENRY (2011): "Comonotonic Measures of Multivariate Risks," *Mathematical Finance*, 22, 109–132.

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.

GUNSILIUS, F. (2019): "On the convergence rate of potentials of Brenier maps," Working Paper, Brown University.

GUNSILIUS, F., AND S. SCHENNACH (2019): "Subsampling validity for estimated Brenier maps," Working Paper, Brown University.

HORN, R. A., AND C. R. JOHNSON (1991): *Matrices and functions*p. 382–560. Cambridge University Press, Cambridge.

HÜTTER, J.-C., AND P. RIGOLLET (2019): "Minimax rates of estimation for smooth optimal transport maps," Discussion Paper 1905.05828, ArXiv preprint.

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

JURADO, K., S. LUDVIGSON, AND S. NG (2015): "Measuring Uncertainty," *American Economic Review*, 105, 1177–1215.

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.

LEE, Y. J. (2018): "Sieve estimation of optimal transport with applications to multivariate quantiles and matching," Working Paper, UCL.

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.

LUCET, Y. (1997): "Faster than the fast Legendre transform, the linear-time Legendre transform," *Numerical Algorithms*, 16, 171–185.

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

LUDVIGSON, S. C., AND S. NG (2009): "Macro Factors in Bond Risk Premia," *Review of Financial Studies*, 22, 5027–5067.

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.

SCHENNACH, S. M. (2014): "Entropic Latent Variable Integration via Simulation," *Econometrica*, 82, 345–386.

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.

STEWART, G. W., AND J. SUN (1990): *Matrix perturbation Theory*. Academic Press, Boston.

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*, 44, 293–335.

STOCK, J. H., AND M. W. WATSON (2002): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business & Economic Statistics*, 20, 147–162.

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.