EFFICIENT CONVEX PCA WITH APPLICATIONS TO WASSERSTEIN GEOIDESIC PCA AND RANKED DATA

STEVEN CAMPBELL AND TING-KAM LEONARD WONG

Abstract. Convex PCA, which was introduced in Bigot et al. [6], is a dimension reduction methodology for data with values in a convex subset of a Hilbert space. This setting arises naturally in many applications, including distributional data in the Wasserstein space of an interval, and ranked compositional data under the Aitchison geometry. Our contribution in this paper is threefold. First, we present several new theoretical results including consistency as well as continuity and differentiability of the objective function in the finite dimensional case. Second, we develop a numerical implementation of finite dimensional convex PCA when the convex set is polyhedral, and show that this provides a natural approximation of Wasserstein geodesic PCA. Third, we illustrate our results with two financial applications, namely distributions of stock returns ranked by size and the capital distribution curve, both of which are of independent interest in stochastic portfolio theory.

1. Introduction

Principal component analysis (PCA) is a fundamental dimension reduction technique for revealing the underlying structure of the data. The standard Euclidean formulation of PCA operates by projecting the data orthogonally onto a lower dimensional subspace which minimizes the residual error, and can be solved analytically via the eigenvalue decomposition of the empirical covariance matrix. Modern challenges in data science motivated various extensions of Euclidean PCA, such as sparsity [38], exponential family PCA [13], functional PCA [34], principal geodesic analysis for manifold-valued data [21], to name a few. For systematic overviews of PCA and its variants we refer the reader to [1, 24].

In this paper, we study convex PCA (CPCA) where the data is constrained to lie in a convex subset of a Hilbert space. CPCA was first introduced in [6] to study a notion of geodesic PCA (GPCA) on the Wasserstein space on an interval, where each data point is a probability distribution. This approach is based on the fact that the Wasserstein space (of order $p = 2$) on an interval is isometric to a convex subset of an $L^2$ space (see Section 5.1). Statistical analysis of distributional data using optimal transport, and principal component analysis in particular, has attracted a lot of attention in recent years. For example, a “log” PCA on the Wasserstein space was studied in [9], while [28] introduced projection methods to perform PCA and regression for one-dimensional distributional data. Other recent statistical works include Wasserstein autoregressive models for distributional time
2 STEVEN CAMPBELL AND TING-KAM LEONARD WONG

Figure 1. Euclidean and convex PCA for a simulated two dimensional data-set with values in a cone (shaded area). In Euclidean PCA, the orthogonal projection onto the principal component may lie outside of the convex set. Restricting the projection to the convex set leads to convex PCA.

series [37], Wasserstein PCA for circular measures [5], and regression models [12]. Also see [29] for an accessible introduction to optimal transport and its applications in statistics and machine learning. Apart from Wasserstein geodesic PCA, convex PCA is also useful in other settings. For example, in Section 6 we apply CPCA to a ranked data-set where each observation is a vector \((y_1, \ldots, y_d)\) satisfying the convex constraint \(y_1 \geq \cdots \geq y_d\). In Figure 1 we compare Euclidean and convex PCA using a simulated two dimensional data-set, where the data is restricted to lie in a convex cone. Geometrically, the projection onto the Euclidean principal component (PC) may not lie in the convex set, making it difficult to interpret the result. By restricting the projected points to the convex set, we obtain (using the algorithm developed in Section 4) the convex PC. Note that the computation is more involved since the projection is no longer a linear operator.

1.1. Main contributions and organization of the paper. Our contribution in this paper is threefold. First, we obtain new theoretical results about CPCA concerning consistency and analytical properties of the objective function. Second, we present a numerical implementation of finite-dimensional CPCA when the convex state space is polyhedral, and apply it to solve a finite-dimensional approximation of Wasserstein geodesic PCA on an interval. Third, we provide two novel financial applications of CPCA motivated by stochastic portfolio theory [19] in mathematical finance.

We begin by recalling in Section 2 the formulation of convex PCA following the treatment of [6]. Convex PCA has two formulations, namely global and nested. In this paper, we focus on the nested version for interpretability and computational purposes. We also define a notion of explained variation in this context.
In Section 3, we study theoretical properties of CPCA, mostly in the finite dimensional case. We consider (i) consistency in measure, (ii) consistency of finite dimensional approximations of the underlying Hilbert space, and (iii) analytical properties of the objective function of an equivalent formulation of nested CPCA. These results place CPCA on a rigorous basis and pave the way for numerical implementation.

A computational algorithm for CPCA is the subject of Section 4. When the Hilbert space is finite dimensional and the convex set is polyhedral (i.e., given by the intersection of a finite collection of half spaces), we present an efficient algorithm which can be readily implemented. For especially large and high dimensional problems, we also provide a C++ implementation that takes advantage of parallel computing techniques.

Wasserstein geodesic PCA on an interval is studied in Section 5. After recalling the isometry between the Wasserstein space and a closed convex subset of an $L^2$ space, we propose a natural finite dimensional approximation of Wasserstein geodesic PCA by projecting the data onto suitable subspaces of the underlying Hilbert space, and note that this approximation satisfies the consistency result in Section 3.2. Numerical implementation of Wasserstein GPCA was considered (among related approaches such as projected PCA) in [28] using a quadratic B-spline expansion of the distribution function. Here, our approximation is based on discretizing the transport map and is exact when all data points are in the given finite dimensional subspace.

Our work on CPCA was motivated by our previous paper [8] on stochastic portfolio theory (SPT). SPT is a mathematical framework introduced by Fernholz [19] for studying macroscopic properties of equity markets and portfolio selection. A key empirical observation of this theory is that the behaviours of stocks have a systematic dependence on their relative ranks (measured by market capitalization), and this dependence may be exploited by carefully constructed portfolios to outperform the market. In [8], we studied a portfolio optimization problem using rank-based portfolios and were led to quantify statistically the rank-based behaviours. In Section 5.3, we apply Wasserstein geodesic PCA to distributions of US stock returns ranked by size, and show that the first two convex principal components can be interpreted in terms of volatility and skewness. In Section 6, we consider the US capital distribution curve which captures market capitalization weights arranged in descending order. By viewing them as (ranked) compositional data under the Aitchison geometry of the simplex [18] and applying CPCA, we show that the projection onto the first convex principal component corresponds closely to market diversity, an entropy-like quantity which measures the concentration of market capital and correlates significantly with the relative performance of active portfolio managers. These properties have significant implications in the construction of macroscopic stochastic models of equity markets as well as portfolio selection.

Finally, in Section 7, we summarize our contributions and discuss possible directions for future research. The Appendix contains proofs of most technical results as well as further details of our numerical implementation.

1The codes are available at: https://github.com/stevenacampbell/ConvexPCA
2. Convex PCA

Convex PCA was originally proposed in [6] to study Wasserstein geodesic PCA on the real line. In this section we recall the general formulation of CPCA which has two versions, namely global and nested. We also give a definition of explained variation in this context. Wasserstein geodesic PCA will be revisited in Section 5 where we propose and implement a finite dimensional approximation.

Let $H$ be a Hilbert space (over $\mathbb{R}$) with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. The dimension of $H$ can be finite or infinite. In the latter case, we assume that $H$ is separable. For $x, y \in H$ and $E \subset H$, let $d(x, y) = \|x - y\|$ and $d(x, E) = \inf_{z \in E} d(x, z)$. Next, let $X$ be a given nonempty closed convex subset of $H$ which will serve as the state space of the data. In some theoretical results we will further assume that $X$ is compact.

If $A$ and $B$ are nonempty closed subsets of $X$, we let

$$ h(A, B) = \max \left\{ \sup_{x \in A} d(x, B), \sup_{y \in B} d(y, A) \right\} $$

be the Hausdorff distance between $A$ and $B$. It is well known that $h$ defines a metric on the space of nonempty and compact subsets of $X$. In what follows $k$ is always a non-negative integer. For $A \subset H$, we let $\dim A$ be the dimension of the smallest affine subspace of $H$ containing $A$.

**Definition 2.1.** We endow the following spaces with the Hausdorff metric $h$.

(i) Let $CL(X)$ be the space of nonempty and closed subsets of $X$.

(ii) Let $CC_k(X)$ be the family of convex sets $C \in CL(X)$ such that $\dim C \leq k$.

(iii) Given a reference point $\pi \in X$, we let $CC_{\pi,k}(X)$ be the family of sets $C \in CC_k(X)$ such that $\pi \in C$.

By [3] Proposition 3.2, if $X$ is compact then all three spaces in Definition 2.1 are compact. Also, by [6] Proposition A.3, $d(\cdot, \cdot)$ is continuous on $CL(X)$. If $C \subset H$ is a (nonempty) closed convex set, we let $\Pi_C : X \to C$ be the orthogonal projection which is well-defined by the Hilbert projection theorem.

We are almost ready to state the two versions of CPCA. Let $\mu$ be a Borel probability measure on $X$ representing the distribution of the data; in applications, $\mu$ is usually an empirical measure, say $\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}$, where $\delta_{x_i}$ is the point mass at the data point $x_i \in X$. Let $x$ be an $X$-valued random element whose distribution is $\mu$. We assume $\mu$ has finite second moment, i.e., $\mathbb{E}_\mu[\|x\|^2] < \infty$. We also let $\pi \in X$ be a fixed reference point which is typically the mean $\mathbb{E}_\mu[x]$ of $\mu$.

**Definition 2.2** (Convex PCA). Let $\mu$ be a Borel probability measure on $X$ with finite second moment, $x \sim \mu$ and $\pi \in X$. For $C \in CL(X)$, define

$$ J(C; \mu) = \mathbb{E}_\mu[d^2(x, C)]. $$

(i) (Global CPCA) Given $k \geq 0$, a $(k, \pi, \mu)$-global convex principal component (CPC) is a set $C_k$ such that

$$ C_k = \arg \min_{C \in CC_{\pi,k}(X)} J(C; \mu). $$

(ii) (Nested CPCA) Let $C_0 = \{\pi\}$. For $k \geq 1$, we say that $C_k$ is a $(k, \pi, \mu)$-nested convex principal component (CPC) if there exist $C_1, \ldots, C_{k-1}$ such
that \( C_0 \subset C_1 \subset \cdots \subset C_{k-1} \subset C_k \) and, for each \( \ell = 1, \ldots, k \),

\[
(2.4) \quad C_\ell \in \arg\min_{C \in CC_{x,\ell}(X), C \supset C_{\ell-1}} J(C; \mu).
\]

In the Euclidean setting (i.e., \( X = H \)), the global and nested PCA problems are equivalent. Here, they are equivalent only when \( k = 1 \). Although solutions to the nested CPCA problem are generally not optimal in the global sense, they are easier to compute and interpret. Hence, we focus on nested CPCA in later sections. Using the compactness and continuity results cited above, it can be shown (see [6, Theorem 3.1]) that when \( X \) is compact both the global and nested CPCA problems \((2.3)\) and \((2.4)\) have non-empty solution sets for any \( k \). Also, as long as \( \mu \) is not supported on a \((k-1)\)-dimensional convex subset of \( X \), in \((2.3)\) and \((2.4)\) the objective values are unchanged if we restrict to sets of the form

\[
(2.5) \quad C_k = (\bar{x} + \text{span}\{p_1, \ldots, p_k\}) \cap X,
\]

where \( \{p_1, \ldots, p_k\} \) is a sequence of orthonormal vectors in \( H \) [6, Propositions 3.3 and 3.4]. This representation will be used in the analysis in Sections 3 and 4. In \((2.5)\), for an optimal set \( \{p_1^*, \ldots, p_j^*\} \), we call \( p_j^* \) the \( j \)-th convex principal direction

A useful concept in PCA is the proportion of explained variation. In the classical setting, it can be expressed in terms of the eigenvalues of the covariance matrix. Since the projection map \( \Pi_C \) is nonlinear when \( C \) is not a subspace, eigenvalue decomposition is no longer available. Nevertheless, we may still define a notion of explained variation.

**Definition 2.3** (Total and explained variation). The total variation of the data \( x \sim \mu \) with respect to a given reference point \( \bar{x} \in X \) is defined by

\[
(2.6) \quad TV = \mathbb{E}_\mu [d(x, \bar{x})^2].
\]

To avoid triviality, we assume \( TV > 0 \). Let \( C \in CC_{x,k}(X) \). The proportion of variation explained by \( C \) is defined by

\[
(2.7) \quad EV(C) = \frac{\mathbb{E}_\mu [d(\Pi_C x, \bar{x})^2]}{TV}.
\]

**Remark 2.4.**

(i) Clearly \( EV(C) \leq 1 \) and, if \( \mu(C) = 1 \), then \( EV(C) = 1 \). More generally, if \( C_1 \subset C_2 \subset \cdots \) is a sequence of global or nested CPCs, then \( EV_k = EV(C_k) \) is non-decreasing in \( k \). In this case, we may interpret \( (EV_k - EV_{k-1})TV \) as the additional variation explained by the \( k \)-th component.

(ii) Definition \((2.3)\) reduces to the usual definition for Euclidean PCA if we let \( X = H \), \( \bar{x} = \mathbb{E}_\mu[x] \) and \( C \) be an affine subspace containing \( \bar{x} \).

(iii) Since the definition of \( EV(C) \) requires evaluating the projection map \( \Pi_C \) for individual data points, it is generally costly to compute \( EV(C) \) for large data sets.

**Proposition 2.5.** Let \( X \) be compact, and let \( C_k, k = 1, 2, \ldots, \), be an increasing sequence of convex subsets of \( X \). If \( \bigcup_k C_k = X \), then \( \Pi_{C_k} \to \Pi_X \) pointwise on \( X \). Consequently, we have \( EV(C_k) \to 1 \) as \( k \to \infty \).

**Proof.** See Appendix \[A.1\] \( \square \)

\(^2\)By an abuse of terminology, we also call it the \( j \)-th convex principal component which is often used in the Euclidean context.
3. Finite dimensional convex PCA

While infinite dimensional Hilbert spaces arise naturally in theory (e.g., Wasserstein geodesic PCA), in implementations we almost always work with a finite dimensional space due to approximation or discretization. In this section we establish some theoretical properties of CPCA mostly in the finite dimensional case. Sections 3.1 and 3.2 study consistency properties in measure and in the dimension of $H$. In Section 3.3 we consider an equivalent formulation of nested CPCA and prove continuity and differentiability of the objective function.

3.1. Consistency in measure. In principal component analysis, as well as other statistical methodologies, consistency is a desirable, if not indispensable, property. In our context, by consistency we mean the stability of the optimization problems (2.3) and (2.4) as the input measure $\mu$ varies. For example, under i.i.d. sampling the empirical measure converges to the population as the sample size tends to infinity. In this subsection we provide consistency results in the spirit of [6, Section 6], thereby fixing a technical gap in [6, Lemma 6.1]. Using similar techniques, we study in Section 3.2 the stability of CPCA when the underlying Hilbert space $H$ is approximated by a sequence of finite dimensional subspaces.

In this subsection we assume that $X \subset H$ is a compact convex set. Let $P(X)$ be the space of Borel probability measures on $X$ and equip it with the topology of weak convergence. Since $X$ is compact, weak convergence is equivalent to convergence in the $p$-Wasserstein distance, for any $p \geq 1$ (see e.g., [36, Theorem 6.9]). For later use, we recall that for $\mu, \nu \in P(X)$, the $p$-Wasserstein distance is defined by

\[
W_p(\mu, \nu) = \left( \inf_{\pi \in \Pi(\mu, \nu)} \int d(x, y)^p d\pi(x, y) \right)^{1/p},
\]

where $\Pi(\mu, \nu)$ is the set of couplings of the pair $(\mu, \nu)$. Also recall the functional $J(C; \mu)$ defined by (2.2).

When stating and proving consistency results, it is convenient to formulate the optimization problems in terms of indicator functions. Given a subset $A$ of an ambient space, let $\chi_A$ be its indicator function which is 0 on $A$ and $+\infty$ on its complement. Then the global CPCA problem (2.3) is equivalent to

\[
(3.2) \quad \inf_{C \in CL(X)} \left\{ J(C; \mu) + \chi_{CC_{k-1}}(C) \right\}.
\]

Similarly, the nested CPCA problem (2.4) can be written as

\[
(3.3) \quad \inf_{C \in CL(X)} \left\{ J(C; \mu) + \chi_{C \subset CC_{k-1}}(C) \right\},
\]

where $C_{k-1}$ is a given $(k-1, \pi, \mu)$-nested CPC. Also, we will be using the notion of $\Gamma$-convergence (denoted by $\Gamma$-lim) whose definition is recalled in Appendix C. It also contains other relevant definitions and results, such as Kuratowski convergence of sets, that are used in the proofs.

We first give a fairly standard but general result which will serve as a framework for proving specific consistency statements.

**Theorem 3.1** (General consistency). Let $X \subset H$ be compact and convex. Let $\mathcal{C}_n, \mathcal{C} \subset CL(X)$ and $\mu_n, \mu \in \mathcal{P}(X)$. Suppose $\Gamma$-lim$_{n \to \infty} \chi_{\mathcal{C}_n} = \chi_{\mathcal{C}}$ and $\mu_n \to \mu$ in $\mathcal{P}(X)$. Then

\[
(3.4) \quad \lim_{n \to \infty} \inf_{C \in CL(X)} \left\{ J(C; \mu_n) + \chi_{\mathcal{C}_n}(C) \right\} = \inf_{C \in CL(X)} \left\{ J(C; \mu) + \chi_{\mathcal{C}}(C) \right\}.
\]
Moreover, if \( C_n \in \arg\min_{C \in \mathcal{CL}(X)} \{ J(C; \mu_n) + \chi_{\epsilon_n}(C) \} \) for \( n \geq 1 \), then any accumulation point of \((C_n)_{n \geq 1}\) (under the Hausdorff distance) attains the infimum on the right hand side of (3.4).

To apply Theorem 3.1, we need to verify the condition \( \Gamma\lim_{n \to \infty} \chi_{\epsilon_n} = \chi_{\epsilon} \). When this holds, we have convergence of the objective value as well as convergence of the solution in a suitable sense. Unfortunately, the proof of [6, Lemma 6.1], which verifies this condition and underlies the consistency statements there, contains a technical gap which we were able to fix only when \( H \) is finite dimensional. The proofs of Theorem 3.1 and the following results are given in Appendix A.2 which also contains more discussion about the technical issues involved in [6, Lemma 6.1].

**Theorem 3.2** (Consistency of global CPCA). Suppose \( H \) is finite dimensional and \( X \) has non-empty interior. Let \( \bar{\tau} \in \text{int } X \), \( \tau_n \to \bar{\tau} \) in \( X \) and \( \mu_n \to \mu \) in \( \mathcal{P}(X) \). For \( k \geq 1 \), consider

\[
\epsilon_n = CC_{\tau_n,k}(X) \quad \text{and} \quad \epsilon = CC_{\bar{\tau},k}(X).
\]

Then \( \Gamma\lim_{n \to \infty} \chi_{\epsilon_n} = \chi_{\epsilon} \) and hence the conclusions of Theorem 3.1 hold for the corresponding global CPCA problems formulated via (3.2).

We also give an analogous consistency result for nested CPCA which is not explicitly treated in [6]. Note that it is formulated in terms of a subsequence because the convex PCs are not necessarily unique.

**Theorem 3.3** (Consistency of nested CPCA). Suppose \( H \), \( X \), \( \bar{\tau} \) and \( k \) are as in Theorem 3.2. Suppose \( \tau_n \to \bar{\tau} \) in \( X \), \( \mu_n \to \mu \) in \( \mathcal{P}(X) \) and that \( \dim(\text{supp}(\mu)) \geq k \). For each \( n \), suppose

\[
\{ \tau_n \} = C_{n,0} \subset C_{n,1} \subset \cdots \subset C_{n,k}
\]

is a sequence such that each \( C_{n,j} \) is a \( (j, \tau_n, \mu_n) \)-nested convex principal component. Then there is a subsequence \( n' \) along which the following statements hold:

(i) For each \( j \) we have \( C_{n',j} \to C_j \in CC_{\tau,j}(X) \).

(ii) The sequence \( (C_{j,k})_{j=0}^{k} \) of sets is increasing in \( j \) and forms a sequence of nested convex principal components with respect to \( \mu \) and \( \bar{\tau} \).

(iii) For each \( n' \) and \( j \), let

\[
\epsilon_{n',j} = \{ C \in CC_{\tau',j}(X) : C \supset C_{n',j-1} \},
\]

\[
\epsilon_j = \{ C \in CC_{\tau,j}(X) : C \supset C_{j-1} \}.
\]

Then \( \Gamma\lim_{n' \to \infty} \chi_{\epsilon_{n',j}} = \chi_{\epsilon_j} \) and hence the conclusions of Theorem 3.1 hold for the corresponding nested CPCA problems formulated via (3.3).

### 3.2 Consistency of finite dimensional approximation

Convex PCA and Wasserstein geodesic PCA on the real line are generally infinite dimensional optimization problems. To implement them in practice, finite dimensional approximations are necessary. Using similar ideas as in Section 3.1, we formulate such an approximation and show its consistency as the dimension tends to infinity.

Let \( H \) be an infinite dimensional separable Hilbert space and \( X \subset H \) be a compact convex set. Let \( \{ H_n \}_{n \geq 1} \) be an increasing sequence of finite dimensional vector subspaces of \( H \), such that \( \bigcup_{n=1}^{\infty} H_n \) is dense in \( H \). For example, we may consider an orthonormal basis \( \{ \varphi_n \}_{n \geq 1} \) and let \( H_n = \text{span}\{ \varphi_1, \ldots, \varphi_n \} \). Let \( \Pi_n : H \to H_n \) be the orthogonal projection onto \( H_n \). We impose the simplifying assumption that
Π_n X ⊂ X for each n. Note that this condition holds in our finite dimensional approximation of Wasserstein geodesic PCA which will be treated in Section 5.

Let π ∈ X and μ ∈ ℙ(X), and consider the corresponding CPCA problems (3.2) and (3.3). Given {H_n}, natural finite dimensional approximations of the CPCA problems can be formulated as follows. Let X_n = Π_n(X) which is a compact convex subset of H_n. Let x_n = Π_n x ∈ X_n and let μ_n = (Π_n)#μ ∈ ℙ(X_n) be the pushforward of μ under Π_n. By definition, μ_n is defined by μ_n(A) = μ(Π_n^{-1}(A)), for A ⊂ X_n measurable. It is clear that μ_n → μ. Now we may define the finite dimensional approximations by

\[ \inf_{C \in CL(X)} \{ J(C; μ_n) + χ_{CCxn,k(X_n)}(C) \}, \quad \text{(global)} \]

\[ \inf_{C \in CL(X)} \{ J(C; μ_n) + χ_{\{C \in CCxn,k(X_n): C ⊃ C_{k-1} \}}(C) \}, \quad \text{(nested)} \]

where C_{k-1} ∈ CCπn,k(X_n) is given.

We give a consistency result for the global CPCA which is equivalent to nested CPCA when k = 1. While we expect that analogous results hold for the higher CPCs of the nested problem, the mathematical statements and proofs become much more messy and technical. To focus on the algorithms and applications, we do not pursue consistency further in this paper.

**Theorem 3.4** (Consistency of finite dimensional approximation for global CPCA). Suppose X is compact and Π_n X ⊂ X for each n. For k fixed, let
\[ C_n = CCπ_{n,k}(X_n), \quad C = CCπ,k(X). \]
Then \( \Gamma^- \lim_{n \to \infty} χ_{C_n} = χ_C \) and hence the conclusions of Theorem 3.1 hold.

**Proof.** See Appendix A.2. □

### 3.3. Analytical properties for nested CPCA.
Now we focus on nested CPCA where H is finite dimensional. Under suitable conditions, we establish continuity and differentiability of quantities involved in the optimization. This will allow us to study, in Section 4, numerical algorithms to implement CPCA in practice.

In this subsection we work under the following setting:

**Condition 3.5** (Conditions for finite dimensional nested CPCA).

\( (i) H \) is finite dimensional and so is isomorphic to \( \mathbb{R}^d \) for some \( d \geq 1 \).

\( (ii) \) There exist continuously differentiable convex functions \( g_1, \ldots, g_m : H \to \mathbb{R} \) such that the convex state space \( X \) has the form

\[ X = \bigcap_{j=1}^{m} \{ x \in H : g_j(x) \leq 0 \}. \]

\( (iii) \) The data distribution \( μ \) is an empirical measure of the form \( μ = \frac{1}{N} \sum_{i=1}^{N} δ_{x_i} \).

\( (iv) \) The interior of \( X \) is non-empty and the reference point \( π \in X \) satisfies \( g_j(π) < 0 \) for all \( j \).

The form (3.6) is sufficiently flexible to handle many applications, including those in Sections 3 and 6. For example, if each \( g_j \) is affine then \( X \) is polyhedral, and if each \( g_j \) is linear then \( X \) is a polyhedral cone. Note that here we do not require \( X \) to be compact.
To formulate the main results, we use the observation (see \[6, Proposition 3.4\] and (2.5)) that the nested CPCA problem can be solved by constructing a sequence of orthonormal vectors. For $p \in H$, we let

\[
V(p) = J((x + \text{span}(p)) \cap X; \mu) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \inf_{z_i \in (x + \text{span}(p)) \cap X} \|x_i - z_i\|^2 \right\}.
\]

Define

\[
p^*_1 \in \arg \min_{p \in H : \|p\| = 1} V(p),
\]

and, for $k \geq 2$,

\[
p^*_k \in \arg \min_{p \in H : p \in \text{span}(p^*_1, \ldots, p^*_{k-1})^\perp, \|p\| = 1} V(p).
\]

By Theorem 3.6 below, $V(\cdot)$ is continuous on $H \setminus \{0\}$. This guarantees the existence of $p^*_k$ (as long as dim $H \geq k$). By \[6\] Proposition 3.4, the sets

\[
C_k = (x + \text{span}(p^*_1, \ldots, p^*_k)) \cap X \in CC(x,k)(X)
\]

form a sequence of nested convex principal components.

Now we establish some analytical properties of the function $V(\cdot)$.

**Theorem 3.6 (Continuity).** Under Condition 3.5, the function $V$ defined by (3.7) is continuous on $H \setminus \{0\}$. Moreover, each inner minimization problem in (3.7) has a unique optimizer $z^*_i = z^*_i(p) = x + a^*_i(p)p$ which is continuous in $p \in H \setminus \{0\}$.

Using a recent generalized envelope theorem proved in \[26\] and recalled in Appendix C.3, we also study the differentiability of $V$.

**Theorem 3.7 (Differentiability).** Under Condition 3.5, the directional derivative

\[
V'(p; \tilde{p}) = \lim_{t \to 0^+} \frac{1}{t}(V(p + t\tilde{p}) - V(p))
\]

of $V$ exists for any $p \in H \setminus \{0\}$ and $\tilde{p} \in H$. Moreover, if $p \neq 0$ and for each $i$ at most one constraint is binding (i.e., $g_j(z^*_i) = 0$ for at most one $j$), then $V$ is differentiable at $p$.

**Proof.** See Appendix A.2.6 where we also provide expressions of the directional directive and gradient of $V$ in terms of a Lagrangian and its saddle points. \(\square\)

**Remark 3.8.** In particular, if the Hausdorff dimension of $\bigcup \{g_i = 0\} \cap \{g_j = 0\}$ is at most $d - 2$, then $V$ is differentiable almost everywhere on $H \setminus \{0\}$.

4. Implementation for polyhedral domains

In this section we provide an efficient implementation of nested convex PCA when the Hilbert space is finite dimensional and the convex domain is polyhedral. When combined with parallel computing techniques our algorithm can be applied to fairly large and high dimensional data-sets. Illustrative empirical examples, which are of independent financial interest, are presented in Sections 5 and 6.

As in Section 3 we let the Hilbert space $H$ be finite dimensional, so $H$ is isomorphic to a Euclidean space. To simplify the notations we assume without loss

---

\[3\] Our implementation, using R and C++, can be found at: [https://github.com/stevenacampbell/ConvexPCA](https://github.com/stevenacampbell/ConvexPCA)
of generality that $H = \mathbb{R}^d$ for some $d \geq 2$ (the case $d = 1$ is trivial). Elements of $\mathbb{R}^d$ are regarded as column vectors. We say that a closed convex set $X \subset \mathbb{R}^d$ is polyhedral if there exists an $m \times d$ matrix $A$ and $b \in \mathbb{R}^m$ such that

$$X = \{ x \in \mathbb{R}^d : Ax \geq b \}. \tag{4.1}$$

If $b = 0$, we say that $X$ is a polyhedral convex cone. We remark that if $X \subset \mathbb{R}^d$ is a compact convex set with non-empty interior, then it can be approximated arbitrarily well by convex polyhedrons under the Hausdorff metric (see [7, 33]).

Let $X$ be a convex polyhedron of the form (4.1) and consider, for a given reference point $\bar{x} \in X$ and empirical distribution $\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{z_i}$, the nested CPCA problem which can be stated in the form (see (3.8) and (3.9))

$$p \in H : \| p \| \leq 1, \quad \arg \min_{p \in H} V(p), \quad p_k^* \in \arg \min_{p \in H : p \in P_{k-1}^d, \| p \| = 1} V(p), \tag{4.2}$$

where $P_{k-1} = \text{span}\{p_1^*, \ldots, p_{k-1}^*\}$ and

$$V(p) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \inf_{z_i \in (\bar{x} + \text{span}(p)) \cap X} \| x_i - z_i \|^2 \right\}, \tag{3.7}$$

as given by (3.7). We assume that Condition 3.5 holds.

We begin by parameterizing the quantities involved. Write $z_i = \bar{x} + a_i p$ where $a_i \in \mathbb{R}$ is chosen such that $z_i \in X$. Suppose $p_1^*, \ldots, p_{k-1}^*$ have been found and consider the second minimization problem in (4.2). Given $P_{k-1}$, let $B_k = [b_1 \cdots b_{d-k+1}]$ where $\{b_1, \ldots, b_{d-k+1}\}$ is an orthonormal basis of $P_{k-1}^\perp$. Then $p \in P_{k-1}$ with $\| p \| = 1$ if and only if $p$ has the form $p = B_k w$ where $w$ is a unit vector in $\mathbb{R}^{d-k+1}$. Thus (4.2) is equivalent to solving

$$w_k^* = \arg \min_{w : \| w \| = 1} \hat{V}(w) = \arg \min_{w : \| w \| = 1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \inf_{a_i \in \mathbb{R}, z_i \in X} \| x_i - (\bar{x} + a_i B_k w) \|^2 \right\}, \tag{3.7}$$

where the constraint $z_i \in X$ can be written in terms of $a_i$ and $w$ as

$$A(\bar{x} + a_i B_k w) \geq b. \tag{4.4}$$

When $k = 1$, we may simply let $B_1 = I$ and the equivalence still holds. See Figure 2 for an illustration. Finally, we express the constraint $\| w \| = 1$ using hyperspherical
coordinates. In particular, write \( \mathbf{w} = \omega(\mathbf{\theta}) \) where \( \mathbf{\theta} = (\theta_1, \ldots, \theta_{d-k}) \) and
\[
\omega(\mathbf{\theta}) = \begin{bmatrix}
\sin(\theta_1) \\
\sin(\theta_1) \cos(\theta_2) \\
\sin(\theta_1) \sin(\theta_2) \cos(\theta_3) \\
\vdots \\
\sin(\theta_1) \cdots \sin(\theta_{d-k-1}) \cos(\theta_{n-k}) \\
\sin(\theta_1) \cdots \sin(\theta_{d-k-1}) \sin(\theta_{n-k})
\end{bmatrix}.
\]
Now (4.3) is an unconstrained optimization problem in \( \mathbf{\theta} \in \mathbb{R}^{d-k} \).

We proceed to discuss the algorithm, for which more details are given in Appendix B. For each \( i \), the inner minimization problem in (4.2) is easy to solve since \( L = (\bar{x} + \text{span}\{p\}) \cap X \) can be identified with an interval. Specifically, we first compute the usual orthogonal projection, say \( \tilde{z}_i^* \), of \( x_i \) onto \( L \). If \( \tilde{z}_i^* \) is in \( X \), then \( z_i^* = \tilde{z}_i^* \) is the optimal point. Otherwise, \( z_i^* \) is the boundary point of \( L \) which is closest to \( \tilde{z}_i^* \). This allows us to evaluate the objective \( \tilde{V}(\omega(\mathbf{\theta})) \). Algorithm B.1 gives the procedure for finding the boundary points. Using the differentiability of \( V \) (and hence \( \tilde{V} \)) in \( p \) (and hence in \( \mathbf{\theta} \)), we can use standard gradient approximation (e.g. central differences) and gradient-based optimization methods to solve (4.2).

**Example 4.1.** We illustrate our CPCA algorithm with the simulated data set shown in Figure 1. Here \( d = 2 \) and \( X \) is a convex cone indicated by the shaded region. The data is intentionally constrained to lie in a vertical strip so that the first Euclidean PC (dashed purple line) is almost vertical. Note that for data points near the lower left corner the Euclidean projections are outside \( X \). As shown in the figure, the first convex PC (solid blue line) is quite different from the Euclidean PC. Now the data points near the lower left corner are projected to the endpoint of the convex PC. We also compute the proportion of explained variation (Definition 2.3), which is 98% for the convex PC. This is slightly lower than 99% which is the variation explained by the Euclidean PC.

## 5. Wasserstein geodesic PCA

*Distributional data* arises when each data point can be regarded as a probability distribution. For instance, age distributions of countries, and empirical return distributions of financial assets, can be regarded as distributional data-sets. The Wasserstein metric (3.1) in optimal transport [35, 36] provides a natural geometric structure on spaces of probability distributions. As mentioned in Section 1, convex PCA was first introduced in [6] to study Wasserstein geodesic PCA on the real line. This is because the Wasserstein space \( (\mathcal{P}_2(\Omega), W_2) \), where \( \Omega \subset \mathbb{R} \) is an interval and \( \mathcal{P}_2(\Omega) \) is the space of Borel probability measures on \( \Omega \) with finite second moment, is isometric to a closed convex subset of an \( L^2 \) space. After recalling the isometry and the relevant notations, we introduce a finite dimensional approximation for which the algorithms developed in Section 4 apply. We then apply Wasserstein geodesic PCA to analyze return distributions of stocks ordered by size.

### 5.1. Wasserstein space on an interval and geodesic PCA.

For technical purposes, we let \( \Omega \subset [a, b] \) be a compact interval and leave the unbounded case for future research⁴. In this case \( \mathcal{P}_2(\Omega) = \mathcal{P}(\Omega) \) but we will still write the former

---

⁴At a practical level, if the data set consists of compactly supported distributions (say empirical distributions), we can always choose a sufficiently large interval.
since we are using the 2-Wasserstein metric. Let \( \nu \in \mathcal{P}_2(\Omega) \) be a reference measure which is absolutely continuous with respect to the Lebesgue measure and has full support. A simple but useful example is the uniform distribution on \( \Omega \). For any \( \mu \in \mathcal{P}_2(\Omega) \), there exists a non-decreasing mapping \( T_\mu : \Omega \to \Omega \), which is unique \( \mu \)-a.e., such that \( (T_\mu)_\# \nu = \mu \). Then, it is a well-known result in optimal transport that the Wasserstein distance between \( \mu_0, \mu_1 \in \mathcal{P}_2(\Omega) \) is given by

\[
W_2^2(\mu_0, \mu_1) = \int_{\Omega} |T_{\mu_0}(x) - T_{\mu_1}(x)|^2 d\nu(x).
\]

The mapping \( \mu \mapsto T_\mu \) defines a bijection from \( \mathcal{P}_2(\Omega) \) onto the space \( \mathcal{T}_\nu(\Omega) = \{ T : \Omega \to \Omega : T \text{ is } \nu \text{-a.e. non-decreasing} \} \), which can be shown to be a compact convex subset of the separable Hilbert space \( \mathcal{H} = L^2_\nu(\Omega) \). By (5.1), \( \mathcal{P}_2(\Omega) \) is isometric to the vector field \( \mathcal{T}_\nu(\Omega) \) via this mapping. (In [6] the isometry is defined in terms of \( v_\nu(x) = T_\nu(x) - x \) which is also in \( L^2_\nu(\Omega) \).)

We then define Wasserstein geodesic PCA on \( \mathcal{P}_2(\Omega) \) to be the convex PCA on \( \mathcal{X} = \mathcal{T}_\nu(\Omega) \subset \mathcal{H} = L^2_\nu(\Omega) \). Note that a line segment in \( \mathcal{X} \) corresponds via the isometry to a geodesic in the Wasserstein space \( \mathcal{P}_2(\Omega) \) which is given in terms of McCann’s displacement interpolation [27]. In particular, the geometry does not depend on the choice of the absolutely continuous reference measure \( \nu \).

5.2. Finite dimensional approximation. Consider a collection \( \mu_1, \ldots, \mu_N \in \mathcal{P}_2(\Omega) \) of distributional data and a reference point \( \mu \in \mathcal{P}_2(\Omega) \) (corresponding to \( x \) in the general formulation). Let \( T_1, \ldots, T_N, T \) be the corresponding elements in \( X = \mathcal{T}_\nu(\Omega) \subset \mathcal{H} = L^2_\nu(\Omega) \). Note that a line segment in \( X \) corresponds via the isometry to a geodesic in the Wasserstein space \( \mathcal{P}_2(\Omega) \) which is given in terms of McCann’s displacement interpolation [27]. In particular, the geometry does not depend on the choice of the absolutely continuous reference measure \( \nu \).

The idea is simple: we approximate the interval \( \Omega = [a, b] \) using a refining sequence of partitions. For expositional simplicity and concreteness we focus on the dyadic partition \( \mathcal{D}_n = \{ a + (b-a)j2^{-n} \}_{j=0}^{2^n} \), but other partitions may be used for computational purposes. For \( n \geq 1 \), let \( \mathcal{H}_n \subset \mathcal{H} \) be the finite dimensional subspace consisting of functions \( T \in \mathcal{H} \) such that \( T \) is \( \nu \)-a.e. constant on each subinterval corresponding to \( \mathcal{D}_n \). Clearly \( \mathcal{H} = \bigcup_n \mathcal{H}_n \) by a standard density argument. Let \( \Pi_n : \mathcal{H} \to \mathcal{H}_n \) be the orthogonal projection onto \( \mathcal{H}_n \). By projecting \( T_i \) and \( T \) onto \( \mathcal{H}_n \) (see (5.3) below), we get along the lines of Section 3.2 a finite dimensional approximation of Wasserstein geodesic PCA. In particular, since \( T_i \) (which corresponds to the transport map from \( \nu \) to \( \mu_i \)) is piecewise constant, we approximate each \( \mu_i \) by a discrete distribution supported on a set with at most \( 2^n \) elements. The following lemma shows that our approximation satisfies the condition \( \Pi_n X \subset X \) and hence the consistency result in Theorem 3.4 applies.

**Lemma 5.1.** For \( n \geq 1 \) we have \( \Pi_n X \subset X \).

**Proof.** Given \( n \), consider the orthogonal collection \( (\varphi_{n,j})_{j=1}^{2^n} \subset \mathcal{H}_n \) given by

\[
\varphi_{n,j} = 1_{I_j},
\]
where $I_j = (a + (b - a)(j - 1)2^{-n}, a + (b - a)2^{-n}]$ is the $j$-th subinterval of $D_n$, and $1_{I_j}$ is 1 on $I_j$ and 0 elsewhere. Then for $T \in H$ we have
\begin{equation}
\Pi_n T = \sum_{j=1}^{2^n} \frac{(T, \varphi_{n,j})_{L^2(\Omega)}}{\|\varphi_{n,j}\|_{L^2(\Omega)}} \varphi_{n,j} = \sum_{j=1}^{2^n} \left( \frac{1}{\nu(I_j)} \int_{I_j} T d\nu \right) 1_{I_j}.
\end{equation}

Clearly, if $T \in X$ then $\Pi_n T$ is non-decreasing and maps $\Omega$ into $\Omega$. So $\Pi_n T \in X$ and $X$ is invariant under $\Pi_n$. □

Each $T \in H_n$ (has a version which) can be written in the form
\begin{equation}
T = \sum_{j=1}^{2^n} t_j 1_{I_j},
\end{equation}
for some $t = (t_1, \ldots, t_{2^n}) \in \mathbb{R}^{2^n}$. Then $T \in \Pi_n X$ if and only if
\begin{equation}
a \leq t_0 \leq t_1 \leq \cdots \leq t_{2^n} \leq b.
\end{equation}

Thus $X_n = \Pi_n X$ can be identified with a convex polyhedral domain \{ $t \in \mathbb{R}^{2^n}$ : $At \geq b$\} for appropriately defined $A$ and $b$, and hence our algorithm developed in Section 4 applies. From (5.5), the matrix $A$ is bidiagonal, a feature which simplifies some of the computations. We remark that if each data point $\mu_i$ (more precisely each $T_i = T_{\mu_i}$) is already in $H_n$, then our formulation is exact.

5.3. **Ranked stock return distributions.** In both empirical examples (here and in Section 6) of this paper we consider the US stock market using data provided by The Center for Research in Security Prices (CRSP) [10]. In this example, we consider daily market capitalizations and dividend-adjusted stock returns from January 1962 to December 2021.

---

Figure 3. Quantiles of the return distributions for each rank.
On each trading day, we consider the largest $N = 1000$ stocks (which differ from day to day) and consider their daily (arithmetic) returns. We arrange the returns according to the initial ranks of the stocks. Repeating this for each day, we obtain, for each rank $k = 1, \ldots, N$, an empirical return distribution $\mu_k \in \mathcal{P}(\mathbb{R})$. More precisely, we have $\mu_k = \frac{1}{T} \sum_{t=1}^{T} \delta_{R_{kt}}$ where $R_{kt}$ is the return of the stock which was at rank $k$ on day $t$. In stochastic portfolio theory it was observed that rank-based properties of stocks are useful in portfolio selection. For example, parameters (e.g. growth rates) of individual stocks are difficult to estimate but the corresponding rank-based quantities are more often much more stable. This phenomenon has inspired a large literature of rank-based diffusion models where the dynamics of a stock depends on its current rank (see e.g. [22] and the references therein). Understanding of the rank-based return distributions is relevant in constructing realistic rank-based models of the stock market. The data-set (consisting of $N = 1000$ empirical return distributions) is visualized in Figure 3. Most importantly, we observe that small stocks tend to be more volatile and their return distributions are more positively skewed.

We perform nested Wasserstein geodesic PCA to analyze this data-set. We choose a sufficiently large interval $[a, b]$ which covers all returns, and use a dyadic partition ($\mathcal{P}_n$ with $n = 10$) to obtain a finite dimensional nested CPCA problem as described in Section 2.2. We let the reference distribution $\nu$ be the uniform distribution on $[a, b]$, and let $\overline{\nu}$ be the (Wasserstein) barycenter of the return distributions. In Figure 4 we visualize the first two convex PCs (principal directions) which we also call the Wasserstein GPCs. In terms of the general notations of CPCA in Section 2 we visualize the curve $t \mapsto \overline{\nu} + t\mathbf{p}_k^* \in X$ for $t \in [-\epsilon, \epsilon]$ and $k = 1, 2$. We also call this the perturbation of $\overline{\nu}$ along $\mathbf{p}_k^*$. The first principal direction $\mathbf{p}_1^*$ captures the volatility and the overall skewness of the distribution, and explains alone 89.9% of the variation (in the sense of Definition 2.3). The second principal direction $\mathbf{p}_2^*$ can be interpreted in terms of additional skewness effects on the left and right tails of the distribution (note that the middle quantiles are almost unaffected), and contributes to a small increase of about 3.5% in the explained variation. The projection of the data-set onto $C_2^* = (\overline{\nu} + \text{span}\{\mathbf{p}_1^*, \mathbf{p}_2^*\}) \cap X$ (identified with a subset of $\mathbb{R}^2$) is shown in Figure 5(left).

Note that in Figure 4 each quantile moves at constant velocity as the perturbation parameter $t$ varies. This is a feature of McCann’s displacement interpolation, under which each “particle” travels along a constant velocity straight line. More precisely, for each $k$ and $t$, the perturbed distribution (corresponding to $\overline{\nu} + t\mathbf{p}_k^*$) has the form $(\text{Id} + tv_k)_{\#}\overline{\nu}$ for some vector field $v_k$ on the interval. In Figure 5(right) we plot the initial velocity fields $v_1$ and $v_2$ as functions of the return. Consistent with above, we see that $v_1$ is an asymmetric size effect and $v_2$ acts mostly on the tails.

6. CAPITAL DISTRIBUTION CURVE UNDER THE AITCHISON GEOMETRY

6.1. Financial background. Our second empirical example is about the capital distribution curve of the US stock market. We begin by defining this quantity and explaining its financial relevance. Let $m_i(t) > 0$ be the market capitalization of

---

Footnotes:

6. Returns correspond to the ret variable in the CRSP database which is adjusted for stock splits and dividends.

7. We show here quantiles for $q$ and $1 - q$ in $\{.5, .55, .6, .65, .7, .75, .775, .8, .825, .875, .9, .925, .94, .955, .965, .975, .98, .985, .99, .994\}$. These values are chosen for visualization purposes.
stock \( i \) at time \( t \). The market weight of stock \( i \) (relative to a given universe with \( n \) stocks) is given by

\[
\mu_i(t) = \frac{m_i(t)}{m_1(t) + \cdots + m_n(t)}, \quad i = 1, \ldots, n.
\]

The vector \( \mu(t) = (\mu_1(t), \ldots, \mu_n(t)) \) takes values in the open unit simplex

\[
\Delta_n = \{ \mathbf{p} = (p_1, \ldots, p_n) \in (0, 1)^n : p_1 + \cdots + p_n = 1 \}.
\]

The capital distribution at time \( t \) is given by the vector of ordered market weights given by

\[
\mu(\cdot)(t) = (\mu(1)(t), \ldots, \mu(n)(t)),
\]
where \( \mu_1(t) \geq \cdots \geq \mu_n(t) \) are the (decreasing) order statistics of the market weights. Note that \( \mu_k(t) \) takes values in the ordered unit simplex \( \Delta_n, \geq = \{ \mathbf{p} \in \Delta_n : p_1 \geq \cdots \geq p_n \} \). In Figure 6 (left) we plot the distribution curves (in log-log scale, i.e., \( \log k \mapsto \log \mu_k(t) \)) of the whole US universe, where one curve is drawn for (early January of) each year from 1962 to 2021. A general observation is that the shape of the capital distribution is remarkably stable over time; capturing this stability is one of the main goals of the rank-based diffusion models mentioned in Section 5.3. Nevertheless, fluctuations of the capital distribution curve have significant financial implications. In particular, consider what is called in stochastic portfolio theory the market diversity which refers to the concentration of market capital. For example, we say that market diversity is low when most market capital is highly concentrated in the largest stocks, and market diversity is high if market capital is distributed more equally among the stocks. One way to measure market diversity is to use the function

\[
D_\lambda(\mathbf{p}) = \left( \sum_{i=1}^{n} p_i^\lambda \right)^{1/\lambda}, \quad \mathbf{p} \in \Delta_n,
\]

where \( \lambda \in (0, 1) \) is a tuning parameter. It was shown empirically, see e.g. [2, 19, 20], that the change in market diversity correlates positively with the performance of active portfolio managers relative to the capitalization-weighted market portfolio. Also see [3] for recent empirical studies using the CRSP universe.

As seen in Figure 6 (left), the number of stocks varies over time. To keep the dimension fixed, we consider in this section the daily capital distribution curves computed using the largest \( n = 1000 \) stocks (which vary each day) from January 1973 to December 2021. These curves are plotted in Figure 6 (right). In December

\[\text{Figure 6. Capital distribution curve of the CRSP universe from 1962 to 2021 (left) and restricted capital distribution curve of the largest 1000 stocks from 1973 to 2021 (right). The curves are coloured according to their chronological order with the most recent curves red, and the oldest curves yellow.}\]
1972, data from the Nasdaq Stock Exchange was added to the CRSP universe and caused a material change in the capital distribution. Hence for consistency we restrict to dates after this addition. On average, over the chosen period the top 1000 stocks capture more than 85% of the total market capitalization. While each $\mu(t)$ may be regarded as a probability distribution on the finite set $X = \{1, \ldots, n\}$, the set $X$ has no natural metric structure and so the Wasserstein metric may not be appropriate. Rather, we endow the simplex $\Delta_n$ with the Aitchison geometry in compositional data analysis [17], and use it to study fluctuations of the capital distribution curve using convex PCA. For our data, we expect that this geometry is more appropriate than the Euclidean one since it respects the relative (rather than absolute) scale of the components in the simplex and clearly distinguishes the boundary of the simplex from the interior.

6.2. Aitchison geometry. In this subsection we review the Aitchison geometry under which $\Delta_n$ is an $(n-1)$-dimensional Hilbert space. More details can be found in [17].

For $x \in (0, \infty)^n$, define its closure $C[x] \in \Delta_n$ by

$$C[x] = \left( \frac{x_1}{\sum_i x_i}, \ldots, \frac{x_n}{\sum_i x_i} \right).$$

Now define the following operations on $\Delta_n$:

(i) (Perturbation) For $p, q \in \Delta_n$, define $p \oplus q = C[(p_1 q_1, \ldots, p_n q_n)]$.

(ii) (Powering) For $p \in \Delta_n$ and $\alpha \in \mathbb{R}$, define $\alpha \odot p = C[(p_1^\alpha, \ldots, p_n^\alpha)]$.

Under these operations, $\Delta_n$ is an $(n-1)$-dimensional real vector space. The identity element is the barycenter $e = \left( \frac{1}{n}, \ldots, \frac{1}{n} \right)$, and the additive inverse of $p$ is $p^{-1} = \ominus p := C[(p_1^{-1}, \ldots, p_n^{-1})]$. Next define the Aitchison inner product by

$$\langle p, q \rangle_A = \frac{1}{2n} \sum_{i,j=1}^n \log \frac{p_i}{p_j} \log \frac{q_i}{q_j}.$$

With this $\Delta_n$ is a Hilbert space. It is isometric with $\mathbb{R}^{n-1}$ via the isometric log ratio transform given by

$$\text{ilr}(p) = \left( \sqrt{\frac{i}{i+1}} \log \left( \frac{p_1 \ldots p_i}{p_{i+1}} \right) \right)_{1 \leq i \leq n-1},$$

which has the form $\text{ilr}(p) = ((p, e_i)_A)_{1 \leq i \leq n-1}$ for a specific orthonormal basis.

The relevance of convex PCA stems from the following observation.

Lemma 6.1. Under the Aitchison geometry, the ordered simplex $\Delta_{n, \geq}$ is a polyhedral convex cone. In particular, we have $\text{ilr}(\Delta_{n, \geq}) = \{ x \in \mathbb{R}^{n-1} : Ax \geq 0 \}$, where

$$A = (a_{ij})_{i,j=1}^{n-1} = \left( \sqrt{\frac{j+1}{j}} \delta_{i,j} - \sqrt{\frac{j}{j+1}} \delta_{i,j+1} \right)_{i,j=1}^{n-1},$$

and $\delta_{i,j}$ is the Kronecker delta.

9If we endow $X$ with the structure of a weighted graph, we may consider optimal transport along the lines of [25]. But CPCA cannot be applied, at least directly, to the resulting Wasserstein-like space which can be shown to be Riemannian.
Proof. Omitted. Note that $A$ is bidiagonal which is advantageous from the computational point of view. \qed

6.3. Convex principal components and market diversity. To perform nested CPCA for the capital distribution curves, we first map the data to $\mathbb{R}^{n-1}$ via the isometric log ratio transform (6.6). After performing CPCA on the resulting polyhedral convex cone (see Lemma 6.1), we can interpret the result on the (ordered) simplex by applying the inverse transformation.

In Figure 7 we illustrate the perturbations about the average capital distribution curve (under the Aitchison geometry) along each of the first two convex principal directions. The proportion of explained variation is 62.5% for the first principal direction and 78.8% for the first two. Perturbations along the first principal direction correspond to tilting of the curve about a point near the center (in log-log scale) while fixing for the most part the largest stock. On the other hand, perturbations in the second principal direction may be interpreted in terms of (correlated) fluctuations of the largest few stocks relative to the rest of the universe.

Remarkably, the first principal direction is closely related to the change in market diversity. In Figures 8 and 9 (left) we compare (i) the projected coordinates of the capital distribution curves on the first principal direction and (ii) market diversity measured by $D_{\lambda}$ (see (6.4)) with $\lambda = 0.5$. We observe that the two series are very strongly correlated. This finding gives a statistical justification for market diversity and suggests that it is an excellent one-dimensional summary of the capital distribution curve. We also observe that there appears to be a change in the relationship during the COVID health crisis. The relationship in Figure 8 seems to experience a parallel shift and the resulting points appear similar to those observed in the 1970s and 1980s. Turning to the second principal component (shown in Figures 8 and 9 (right)), we observe a fairly significant positive correlation between the projected coordinate and the (relative and combined) weight of the largest...
two stocks. These findings suggest that the capital distribution curve is not as stationary as what Figure 6 and the rank-based diffusion models might suggest.

Taken together, the results of our CPCA suggest that the dominant features of the capital curve are the change in market diversity and the idiosyncratic fluctuations of the largest stocks. In future research, we plan to use these insights to improve existing rank-based models and their calibration.

7. Conclusion and discussion

In this paper we contribute to both the theory and application of convex PCA which was first introduced in [6]. We establish several new theoretical results mostly in the finite dimensional setting, and develop an efficient algorithm for solving large CPCA problems that is amenable to parallelization and complements the approach of [28]. This algorithm can be readily specialized to solve (a finite dimensional approximation of) Wasserstein geodesic PCA on a finite interval. We also provide two interesting financial applications.

We discuss here several directions for future research.

(i) While we provide several consistency results in the finite dimensional case, infinite dimensional analogues of Theorems 3.2 and 3.3 remain open. In particular, we rely crucially on the compactness of the state space $X$ and that it has a nonempty interior. It appears that a different approach is needed for the general case.

(ii) Existing works on Wasserstein geodesic PCA focus on the one-dimensional setting because the optimal transport maps are simply non-decreasing functions. In general, optimal transport maps are gradients of convex functions by Brenier’s theorem, but they are much less tractable when $d \geq 2$. Also,
computation of the optimal transport map is much more expensive in multiple dimensions. Another direction of interest is to formulate probabilistic models for Wasserstein geodesic PCA and related algorithms.

(iii) Our empirical study highlights behaviours of the capital distribution curve and rank-based properties of stocks. Naturally, we hope that these stylized properties are preserved in stochastic models of the equity market. While the statistical properties of individual asset prices have been intensively studied in financial econometrics (see e.g. [14]), it is not easy to combine them to formulate realistic macroscopic models involving hundreds or thousands of stocks. The rank-based models in stochastic portfolio theory, which are based on systems of interacting Itô processes, fail to capture short-term fluctuations that are relevant in portfolio selection, as they rely on the ergodic properties of Markov processes. On the other hand, dynamic factor models (see e.g. [4] and the references therein) capture the (possibly dynamic) correlation structure among asset returns but fail to reproduce macroscopic behaviours such as the stability of the capital distribution curve when compounded over time. The construction of realistic macroscopic models, with applications in portfolio management, is an important direction which we hope to address in future research.
Appendix A. Proofs

A.1. Proofs from Section 2

A.1.1. Proof of Proposition 2.5

Lemma A.1. Suppose $X$ is compact. Let $(C_k)_{k \geq 1}$ be an increasing sequence of non-empty compact sets contained in $X$ and $C = \bigcup_k C_k$. Then $h(C_k, C) \to 0$ where $h$ is the Hausdorff metric.

Proof. By [6 Proposition 3.2]), $(CL(X), h)$ is a compact metric space. Along a fixed subsequence $k'$ there is a further subsequence $k''$ along which $h(C_{k''}, C^*) \to 0$ for some compact $C^* \in CL(X)$. Since $X$ is compact, convergence under Hausdorff metric is equivalent to Kuratowski convergence (see Lemma C.5). By the definition of Kuratowski convergence (Definition C.4), for any $x \in C^*$ there exists a sequence $(x_{k''})$ such that $x_{k''} \in C_{k''}$ for each $k''$ and $x_{k''} \to x$. It follows that $C^* \subseteq \bigcup_k C_k$. On the other hand, it is easy to see that $\bigcup_k C_k \subseteq C^*$. Thus $C^* = \bigcup_k C_k$. Since the subsequence $k'$ is arbitrary, we have that $h(C_k, C) \to 0$.

Proof of Proposition 2.5. By Lemma A.1 we have $h(C_k, X) \to 0$. Since $\Pi_{C_k}(x) \in C_k$ by construction, we have

$$d(\Pi_{C_k}(x), x) \leq h(C_k, X) \to 0, \quad x \in X.$$ 

Hence $\Pi_{C_k} \to \Pi_X = \text{Id}$ pointwise on $X$. By the bounded convergence theorem, we have

$$\text{EV}(C_k) = \frac{1}{TV} \mathbb{E}_\mu [d(\Pi_{C_k} x, \tau)^2] \to \frac{1}{TV} \mathbb{E}_\mu [d(\Pi_X x, \tau)^2] = 1.$$

\hfill \Box

A.2. Proofs from Section 3

A.2.1. Proof of Theorem 3.1

Lemma A.2. Suppose $X$ is compact. If $\mu_n \to \mu$ in $\mathcal{P}(X)$ then $J(\cdot; \mu_n) \to J(\cdot; \mu)$ uniformly on $CL(X)$.

Proof. Note that $d(\cdot, C)$ is 1-Lipschitz for any $C \in CL(X)$. Since $X$ is compact, $d(\cdot, C)$ is uniformly bounded above. Thus, there exists a constant $L > 0$, independent of $C \in CL(X)$, such that $d^2(\cdot, C)$ is $L$-Lipschitz on $X$. By the Kantorovich-Rubenstein duality (see [36 (6.3)]), we have, for any $C \in CL(X)$,

$$|J(C; \mu_n) - J(C; \mu)| = |\mathbb{E}_{\mu_n}[d^2(x, C)] - \mathbb{E}_\mu[d^2(x, C)]| \leq LW_1(\mu_n, \mu).$$ 

This completes the proof since by compactness of $X$ weak convergence is equivalent to convergence in $W_1$.

\hfill \Box

Proof of Theorem 3.1. We proceed analogously to [6 Theorem 6.1]. As $J(\cdot; \mu)$ is continuous, by [15 Remark 4.8] the uniform convergence $J(\cdot; \mu_n) \to J(\cdot; \mu)$ (proved in the lemma above) implies continuous convergence (Definition C.1). Moreover, by [15 Remark 4.9] continuous convergence implies $\Gamma$-convergence. It follows that

$$\Gamma \lim_{n \to \infty} J(\cdot; \mu_n) = J(\cdot; \mu).$$
Now [10] Proposition 6.20] gives, by the continuous convergence of \( J(\cdot, \mu_n) \), the finiteness of \( J(\cdot, \cdot) \) (by the boundedness of \( d(\cdot, \cdot) \)), and the \( \Gamma \)-convergence of the indicators, that
\[
\Gamma\text{-}\lim_{n \to \infty} \{ J(\cdot, \mu_n) + \chi_{\mathcal{E}_n}(\cdot) \} = J(\cdot, \mu) + \chi_{\mathcal{E}}(\cdot).
\]
In view of the compactness of \( C\mathcal{L}(X) \) and Proposition C.3 this completes the proof. \( \square \)

**A.2.2. Proof of Theorem 3.2.** We begin by noting a technical gap in [6] Lemma 6.1.

**Remark A.3 (Technical gap in [6]).** In [6] Lemma 6.1, the authors constructed the translated sets \( C_n := C + x_n - x \) and use them in the proof of \( \Gamma \)-convergence in their Theorem 6.1. However, these sets may not be contained in \( X \). Indeed, for a counterexample it suffices to take \( X \) to be the unit square in \( \mathbb{R}^2 \) and \( C \) to be one of its diagonals. Any translation of \( C \) moves it, at least partially, outside of \( X \).

To overcome this technical issue, we assume that \( H \) is finite dimensional and that \( X \) has a nonempty interior containing the reference element \( \pi \). With these additional assumptions we are able to establish Theorem 3.2 and Theorem 3.3.

The following results are needed in the proof of Theorem 3.2. We say that sets \( A, B \subset H \) are weakly linearly separable if there exists a non-zero linear functional \( f : H \to \mathbb{R} \) and \( c \in \mathbb{R} \) such that \( f(a) \geq c \geq f(b) \) for all \( a \in A \) and \( b \in B \).

**Theorem A.4.** Let \( H \) be finite dimensional and \( K, L \subset H \) be non-empty, compact and convex sets that are not weakly linearly separable. If \( K_i, L_i \subset H \) are also non-empty, compact and convex sets with \( h(K_i, K) \to 0 \), \( h(L_i, L) \to 0 \) as \( i \to \infty \), then \( K_i \cap L_i \neq \emptyset \) for almost all \( i \) and \( h(K_i \cap L_i, K \cap L) \to 0 \) as \( i \to \infty \).

**Proof.** This is a restatement of [12] Theorem 1.8.8] after noting that it holds for any finite dimensional real Hilbert space \( H \). \( \square \)

**Lemma A.5.** If \( H \) is finite dimensional and \( A, B \subset H \) are sets with \( \text{int}(A) \neq \emptyset \), \( B \subset A \) and \( B \cap \text{int}(A) \neq \emptyset \), then \( A \) and \( B \) are not weakly linearly separable.

**Proof.** We proceed by contradiction. If \( A \) and \( B \) are weakly linearly separable, there exists a non-zero linear functional \( f \) and \( c \in \mathbb{R} \) such that \( f(a) \geq c \geq f(b) \) for all \( a \in A \) and \( b \in B \). As \( B \subset A \), we must have \( f(b) = c \) for all \( b \in B \). Choose \( b_0 \in B \cap \text{int}(A) \) and an \( \epsilon \)-ball \( B_\epsilon(b_0) \subset \text{int}(A) \). Since \( f \) is nonzero, there exists \( a \in B_\epsilon(b_0) \) such that \( f(a) < c \). Thus we get a contradiction. \( \square \)

**Proof of Theorem 3.2.** By Lemma C.6 it suffices to show that
\[
CC_{\pi, k}(X) = K\text{-}\lim_{n \to \infty} CC_{\pi_n, k}(X).
\]
We prove this by verifying the properties in Definition C.4

(i) Let \( C \in CC_{\pi, k}(X) \) be given. For \( n \geq 1 \), let \( C'_n = C + \pi_n - \pi \subset H \). Clearly \( h(C, C'_n) \leq \| \pi - \pi_n \| \to 0 \). Define \( C_n = C'_n \cap X \in CC_{\pi_n, k}(X) \). We claim that \( h(C_n, C) \to 0 \). Since \( \pi \in \text{int}(X) \) we have \( C \cap \text{int}(X) \neq \emptyset \) as \( \pi \in C \). By Lemma A.5 \( C \) cannot be (weakly) linearly separated from \( X \). Taking \( L_n = L \subset X \) in Theorem A.4 gives that \( h(C_n, C) = h(C'_n \cap X, C \cap X) \to 0 \).

(ii) The proof of this part is the same as that of [6] Lemma 6.1. We include it for completeness. Let \( C \) be an accumulation point of \( (C_n)_{n \geq 1} \). As \( \pi_n \in C_n \) and
By compactness, there exists a subsequence \( n \) each \( \in B \) for (ii) is analogous. Suppose that \( \dim(\alpha) \) for \( n \geq 1 \), be compact convex sets and suppose there exists a compact set \( C \) such that
\[
A, B, A_n, B_n, \text{conv}(A_n \cup B_n), \text{conv}(A \cup B) \subset C,
\]
for all \( n \geq 1 \). If \( A_n \to A \) and \( B_n \to B \) in Hausdorff distance then \( \text{conv}(A_n \cup B_n) \to \text{conv}(A \cup B) \) in Hausdorff distance.

Proof. (i) Clearly \( A \cup B \) is compact. By [11] Theorem 5.35 \( \text{conv}(A \cup B) = \text{conv}(A \cup B) \). To prove the reverse inclusion, let \( x \in \text{conv}(A \cup B) \) be given. Then there exists a sequence \( x_n \in \text{conv}(A \cup B) \) such that \( x_n \to x \). For each \( n \), there exist \( w_n \in [0, 1] \) and \( a_n \in A, b_n \in B \) such that \( x_n = w_n a_n + (1 - w_n) b_n \). By compactness, there exists a subsequence \( n' \) along which \( w_{n'} \to w \in [0, 1] \), \( a_{n'} \to a \in A \) and \( b_{n'} \to b \in B \). Thus \( x = w a + (1 - w)b \in \text{conv}(A \cup B) \).

(ii) First we note that \( A_n \cup B_n \to A \cup B \) in Hausdorff distance. Since \( C \) is compact, by Lemma C.5 it suffices to show that \( K, \lim_{n \to \infty} \text{conv}(A_n \cup B_n) = \text{conv}(A \cup B) \). We now check the relevant conditions. (i) If \( x \in \text{conv}(A \cup B) \) then \( x = w a + (1 - w)b \) for some \( w \in [0, 1] \), \( a \in A \) and \( b \in B \). By Hausdorff convergence there exists \( a_n \in A_n, b_n \in B_n \) such that \( a_n \to a \) and \( b_n \to b \). Let \( x_n = w a_n + (1 - w)b_n \in \text{conv}(A_n \cup B_n) \). Then \( x_n \to x \). (ii) Suppose \( x_n \in \text{conv}(A_n \cup B_n) \) for \( n \geq 1 \) and \( x \) is an accumulation point of \( (x_n)_{n \geq 1} \). Then there exists a subsequence such that \( x_{n'} \to x \). Write \( x_{n'} = w_{n'} a_{n'} + (1 - w_{n'}) b_{n'} \) where \( w_{n'} \in [0, 1], a_{n'} \in A \) and \( b_{n'} \in B \). Again by compactness there exists a further subsequence such that \( w_{n''} \to w \in [0, 1], a_{n''} \to a \) and \( b_{n''} \to b \). So
\[
x = \lim_{n'' \to \infty} x_{n''} = \lim_{n'' \to \infty} (w_{n''} a_{n''} + (1 - w_{n''}) b_{n''}) = w a + (1 - w)b \in \text{conv}(A \cup B).
\]

\[\square\]

Lemma A.7. Suppose \( \dim(\text{supp}(\mu)) \geq k \). Let \( j \leq k \). If \( C_j^* \) is either (i) a global or (ii) nested \((j, \bar{\pi}, \mu)\)-convex principal component, then \( \dim(C_j^*) = j \).

Proof. We will prove this result by contradiction. We will treat (i) as the result for (ii) is analogous. Suppose that \( \dim(C_j^*) < j \leq k \). Then, there must exist an \( x \in \text{supp}(\mu) \) such that \( x \notin \text{aff}(C_j^*) \), where \( \text{aff}(C_j^*) \) is the affine hull of \( C_j^* \). Additionally, we can take an open ball \( B_e(x) \) of sufficiently small radius \( \varepsilon > 0 \) about \( x \) such that \( B_e(x) \cap C_j^* \neq \emptyset \) and \( \text{dist}(B_e(x), C_j^*) > 2 \varepsilon \). As \( x \in \text{supp}(\mu) \) it follows that \( \mu(B_e(x)) > 0 \). Now consider \( \tilde{C} = \text{conv}(C_j^* \cup \{x\}) \). It is clear that \( \dim(\tilde{C}) = \dim(C_j^*) + 1 \leq j \). Moreover, \( \bar{\pi} \in \tilde{C}, \tilde{C} \subset X \) (as \( \text{supp}(\mu) \subset X \) by
assumption), $\tilde{C}$ is convex, and $\tilde{C}$ is compact (see Lemma A.6). Hence $\tilde{C} \in CC_{x, j}(X)$ and thus $J(C^*_j; \mu) \leq J(\tilde{C}, \mu)$ by definition of $C^*_j$. However, for any $y \in B_\epsilon(x)$

$$d(y, \tilde{C}) = ||\Pi_\tilde{C} y - y|| \leq ||\Pi_C y - x|| + ||x - y|| \leq 2\epsilon < \text{dist}(B_\epsilon(x), C^*_j) \leq d(y, C^*_j).$$

Note here we have used that $||\Pi_\tilde{C} y - x|| \leq ||y - x||$ as $x \in \tilde{C}$. Also, as $C^*_j \subset \tilde{C}$ for any $z \in H$, $d(z, C^*_j) \geq d(z, \tilde{C})$. We conclude that

$$J(C^*_j; \mu) = E_\mu[d(x, C^*_j)^2] > E_\mu[d(x, \tilde{C})^2] = J(\tilde{C}; \mu),$$

which contradicts the optimality of $C^*_j$. \qed

We introduce a linear transformation which will be useful in our argument.

**Definition A.8.** Let $K$ be a compact and convex set in $H$ containing 0. For \(\text{aff}(K)\) the affine span of $K$ and $v \in H$ a vector, we define the linear transformation $T(\cdot; K, v) = \Pi_{\text{aff}(K)}(\cdot) + \Pi_{\text{span}(v)}(\cdot)$. To prepare for the induction step, suppose Theorem 3.3 holds up to some $j - 1 \leq k$. Let $C \in \mathcal{C}_j$ and take $C_{j-1}$ with $C_{j-1} \subset C$ according to the definition of $\mathcal{C}_j$. Similarly, take $C_{n,j-1} \subset CC_{x, j-1}(X)$ as in the statement of the Theorem. Define $K_{n,j-1} = C_{n,j-1} - \pi_n$, $K_{\infty,j-1} = C_{j-1} - \pi$ and $K = C - \pi$. Note that $K_{n,j-1}, K_{\infty,j-1}, K \subset X := X - X$.

It is easy to see from the compactness of $X$ that $X$ is compact.

**Lemma A.9.** If $h(C_{n,j-1}, C_{j-1}) \to 0$ then $h(K_{n,j-1}, K_{\infty,j-1}) \to 0$.

**Proof.** By Lemma C.5 it suffices to check Kuratowski convergence (Definition C.4) as $X$ is compact. The proof is then straightforward. \qed

We will use these sets to define linear transformations using Definition A.8. The required vector $v$ will be constructed as follows. By Lemma A.7 we have $\dim C_{j-1} = j - 1$, and by definition, we have $\dim C_{n,j-1} \leq j - 1$, $\text{dim} C \leq j$. Consider the subspaces of $H$ given by $A_n := \text{aff}(K_{n,j-1})$ and $A_{\infty} := \text{aff}(K_{\infty,j-1})$. Here, and throughout, for two subspaces $A, A'$ of $H$ we will define $A \oplus A'$ to be their direct sum:

$$A \oplus A' = \{ x \in H : x = a + a', a \in A, a' \in A' \}.$$ 

Since $A_{\infty} \subset \text{aff}(K)$ we have that either (i) $A_{\infty} = \text{aff}(K)$, or (ii) as $\text{dim}(K_{\infty,j-1}) = j - 1$, there exists a $b \in \text{aff}(K)$ such that $b$ is orthogonal to $A_{\infty}$ and $\text{aff}(K) = A_{\infty} \oplus \text{span}\{b\}$. In this case, we fix such a $b$ to play the role of the defining vector $v$ in Definition A.8. In the former case (i), we instead take $b = 0$.

With this, we will now finally define

(A.1) \(T_n := T(\cdot; K_{n,j-1}, b) = \Pi_{A_n} + \Pi_{\text{span}\{b\}}\)

and

(A.2) \(T_\infty := T(\cdot; K_{\infty,j-1}, b) = \Pi_{A_{\infty}} + \Pi_{\text{span}\{b\}}\).

Next, it is important when taking the inductive step that the $T_n$ are consistent with $T_\infty$. We show below that $\Pi_{A_n} \to \Pi_{A_{\infty}}$ pointwise, and therefore, $T_n \to T_\infty$ pointwise.
Lemma A.10. If \( h(C_{n,j-1}, C_{j-1}) \to 0 \) we have \( \Pi_{A_n}(x) \to \Pi_{A_\infty}(x) \) for all \( x \in H \).

Proof. We begin by showing the affine subspaces \((A_n)_{n \geq 1}\) converge to \( A_\infty \) in an appropriate sense. We recall that we must have \( K_{n,j-1} \to K_{\infty,j-1} \) in the Kuratowski sense (see Lemmas A.9 and C.5). Let \( j_n := \dim A_n \leq j - 1 \). Consider an arbitrary orthonormal basis \((\phi_{n})^{j_n}_{\ell=1}\) of \( A_n \) and extend it to an orthonormal basis of \( H \), \((\phi_{n}^{(n)})^{d}_{\ell=1} \) (where \( d = \dim H \)). By definition every \( y \in K_{n,j-1} \) can be written as \( y = \sum_{\ell=1}^{d} w_{\ell} \phi_{n}^{(n)} \) for some \( (w_{\ell})_{\ell=1}^{d} \) with \( w_{\ell} = 0 \) if \( \ell > j_n \).

Let \( x \in K_{\infty,j-1} \) be arbitrary. By the Kuratowski convergence (Definition C.4) there exists a sequence \( y_{n} \in K_{n,j-1} \) with \( y_{n} \to x \). By the compactness of the unit sphere we may extract a subsequence of the bases \((\phi_{n}^{(n)})^{d}_{\ell=1}\) such that \( \phi_{\ell}^{(n)} \to \phi_{\ell}^{*} \) for some \( \phi_{\ell}^{*} \) and all \( \ell = 1, \ldots, d \). Write \( y_{n}^{\prime} = \sum_{\ell=1}^{d} w_{\ell}^{\prime} \phi_{\ell}^{(n)} \). By the uniform boundedness of \((K_{n,j-1})_{n \geq 0}, K_{\infty,j-1} \) (as \( K_{n,j-1}, K_{\infty,j-1} \subset \mathcal{X} \)) and the Bolzano-Weierstrass Theorem we may choose a further subsequence so that \( w_{\ell}^{\prime} \to w_{\ell}^{*} \) for some \( w_{\ell}^{*} \in \mathbb{R} \) and all \( \ell = 1, \ldots, d \).

Using the convergence \( y_{n} \to x \) we write

\[
y_{n}^{\prime} = \sum_{\ell=1}^{d} w_{\ell}^{\prime} \phi_{\ell}^{(n)} \to \sum_{\ell=1}^{d} w_{\ell}^{*} \phi_{\ell}^{*} = x
\]

As indices \( \ell > j_{n}^{\prime} \) are such that \( w_{\ell}^{(n)\prime} = 0 \) for all \( n^{\prime} \) we have that we may write \( x = \sum_{\ell=1}^{j_{n}^{\prime}-1} w_{\ell}^{\prime} \phi_{\ell} \) (recall \( j_{n}^{\prime} \leq j - 1 \)). Notice that the convergent subsequence of bases is independent of the choice of \( y_{n} \) (it relates only to \( K_{n,j-1} \)) so the same basis subsequence can be chosen for arbitrary \( x \in K_{\infty,j-1} \). Thus, the representation \( x = \sum_{\ell=1}^{j_{n}^{\prime}-1} w_{\ell}^{\prime} \phi_{\ell} \) for \( (w_{\ell})_{\ell=1}^{j_{n}^{\prime}-1} \) depending on \( x \) holds for arbitrary \( x \in K_{\infty,j-1} \). We conclude that \( A_\infty \subset \operatorname{span}\{\tilde{\phi}_{1}, \ldots, \tilde{\phi}_{j-1}\} \). On the other hand, the orthogonality of \((\phi_{\ell})^{j_{\infty}}_{\ell=1}\) is assured by the limit of an orthonormal sequence. Consequently, we have \( \dim(\operatorname{span}\{\tilde{\phi}_{1}, \ldots, \tilde{\phi}_{j-1}\}) = j - 1 \). As \( \dim A_\infty = j - 1 \) we conclude \( \operatorname{span}\{\tilde{\phi}_{1}, \ldots, \tilde{\phi}_{j-1}\} = A_\infty \).

We will now collect these observations to prove the convergence of the projection. As the choice of basis is immaterial to the definition of the projection, the above tells us that for any subsequence \( (\Pi_{A_{n}^{\prime}}) \) we may find a further subsequence \( (\Pi_{A_{n^{\prime}}}) \) of projections that can be written in terms of some converging orthonormal bases \((\phi_{\ell}^{(n^{\prime})})^{j_{\infty}}_{\ell=1}\) of \( A_{n^{\prime}} \) where \( \dim(A_{n^{\prime}}) = j - 1 \) (we can pick the subsequence so that \( \operatorname{span}\{\phi_{1}^{(n^{\prime})}, \ldots, \phi_{j-1}^{(n^{\prime})}\} = A_{n^{\prime}} \)). Therefore,

\[
\Pi_{A_{n^{\prime}}}(x) \to \Pi_{A_{\infty}}(x), \quad \forall x \in H.
\]

As the subsequence was arbitrary, we are done. \( \square \)

We pause here to briefly remark on the properties of our linear transformations.

Remark A.11 (Properties of \( T_{n} \) and \( T_{\infty} \)). We can notice that

(i) \( T_{\infty} = \Pi_{A_{\infty}} + \Pi_{\operatorname{span}\{b\}} = \Pi_{\operatorname{aff}\{K\}} \),

(ii) \( T_{n} [K_{\infty,j-1}] \subset A_{n} \),

(iii) \( \dim(T_{n}[K]) \leq j \), and;

(iv) \( \dim(T_{n}[K] \cup K_{n,j-1}) \leq \dim(A_{n} \oplus \operatorname{span}\{b\}) \leq j \).

In (i) and (ii) we have used that \( A_{\infty} \perp \operatorname{span}\{b\} \). Before proceeding to our final analysis, we highlight two more important lemmas whose conclusions hold
Lemma A.12. If \((T_n)_{n \geq 1}\) are pointwise bounded continuous linear operators with \(T_n : H \to H\), and \(T_\infty\) is continuous with \(T_n(x) \to T_\infty(x)\) for all \(x \in H\) then \(T_n \to T_\infty\) uniformly on compacts.

Second, we have the following standard result.

Lemma A.13. If \((T_n)_{n \geq 1}\) is such that \(T_n \to T_\infty\) uniformly on \(X \subset H\) and \(x_n \in X\) is such that \(x_n \to x \in X\) then \(T_n(x_n) \to T_\infty(x)\).

Lemma A.14. Let \(j \geq 1\), \(C_{n,j-1} \in CC_{\tau,j-1}(X)\), and \(C_{j-1} \in CC_{\tau,j-1}(X)\) with \(\dim(C_{j-1}) = j - 1\). If \(h(C_{n,j}, C_{j-1}) \to 0\) then, \(\Gamma\)-\(\lim_{n \to \infty} \chi_{C_{n,j}} = \chi_{j}\).

Proof. As with the previous consistency results it suffices to show that \(C_j = K\)-\(\lim_{n \to \infty} \chi_{C_{n,j}}\) (see Lemma C.6). We check the conditions in the definition (Definition C.4).

(i) Suppose \(C \in C_j\). Define \(C''_n := T_n(C - \overline{x}) + \overline{x}_n\) where \(T_n\) and \(T\) are determined through equations (A.1) and (A.2). We see that \(\dim(C''_n) \leq j\) and that the convexity and compactness of \(C''_n\) is inherited by the linearity and continuity of \(T_n\). Next, we will see that \(C''_n \to C\) in Hausdorff distance. By the boundedness (in operator norm) of the family \((T_n)_{n \geq 1}\), the set \(\bigcup_n T_n[X - \overline{x}] + X\) is contained in a (large) bounded set in \(H\). Since \(H\) is finite dimensional, we can further take it to be compact. Thus, to get \(h(C''_n, C) \to 0\) it suffices to show the Kuratowski convergence of \(C''_n\) to \(C\) (see Lemma C.5).

\(i')\) Let \(x \in C\) and define \(x_n = T_n(x - \overline{x}) + \overline{x}_n \in C''_n\). We have, in view of Lemma A.12 and Remark A.11(i), that \(T_n \to \Pi_{aff(C - \overline{x})}\) uniformly on \(x - \overline{x}\). Consequently, as \(\overline{x}_n \to \overline{x}\), the convergence \(x_n \to x\) holds.

\(i'')\) Let \(y_n \in C''_n\) and suppose \(\tilde{y}\) is an accumulation point of \((y_n)_{n \geq 1}\). We have there exists a subsequence such that \(y_{n'} \to \tilde{y}\). By definition of \(C''_n\) there exists an \(x_{n'} \in C\) such that \(y_{n'} = T_{n'}(x_{n'} - \overline{x}) + \overline{x}_n\). By compactness of \(C\) extract a further subsequence such that \(x_{n''} \to \tilde{x} \in C\). Applying the properties of \(T_n\) (particularly Lemma A.13) we have

\[
\|\tilde{y} - \tilde{x}\| = \lim_{n' \to \infty} \|y_{n''} - x_{n''}\| = \lim_{n' \to \infty} \|T_{n'}(x_{n'} - \overline{x}) + \overline{x}_n - x_{n''}\| = \|\tilde{x} - \overline{x} + \overline{x} - \tilde{x}\| = 0.
\]

So, \(\tilde{y} = \tilde{x} \in C\) and we conclude \(h(C''_n, C) \to 0\).

With this established, let \(C_n := \text{conv}(C''_n \cup C_{n,j-1})\). Note that we must have that \(\dim(C''_n) \leq j\) (see Remark A.11(iv) and argue by translation). By Lemma A.6 we have \(C'_n \to C\) in Hausdorff distance as \(C''_n \to C\) and \(C_{n,j-1} \to C_{j-1} \subseteq C\). Here we also have \(C'_n\) is convex and compact (again by Lemma A.6), and \(\overline{x}_n \in C'_n\). It remains to get a set that has the same properties, but is assured to be in \(X\). Define \(C_n = C'_n \cap X\). Applying the same reasoning as in the proof of Theorem 3.2 we can make use of Theorem A.4 to get \(h(C_n, C) \to 0\). As \(C_n \in C_{n,j}\) we are done with the first condition.

(ii) Let \(C_n \in C_{n,j}\) for \(n \geq 0\). Suppose \(C\) is an accumulation point of the \(C_n\). As \(\overline{x}_n \to \overline{x}\) we have \(\overline{x} \in C\) (by Hausdorff convergence). We can conclude \(C \in CC_{\tau,j}(X)\) by compactness since \(C_n \in CC_{\tau,j}(X)\) for all \(n\). Hence, \(C \in CC_{\tau,j}(X)\). Finally, \(C \supseteq C_{j-1}\) by [6] Lemma A.4 and so, \(C \in C_j\). This completes the proof. \(\square\)
We are now ready to prove the theorem.

Proof of Theorem 3.3. Since global and nested CPCA are equivalent when \( k = 1 \), the case \( j = 1 \) is a consequence of Theorem 3.2.

We then argue by induction on \( j \). Suppose (i)-(iii) hold up to some \( j \). Then, along a subsequence \( C_{n', \ell} \to C_\ell \in \text{CC}_{\pi, \ell}(X) \) for each \( \ell \leq j \) and \( C_0 = C_1 \subset \cdots \subset C_j \) form a sequence of nested convex principal components. By Lemma A.7 the convex principal components are full dimensional. Specifically, \( \dim(C_j) = j \) and so all the assumptions of Lemma A.14 are satisfied for \( C_{n', j+1}, \text{CC}_{j+1} \). Thus, (iii) holds for \( j + 1 \) and we can then extract a further subsequence such that (i) and (ii) also hold up to \( j + 1 \). Thus by induction the theorem holds up to \( k \).

A.2.4. Proof of Theorem 3.4

Proof of Theorem 3.4. Note again by Lemma C.6 it suffices to show

\[ \text{CC}_{x, k}(X) = \text{K-}\lim_{n \to \infty} \text{CC}_{x, n, k}(X_n). \]

We will prove the Kuratowski convergence via the definition (Definition C.4).

(i) Let \( C \in \text{CC}_{x, k}(X) \) be arbitrary and let \( C_n = \Pi_n C \). It is easy to verify that \( C_n \in \text{CC}_{x, n, k}(X_n) \). To prove (i) it is sufficient to show that \( h(C_n, C) \to 0 \).

We will appeal to the equivalence of Kuratowski and Hausdorff convergence in the compact metric space \((X, d)\) (see Lemma C.5). First, if \( x \in C \) we have that \( \Pi_n x \in C_n \) is such that \( \|\Pi_n x - x\| \to 0 \). Second, consider \( y_n \in C_n \) for \( n \geq 1 \) and let \( y \) be an accumulation point. There must exist a subsequence \( (y_{n'}) \) such that \( \|y_{n'} - y\| \to 0 \). By definition we have that each \( y_{n'} \) corresponds to at least one \( x_{n'} \in C \) so that \( y_n = \Pi_n x_n \). By compactness of \( C \) there exists a further subsequence \( (x_{n''}) \) and \( x \in C \) such that \( \|x_{n''} - x\| \to 0 \). Consider also \( \Pi_n x \). For all \( \epsilon > 0 \) there exists an \( N \) such that for \( n'' \) sufficiently large we have \( \|\Pi_{n''} x - x\| < \epsilon/3, \|y_{n''} - y\| < \epsilon/3, \) and \( \|x_{n''} - x\| < \epsilon/3 \). Since \( \Pi_n \) is 1-Lipschitz, we have, for \( n'' \) sufficiently large,

\[
\|y - x\| \leq \|y - y_{n''}\| + \|y_{n''} - \Pi_{n''} x\| + \|\Pi_{n''} x - x\| \\
= 2\epsilon + \|\Pi_{n''} x_{n''} - \Pi_{n''} x\| \leq 2\epsilon + \|x_{n''} - x\| < \epsilon.
\]

Since \( \epsilon > 0 \) is arbitrary, we have \( y = x \). It follows that \( h(C_n, C) \to 0 \).

(ii) Let \( C_n \in \text{CC}_{x, n, k}(X_n) \) for \( n \geq 1 \). Suppose that \( C \) is an accumulation point of \( (C_n)_{n \geq 1} \). Then \( h(C_{n'}, C) \to 0 \) along some subsequence. As \( C_{n'} \ni \pi_{n'} = \Pi_{n'} \pi \to \pi \) we have that \( \pi \in C \) (again by appealing to Kuratowski convergence). On the other hand, note that \( \text{CC}_k(X) \) is compact so as \( C_{n'} \in \text{CC}_k(X_n) \subset \text{CC}_k(X) \) we have that \( C \in \text{CC}_k(X) \). We conclude \( C \in \text{CC}_{x, k}(X) \) which completes the proof.

A.2.5. Proof of Theorem 3.6

We present directly the main argument, and the lemmas needed will be proved in the discussion to follow.

Proof of Theorem 3.6. Write \( z_i = \pi + a_i p \in X \) and consider

\[
V(p) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \inf_{a_i \in \mathbb{R}} \inf_{x \in X} \| x_i - \pi - a_i p \|^2 \right\}.
\]

Since there are only finitely many data points, there exists \( r > 0 \) such that the optimal \( z_i \) has norm at most \( \sqrt{r} \) for all \( i \) and independent of \( p \). Without loss of generality, we may introduce an additional convex constraint \( g_{m+1} = \|x\|^2 - r \leq 0 \), and we may choose \( r \) such that \( g_{m+1}(<0 \). Let \( X' = X \cap \{x \in H : g_{m+1}(x) \leq 0 \} \)
be the new convex state space which does not affect the solution. Define the correspondence
\begin{equation}
\Phi(p) = \bigcap_{j=1}^{m+1} \{ a \in \mathbb{R} : g_j(\bar{x} + ap) \leq 0 \} = \{ a \in \mathbb{R} : \bar{x} + ap \in X' \}
\end{equation}
on $H \setminus \{0\}$. We now have
\begin{equation}
V(p) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \inf_{a_i \in \Phi(p)} \| x_i - \bar{x} - a_ip \|^2 \right\}.
\end{equation}

By Lemma A.15 below, $\Phi$ is continuous, non-empty, and compact valued. This, alongside the continuity of $\| x_i - \bar{x} - a_ip \|^2$ allows us to apply Berge’s Maximum Theorem (see Theorem C.10) to get that $\psi_i(p) = \inf_{a_i \in \Phi(p)} \| x_i - \bar{x} - a_ip \|^2$ is continuous on $H \setminus \{0\}$. This implies that $V(\cdot)$, as an average of the continuous functions $\psi_i(\cdot)$, is continuous. Berge’s Maximum Theorem also tells us that that the set of optimizers $a_i^*(p)$ corresponding to each $\psi_i$ is upper-hemicontinuous, non-empty, and compact. By the strict convexity of $\| x_i - \bar{x} - a_ip \|^2$ in $a_i \in \mathbb{R}$ we have a unique optimizer. Hence, we also have that $a_i^*(p)$ is a singleton. By Lemma C.8 this implies that it is continuous on $H \setminus \{0\}$ when viewed as a function.

**Lemma A.15.** The correspondence $\Phi(p)$ from (A.3) is non-empty, continuous, convex, and compact-valued for every $p \in H \setminus \{0\}$ under the assumptions of Theorem 3.6. In particular, it is a compact interval.

**Proof.** Clearly, $\Phi$ is closed and convex as the intersection of the sub-level sets of continuously differentiable convex functions in $a$. It is non-empty since $0 \in \Phi(p)$ by assumption on $\bar{x}$ and $r$. Moreover, by the last constraint $\Phi$ is bounded for each fixed $p \in H \setminus \{0\}$. Thus, it is compact by the Heine-Borel Theorem. Since $\Phi$ is also convex and compact in $\mathbb{R}$ it must define a closed interval. It remains to show that $\Phi$ is continuous as a correspondence. To see this first consider the correspondence:
\begin{equation}
\phi(p) = \{ \{g(p), \bar{a}(p)\} \},
\end{equation}
where $g(p)$ solves for the intersection of the line $\bar{x} + ap$ with the feasible region boundary when $a < 0$. That is
\[ \bar{x} + g(p)p \in \partial \{ x \in H : g_j(x) \leq 0, \ j = 1, ..., m + 1 \} \]
and $g(p) < 0$. Define $\bar{a}(p)$ similarly for the case when $a > 0$. Note that there are always two such points since by assumption the origin of the line, $\bar{x}$, is always in the interior of the feasible region and by the convexity of the region the line must exit (and hence intersect) the region at two points. By Lemma A.18 below the functions $\{g(\cdot), \bar{a}(\cdot)\}$ are continuous in $p \in H \setminus \{0\}$. As a result, by Theorem C.9 $\phi$ is continuous as a correspondence and again by Theorem C.9 since $\Phi$ is the convex hull of $\phi$ in $\mathbb{R}$, $\Phi$ is also continuous as a correspondence. \hfill \Box

To prove the continuity of $\{g(\cdot), \bar{a}(\cdot)\}$ from (A.5) we will need some preliminary lemmas. The first is a technical result that states we can choose a full dimensional convex cone from the origin to intersect a portion of the boundary of a compact convex set with non-empty interior containing the origin. Heuristically, it says that we can shine a “flashlight” from the interior onto the boundary of a convex set.
Lemma A.16 (Flashlight lemma). Let $X$ be a compact, convex set in $H$ for $\dim(H) \geq 2$ with non-empty interior containing the origin. Suppose $x_0 \in \partial X$ and consider a ball of radius $\delta$ about $x_0$, $B_\delta(x_0)$ that does not contain $0$. There exists a convex cone $K = \{ \lambda v : \lambda \in \mathbb{R}_+, \, v \in V \subset H \}$ of rays from the origin with non-empty interior containing $x_0$ and (necessarily) intersecting $B_\delta(x_0)$ such that $\partial X \cap K \subseteq \partial X \cap B_\delta(x_0)$.

Proof. As $\dim(H) \geq 2$ by the homeomorphism $x/\|x\|$ we can restrict our attention to $\partial X = S^{d-1} = \{ x \in H : \|x\| = 1 \}$. Consider $D := \partial B_\delta(x_0) \cap S^{d-1}$ which is a $d-2$ dimensional sphere if $d > 2$, and 2 distinct points otherwise. Then, $\text{conv}(D)$ is an $d-1$ dimensional ball that lives in an affine subspace $A$ of $H$ that does not contain $0$ and satisfies $\dim(A) = d-1$. Let $K = \text{cone}(\text{conv}(D)) = \bigcup_{\lambda \geq 0} \lambda \text{conv}(D)$ and notice that every ray in $K$ intersects $S^{d-1}$ at a point in $B_\delta(x_0) \cap S^{d-1}$. Furthermore, it is clear $K$ is full dimensional (as $\dim(A) = d-1$) and $x_0 \in \text{int}(K)$. \hfill \Box

The second lemma characterizes the (local) solution $a^*_j(\cdot)$ of the equation defining the points where a constraint $g_j(x + ap)$ from $\Phi(p)$ in [A.3] is binding.

Lemma A.17. Let $\bar{g}_j(p, a) = g_j(x + ap)$ and suppose there exists a non-zero solution $a^*$ to the equation

$$\bar{g}_j(p_0, a) = 0$$

at $p_0 \in H$ for some $j \in \{1, ..., m + 1\}$. Then under the assumptions of Theorem 7.6 there exist an open set $U_j$ containing $p_0$ and a unique $C^1$ function $a^*_j : U_k \to \mathbb{R}$ such that $a^*_j(p_0) = a^*$ and

$$\bar{g}_j(p, a^*_j(p)) = 0, \quad p \in U_k$$

In particular, if $a^* > 0$ (resp. $a^* < 0$) there exists an open set $V_j \subset U_k$ containing $p_0$ such that $a^*_j(p) > 0$ (resp. $a^*_j(p) < 0$) on $V_k$.

Proof. We begin by showing that $\frac{\partial \bar{g}_j}{\partial a}(p_0, a^*) \neq 0$ by contradiction. Suppose that $\frac{\partial \bar{g}_j}{\partial a}(p_0, a^*) = 0$. Since $\bar{g}_k$ is convex in the coordinate corresponding to $a$ we have that $\bar{g}_j(p_0, a^*) = 0$ is a minimum of $\bar{g}_j(p_0, a)$. However, by setting $a = 0$ this minimality contradicts our assumption on $x_0$ which says $\bar{g}_j(p_0, 0) < 0$. Hence, we conclude that $\frac{\partial \bar{g}_j}{\partial a}(p_0, a^*) \neq 0$. With this result, since we have assumed the existence of a solution $a^*$ to the equation $\bar{g}_j(p_0, a) = 0$ we can apply the implicit function theorem to get the first result in the statement of the lemma. The second result follows taking $V_j = a^*_j(1(0, \infty))$ as the preimage which is open by continuity and contains $p_0$ by assumption on $a^* = a^*_j(p_0)$. \hfill \Box

With this we can tackle the final result of this section.

Lemma A.18. The functions $\pi(p)$ and $g(p)$ from [A.5] are continuous in $p \in H \setminus \{0\}$.

Proof. It suffices to consider $\pi(p)$. Begin by fixing a vector $p_0 \in H \setminus \{0\}$. The point $x + \pi(p_0)p_0$ is on the boundary of $X' = \bigcap_{j=1}^{m+1} \{ x \in H : g_j(x) \leq 0 \}$ and so by Lemma C.13 this implies there is at least one $g_j$ such that

$$g_j(x + \pi(p_0)p_0) = 0$$

with $\pi(p_0) > 0$. By Lemma A.17 for each $j$ for which this holds there exist an open set $U_j$ containing $p_0$ and unique $C^1$ functions $a^*_j : U_j \to \mathbb{R}_+$ such that
\(a_j^*(p_0) = \pi(p_0)\) and 
\[g_j(\pi + a_j^*(p)p) = 0, \quad p \in U_j.\]

With this established, let \(\mathcal{B} \subset \{1, \ldots, m + 1\}\) be the set of indices corresponding to the binding/active constraints at \(p_0\) and let \(\mathcal{I}\) be the set of inactive constraints at \(p_0\) (note \(\mathcal{I} \cup \mathcal{B} = \{1, \ldots, m + 1\}\)).

For each \(j \in \mathcal{I}\) consider the set
\[I_j = g_j^{-1}(\{0\}).\]

Clearly \(\pi + \pi(p_0)p_0 \in I_k\) by assumption. Since the \(g_k\) are continuous, each \(I_k\) is open. Let:
\[I = \bigcap_{j \in \mathcal{I}} I_j.\]

As the intersection of finitely many non-empty open sets containing \(\pi + \pi(p_0)p_0\), \(I\) is open and non-empty. Moreover, if \(x\) is in \(I\) we have that \(g_j(x) < 0\) for all \(k \in \mathcal{I}\).

Now consider the functions \(h_j : U_j \to H\) for \(j \in \mathcal{B}\) given by
\[h_j(p) = \pi + a_j^*(p)p.\]

By the continuity of \(a_j^*\) on \(U_j\) this is continuous on \(U_j\). Moreover since \(p_0\) is such that \(h_j(p_0) = \pi + \pi(p_0)p_0 \in I\) we have that \(O_j = h_j^{-1}(I)\) is open and non-empty containing \(p_0\). Let
\[O = \bigcap_{j \in \mathcal{B}} O_j.\]

Again, \(O\) is open and contains \(p_0\). In particular, each of the functions \(a_j^*\) for \(j \in \mathcal{B}\) are continuous on \(O\) and the values of \(h_j(p)\) on \(O\) correspond to boundary values of the level sets
\[\{x \in H : g_j(x) \leq 0\}.

Moreover, each boundary value is contained in \(I\) where no other constraint is active.

Choose \(\delta\) sufficiently small so that the ball \(B_{\delta}(\pi + \pi(p_0)p_0)\) is contained in \(I\) and does not contain \(\pi\). Consider now our compact convex feasible set \(X'\) and its boundary \(\partial X'\). Since \(X'\) has non-empty interior by assumption, by translation we can apply Lemma [A.16] to conclude there exists a cone \(K\) of rays with origin \(\pi\) and non-empty interior containing \(\pi + \pi(p_0)p_0\) such that
\[K \cap \partial X' \subset B_{\delta}(\pi + \pi(p_0)p_0) \cap \partial X'.\]

Hence, since our open ball is in \(I\) by assumption and every ray in \(K\) intersects a boundary point (by the origin being \(\pi\) in our convex set), we have that every ray in \(K\) intersects a boundary point of our feasible set in \(I\). Since \(K\) has non-empty interior and \(\pi + \pi(p_0)p_0 \in K\) we can choose \(\eta > 0\) small enough so that \(B_{\eta}(\pi + \pi(p_0)p_0) \subset K\). Translating by \(\pi\) and scaling we obtain a ball of radius \(\eta'\) about \(p_0\) such that every \(p \in B_{\eta'}(p_0)\) defines a ray from \(\pi\) that intersects a boundary point in \(I\). That is, the boundary point corresponding to such a \(p\) can only be associated with the boundary of a level set for some \(g_k\) with \(k \in \mathcal{B}\).

Define \(U = O \cap B_{\eta'}(p_0)\). Clearly \(U\) is open and contains \(p_0\). For every \(p\) in \(U\), the intersection with the boundary of our feasible set \(\pi + \pi(p)p\) is a boundary point of one of the level sets of \(g_j\) for \(j \in \mathcal{B}\). Additionally, the intersection with each level set is given by a continuous function \(a_j^*\) on \(U\). Taken together we have that for each \(p \in U\), \(\pi(p) = a_j^*(p)\) for some \(j \in \mathcal{B}\).
Finally, to see continuity at \( p_0 \) let \( \epsilon > 0 \) and let \( p_1 \in H \setminus \{0\} \). Let \( \theta > 0 \) be such that \( B_\theta(p_0) \subset U \). Since the \( a_j^* \) are continuous on \( U \) for each \( k \) we have that there exists a \( \delta_j > 0 \) \( (\delta_j < \theta) \) such that if \( \|p_1 - p_0\|_2 < \delta_j \) then \( |a_j^*(p_1) - a_j^*(p_0)| < \epsilon \). Let \( \delta^* = \min_{j \in B} \delta_j \). Since for any \( p_1 \in U \) there is a \( j_1 \in B \) active we have that if \( \|p_1 - p_0\|_2 < \delta^* \) then:
\[
|\tilde{\pi}(p_1) - \tilde{\pi}(p_0)| \leq \max_{j \in B} |a_j^*(p_1) - a_j^*(p_0)| < \epsilon
\]
as required since the same \( j_1 \) is also active at \( p_0 \). Hence \( \tilde{\pi}(p) \) must be continuous at \( p_0 \) by definition of \( U \). Now since the choice of \( p_0 \) above was arbitrary in \( H \setminus \{0\} \) we conclude that \( \tilde{\pi} : H \setminus \{0\} \to \mathbb{R}_+ \) is continuous.

\[
\text{Proof of Theorem 3.7}
\]

Proof. As in the proof of Theorem 3.6 we can again restrict to a feasible set contained in a closed ball with sufficiently large radius. Let \( \Phi(p) \) be similarly defined as in (A.3). For each \( i \), let \( J_i(a, p) = \|x_i - \tilde{\pi} - ap\|^2 \).

We look to apply Theorem C.11 to each subproblem in (A.4) for which we must verify the necessary assumptions. Note that as \( H \) is finite dimensional Theorem C.11 (and Corollary C.12) applies to our setting once we verify the required conditions. For (i) \( A = \mathbb{R} \) is trivially convex. For (ii) it is clear that \( J_i \) and \( g_j(\tilde{\pi} + ap) \), \( j = 1, ..., m+1 \) are jointly continuous in \((a, p)\) by assumption. For (iii) and (iv) we showed in Lemma [A.15] that \( \Phi(p) \) is compact valued and continuous for \( p \in H \setminus \{0\} \). For (v) this follows by our assumption on \( \pi \) by setting \( a = 0 \). For (vi) since \( J_i \) and \( g_j(\tilde{\pi} + ap) \) are convex in \( a \) for \( j = 1, ..., m+1 \) and for \( a = 0 \), \( g_j(\tilde{\pi}) < 0 \), \( j = 1, ..., m+1 \) we have that Slater’s condition holds for every \( p \in H \setminus \{0\} \). This then implies that the saddle points of the Lagrangian are non-empty for every \( p \in H \setminus \{0\} \). Finally, (vii) holds since we have assumed the \( g_j \) are continuously differentiable and it is clear this similarly holds true for the \( J_i \). Taken together, we obtain the result of Theorem C.11 for our problem which gives directional differentiability. Finally, to establish the points of differentiability, it suffices to note that the assumptions of Corollary C.12 hold trivially for our problem when zero or one constraint is binding at the solution corresponding to the choice \( p \neq 0 \). In fact, this is the only case where the assumptions hold since our minimization variable \( a \) is one-dimensional. Averaging over the sub-problems indexed by \( i \) gives the result in the statement of the Theorem.

We close by collecting here the form of the directional derivative and gradient for our value function. For the set of saddle points \( \mathcal{A}^*_j(p) \times \Lambda^*_j(p) \subset \mathbb{R} \times \mathbb{R}^{m+1} \) in Theorem C.11 corresponding to each inner problem \( J_i \) we have that \( V \) admits the directional derivative
\[
V'(p; \hat{p}) = \frac{1}{N} \sum_{i=1}^{N} \max_{a \in \mathcal{A}^*_i(p)} \min_{\lambda \in \Lambda^*_i(p)} \left[ \nabla_p J_i(a, p) + \sum_{j=1}^{m+1} a \lambda_j \nabla g_j(\tilde{\pi} + ap) \right] \hat{p},
\]
where \( \nabla_p J_i(a, p) = -2a(x_i - \tilde{\pi} - ap) \). Moreover, when \( V \) is differentiable at some \( p_0 \in H \setminus \{0\} \) we have for the unique choice \((a^*_i(p_0), \lambda^*_i(p_0)) \in \mathcal{A}^*_i(p) \times \Lambda^*_i(p)\) that the gradient takes the form
\[
\nabla_p V(p_0) = \frac{1}{N} \sum_{i=1}^{N} \left[ \nabla_p J_i(a^*_i(p_0), p_0) + \sum_{j=1}^{m+1} a^*_i(p_0) \lambda^*_{i,j}(p_0) \nabla g_j(\tilde{\pi} + a^*_i(p_0)p_0) \right].
\]
APPENDIX B. ALGORITHMIC CONSIDERATIONS

We discuss here in greater detail the algorithm outlined in Section 4. In particular, we explain the evaluation of the value function and its gradient.

B.1. Optimization. We first touch on function evaluation. As mentioned in Section 4, for each $i$, the inner minimization problem in (4.2) is easy to solve since $L = (\pi + \text{span}\{p\}) \cap X$ can be identified with an interval. We first compute the usual orthogonal projection, say $\tilde{z}_i^*$, of $x_i$ onto $L$ and check if $\tilde{z}_i^*$ is in $X$. If so, then $z_i^* = \tilde{z}_i^*$ is the optimal point and otherwise, $z_i^*$ is the boundary point of $L$ which is closest to $\tilde{z}_i^*$. Hence, to evaluate the objective $\tilde{V}(\omega(\theta))$ we only need to be able to (i) verify if $\tilde{z}_i^*$ is in $X$ and (ii) efficiently find the boundary point(s) of $L$ for a given $p(\theta) = B_k \omega(\theta)$. The former is straightforward as it can be easily checked if $A \tilde{z}_i^* \geq b$. For (ii), we give the details of finding the boundary point(s) given by $(z_0(\theta), z_1(\theta)) = (\pi + t_0(\theta)p(\theta), \pi + t_1(\theta)p(\theta))$ in Algorithm B.1.

Algorithm B.1: Find Boundary Point Coefficient(s)

Result: Output boundary point coefficient(s) $t_0(\theta), t_1(\theta)$.

Initialize $t_0(\theta) = t_1(\theta) = \infty$;

Let $(A_i)_{i=1}^m$ be the $i$th rows of $A$ and set for $i = 1, \ldots, m$

\[
\alpha_i = \left((b_i - A_i^T \pi)/A_i^T B_k \omega(\theta)\right) \mathbb{I}_{A_i^T B_k \omega(\theta) \neq 0} + \infty \mathbb{I}_{A_i^T B_k \omega(\theta) = 0}
\]

to be the intersection points (if they exist) of the line $t \mapsto \pi + tp$ with each hyperplane defined by the rows of $A$ and $b$;

for $i = 1, \ldots, m$ do

if $A(\pi + \alpha_i B_k \omega(\theta)) \geq b$ then

if $t_0(\theta) = \infty$ then

\[ t_0(\theta) = \alpha_i; \]

end

if $\alpha_i \neq t_0(\theta)$ and $t_0(\theta) < \infty$ then

\[ t_1(\theta) = \alpha_i; \]

break;

end

end

Return $(t_0(\theta), t_1(\theta))$;

Notice that the boundary points depend only on $\theta$, and therefore are common for all the inner optimization problems in (4.2). Hence, for problems with an extremely large data set, the inner problems can be solved in parallel. On the other hand, since for extremely high dimensional problems the gradient is similarly high dimensional, it is often convenient to parallelize the gradient estimation process (e.g. by parallelizing independent central difference computations). For our purposes we implement the parallelization and function evaluation in C++ and rely on central differences for the gradient estimation. Since function evaluation is cheap using the aforementioned approach, this is a convenient way to recover the gradient.
Once this computational challenge is taken care of we can solve the resulting unconstrained optimization problem via a gradient descent approach or through other conventional solvers. Our implementation makes use of a variant of the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm in R available through the “nloptr” package [23]. Specifically, we take advantage of the “Rcpp” package [16] to integrate our C++ implementation with R and perform the optimizations.

For algorithms of this type a relevant question is the choice of the initial point or “guess” for the solution. This is particularly important for large problems, as if we can be intelligent about our initialization, we can substantially improve the convergence time. For our approach, we choose an initial point for each successive principal component that is suggested through an iterated application of traditional PCA techniques. The explicit choice, and the rationale behind it, will be made precise in the following discussion.

### B.2. Choice of initialization point.

We begin this discussion with an equivalent representation of the convex PCA objective function.

**Proposition B.1** (Equivalent formulations of PCA and CPCA Problems). Let \( j \geq 1 \) and suppose \( P_{j-1} \) is a \( j-1 \) dimensional subspace of \( H \). For

\[
\hat{V}_j(p; X) = \frac{1}{N} \sum_{i=1}^{N} \inf_{\parallel x_i - p \parallel = 1, \; a_i, p \in X} \| \Pi_{P_{j-1}^\perp}(x_i - \pi) - a_i p \|^2
\]

and denoting the CPCA objective as \( V(p; X) \) to make the dependence on \( X \) explicit we have

(i) \( \arg \min_{p \in P_{j-1}^\perp, \; \|p\| = 1} V(p; X) = \arg \min_{p \in P_{j-1}^\perp, \; \|p\| = 1} \hat{V}_j(p; X), \) and;

(ii) If there exists an \( i \in \{1, \ldots, N\} \) such that \( x_i - \pi \notin P_{j-1} \) then

\[
\arg \min_{p \in P_{j-1}^\perp, \; \|p\| = 1} V(p; H) = \arg \min_{p \in H : \|p\| = 1} \hat{V}_j(p; H).
\]

**Proof.** See Section [B.3] below. \( \square \)

Naturally, we are interested in the case where \( P_{j-1} \) is given by the previous \( j-1 \) convex principal directions \( (p_k^*)_{k=0}^{j-1} \) (where \( p_0^* \equiv 0 \)). The first statement says that to solve the nested CPCA problem one may equivalently first project the data onto the orthogonal complement of the subspace spanned by the preceding principal directions and then perform the minimization. The second statement says that the solution to traditional Euclidean PCA \((X = H)\) with full dimensional data can be recovered in the same way, but that we can further simplify to an *unconstrained* minimization in the nested formulation of the problem. That is, we can simply perform a Euclidean PCA on the projected data.

The key takeaway is given by the following corollary.

**Corollary B.2.** Let \( j \geq 1 \) and suppose there exists an \( i \in \{1, \ldots, N\} \) such that \( x_i - \pi \notin P_{j-1} \). If \( p_j^{*,H} \in \arg \min_{p \in H, \|p\| = 1} \hat{V}_j(p; H) \) then

\[
\hat{V}_j(p_j^{*,H}; H) = \inf_{p \in P_{j-1}^\perp, \|p\| = 1} \hat{V}_j(p; H) \leq \inf_{p \in P_{j-1}^\perp, \|p\| = 1} \hat{V}_j(p; X) \leq \hat{V}_j(p_j^{*,H}; X).
\]

**Proof.** The first equality follows from Proposition [B.1](i) (with \( X = H \)) and (ii), and the choice of \( p_j^{*,H} \). The last inequality follows again from Proposition [B.1](ii)
as we get \( p^{*,H}_j \in P_{j-1} \). The middle inequality then follows by adding the constraint that the projections are in \( X \).

This says that the Euclidean principal component can be used to bound the value of the convex PCA problem using the representation in Proposition B.1. It is also clear that if none of the constraints defining \( X \) are binding for all \( i \) then the optimal values and minimizers of the two problems agree \( \hat{V}_j(p^{*,H}_j; H) = \tilde{V}_j(p^{*,H}_j; X) \). Note that this also illustrates that, as with standard Euclidean PCA problems, the minimizer for convex PCA problems need not be unique. Most importantly, as Proposition B.1(i) says the minimizers of \( \hat{V}_j \) and \( V \) in \( P_{j-1} \) coincide, this observation suggests that a Euclidean principal direction for the projected data is a good initial guess for the convex principal direction. Indeed, they will coincide if no constraints are binding and so we expect that \( p^{*,H}_j \) should be “close” to a true minimizer. Moreover, the direction \( p^{*,H}_j \) is cheap to compute as it is the solution to a regular PCA problem. As a result, in our optimization routine we initialize our search at some \( p^{*,H}_j \in \arg\min_{p \in H: \|p\|=1} \hat{V}_j(p; H) \) to further accelerate convergence.

B.3. Proof of Proposition B.1

To verify this we will use the following notation:

\[
X = \bigcap_{j=1}^m \{ x \in H : g_j(x) \leq 0 \}, \quad \overline{X} = \bigcap_{j=1}^m \{ x \in H : g_j(\overline{x} + x) \leq 0 \},
\]

\( S_p = \text{span}\{p\} \) and \( \overline{S_p} = \overline{x} + \text{span}\{p\} \).

Note that it is easy to see that for \( x \in X \) we have

\[
\inf_{z \in S_p \cap X} \| x - z \|^2 = \inf_{z \in S_p \cap \overline{X}} \| x - \overline{x} + z \|^2 = \inf_{a \in \mathbb{R}: \overline{x} + a p \in X} \| x - \overline{x} - a p \|^2.
\]

Lemma B.3 (Iterated projection). We have \( \Pi_{S_p \cap X} = \Pi_{S_p \cap \overline{X}} \Pi_{S_p} \).

Proof. Omitted.

This will be used now to prove the following result which immediately gives us the first claim in our proposition. Namely, we now show that we may project the data while maintaining the set of minimizers.

Lemma B.4. For \( j \geq 1 \) we have

\[
\arg\min_{p \in P_{j-1} : \|p\|=1} \frac{1}{N} \sum_{i=1}^N \inf_{z \in S_p \cap X} \| x_i - \overline{x} + z \|^2 \\
= \arg\min_{p \in P_{j-1} : \|p\|=1} \frac{1}{N} \sum_{i=1}^N \inf_{z \in S_p \cap \overline{X}} \| \Pi_{P_{j-1}^{\perp}} (x_i - \overline{x}) - z \|^2.
\]
Proof. If $p \in P_{j-1}^{\perp}$ we have, for any $i$,
\[
\inf_{z \in S_p \cap X} \|x_i - \pi - z\|^2 \\
= \|x_i - \pi - \Pi_{S_p \cap X}(x_i - \pi)\|^2 \\
= \|\Pi_{P_{j-1}}(x_i - \pi) + \Pi_{P_{j-1}^{\perp}}(x_i - \pi) - \Pi_{S_p \cap X}(x_i - \pi)\|^2 \\
= \|\Pi_{P_{j-1}}(x_i - \pi) - \Pi_{S_p \cap X}(x_i - \pi)\|^2 + \|\Pi_{P_{j-1}}(x_i - \pi)\|^2 \\
= \|\Pi_{P_{j-1}}(x_i - \pi) - \Pi_{S_p \cap X}(x_i - \pi)\|^2 + \|\Pi_{P_{j-1}}(x_i - \pi)\|^2 \\
= \inf_{z \in S_p \cap X} \|\Pi_{P_{j-1}}(x_i - \pi) - z\|^2 + \|\Pi_{P_{j-1}}(x_i - \pi)\|^2.
\]
Note the fourth equality follows by Lemma B.3 as
\[
\Pi_{S_p \cap X}(x) = \Pi_{S_p \cap X}(\Pi_{P_{j-1}}x + \Pi_{P_{j-1}^{\perp}}x) = \Pi_{S_p \cap X}\Pi_{S_p \cap P_{j-1}}x = \Pi_{S_p \cap X}\Pi_{P_{j-1}^{\perp}}x.
\]
Now, since the second term on the right hand side is constant in $p$ the claim in the lemma then follows after averaging over the $x_i$.
\]
We now proceed to present a result whose statement implies the second claim of the proposition.

Lemma B.5. For $j \geq 1$ we have
\[
\inf_{p \in P_{j-1}^{\perp} : \|p\|=1} \frac{1}{N} \sum_{i=1}^{N} \inf_{z \in S_p} \|\Pi_{P_{j-1}}(x_i - \pi) - z\|^2 \\
= \min_{p \in \mathcal{H} : \|p\|=1} \frac{1}{N} \sum_{i=1}^{N} \inf_{z \in S_p} \|\Pi_{P_{j-1}}(x_i - \pi) - z\|^2,
\]
and if $(x_i - \pi) \notin P_{j-1}$ for at least one $i$ then the minimizers coincide.

Proof. It suffices to show the relation “$\leq$” as the inequality “$\geq$” is immediate. We have using the definition of the projection for $\|p\|=1$ that for any $i$:
\[
\inf_{z \in S_p} \|\Pi_{P_{j-1}}(x_i - \pi) - z\|^2 = \|\Pi_{P_{j-1}}(x_i - \pi) - (\Pi_{P_{j-1}}(x_i - \pi), p)p\|^2 =: (*)\]
Now by the orthogonal decomposition for $p$ and linearity of the inner product
\[
\langle \Pi_{P_{j-1}}(x_i - \pi), p \rangle = \langle \Pi_{P_{j-1}}(x_i - \pi), \Pi_{P_{j-1}^{\perp}}p \rangle + \langle \Pi_{P_{j-1}}(x_i - \pi), \Pi_{P_{j-1}}p \rangle = \langle \Pi_{P_{j-1}}(x_i - \pi), \Pi_{P_{j-1}^{\perp}}p \rangle,
\]
where the last equality follows by orthogonality. Hence we have:
\[
(*) = \|\Pi_{P_{j-1}}(x_i - \pi) - (\Pi_{P_{j-1}}(x_i - \pi), \Pi_{P_{j-1}^{\perp}}p)p\|^2 \\
= \|\Pi_{P_{j-1}}(x_i - \pi) - (\Pi_{P_{j-1}}(x_i - \pi), \Pi_{P_{j-1}^{\perp}}p)p\|^2 \\
+ \|\Pi_{P_{j-1}}(x_i - \pi), \Pi_{P_{j-1}^{\perp}}p\|^2 \\
\geq \|\Pi_{P_{j-1}}(x_i - \pi) - (\Pi_{P_{j-1}}(x_i - \pi), \Pi_{P_{j-1}^{\perp}}p)p\|^2 \\
\geq \inf_{z \in S_{P_{j-1}^{\perp}}} \|\Pi_{P_{j-1}}(x_i - \pi) - z\|^2.
\]
Consequently for any \( p \in H \) with \( \|p\| = 1 \) we get:

\[
\frac{1}{N} \sum_{i=1}^{N} \inf_{z \in S_{p_j^{-1}}(i)} \|p_{\|p\|=1} \cdot (x_i - \mathbf{x}) - z\|^2 \leq \frac{1}{N} \sum_{i=1}^{N} \inf_{z \in S_{p_j^{-1}}(i)} \|p_{\|p\|=1} \cdot (x_i - \mathbf{x}) - z\|^2.
\]

If \( p_{\|p\|=1} \neq 0 \) then \( S_{p_j^{-1}}(i) = S_{p_j^{-1}}(i) / \|p_{\|p\|=1} \cdot \| \|p\| = 1 \). Otherwise, \( p \in P_{j-1} \) and equality holds in the above with \( z = 0 \) which is admissible for any \( S_{p_j} \) with \( p_j' \in P_{j-1}^{-1} \). Hence, it is clear that we have

\[
\inf_{p \in P_{j-1}^{-1}, \|p\|=1} \frac{1}{N} \sum_{i=1}^{N} \inf_{z \in S_{p_j^{-1}}(i)} \|p_{\|p\|=1} \cdot (x_i - \mathbf{x}) - z\|^2 \leq \inf_{\|p\|=1} \frac{1}{N} \sum_{i=1}^{N} \inf_{z \in S_{p_j^{-1}}(i)} \|p_{\|p\|=1} \cdot (x_i - \mathbf{x}) - z\|^2,
\]

from which the claimed equality follows.

It remains to see that if \( (x_i - \mathbf{x}) \notin P_{j-1} \) for at least one \( i \) then the minimizers must coincide. It is clear that if \( (x_i - \mathbf{x}) \notin P_{j-1} \) then \( p_{\|p\|=1} \cdot (x_i - \mathbf{x}) \neq 0 \) and so if \( p \in P_{j-1} \) then it is not a minimizer of the RHS as it can be strictly improved upon. Moreover, if \( p \notin P_{j-1}^{-1} \) and \( p \notin P_{j-1} \) then the chain of inequalities derived above is strict so \( p \) cannot be a minimizer of the RHS as \( \|p_{\|p\|=1} \cdot P_{j-1}^{-1} \|^{-1} P_{j-1}^{-1} \cdot p \in P_{j-1} \) is an improvement. We conclude that any minimizer of the RHS must be in \( P_{j-1} \) from which the claim follows.

With these results in hand, we formally summarize the conclusions below to complete the proof of the proposition.

**Proof of Proposition B.4** By a reparameterization of the infimums (see equation B.1 in the definition of \( V \) and \( V_j \), the statement (i) follows immediately from Lemma B.4 and (ii) similarly follows from Lemma B.5.

### Appendix C: Miscellaneous definitions and results

**C.1. Convergence of sets and functions.** Throughout this subsection we let a metric space \((\mathcal{Y}, d)\) be given. For \( y \in \mathcal{Y} \), we let \( \mathcal{N}(y) \) be the set of all open neighborhoods of \( y \). Let \( F_n, n \geq 1 \), and \( F \) be functions from \( \mathcal{Y} \) to \( \mathbb{R} \).

**Definition C.1** (Continuous convergence). [15, Definition 4.7] We say that \( F_n \) continuously converges to \( F \) if for every \( y \in \mathcal{Y} \) and every neighborhood \( V \) of \( F(y) \) there exists an integer \( k \) and \( U \in \mathcal{N}(y) \), such that for \( n \geq k \) and \( y \in U \) we have \( F_n(y) \in V \).

**Definition C.2** (\( \Gamma \)-convergence). [15, Proposition 8.1] We say that \( (F_n)_{n \geq 0} \) \( \Gamma \)-converges to \( F \), denoted \( \Gamma \lim_{n \to \infty} F_n = F \), if the following holds for all \( y \in \mathcal{Y} \).

(i) \( F(y) \leq \liminf_{n \to \infty} F_n(y_n) \) for any \( y_n \in \mathcal{Y}, n \geq 1 \) with \( y_n \to y \) and;

(ii) There exists \( y_n \in \mathcal{Y}, n \geq 1 \), with \( y_n \to y \) such that \( F(y) = \lim_{n \to \infty} F_n(y_n) \).

**Proposition C.3.** [15, Theorems 7.8 and 7.23] Suppose that \( \mathcal{Y} \) is compact and \( \Gamma \lim_{n \to \infty} F_n = F \). Then \( M(F) = \{y \in \mathcal{Y} : F(y) = \inf_{z \in \mathcal{Y}} F(z)\} \) is non-empty, \( \lim_{n \to \infty} \inf_{y \in \mathcal{Y}} F_n(y) = \min_{y \in \mathcal{Y}} F(y) \), and if \( y_n \in M(F_n), n \geq 1 \), then the accumulation points of \( \langle y_n \rangle_{n \geq 1} \) belong to \( M(F) \).

**Definition C.4** (Kuratowski Convergence). [15, Remark 8.2.] Let \( A, A_n \subset \mathcal{Y}, n \geq 1 \). We say that the \( A_n \) converges to \( A \) in the sense of Kuratowski, denoted by \( K \lim_{n \to \infty} A_n = A \), if
(i) for all \( y \in A \), there exists a constant \( m \in \mathbb{N} \) and \( y_n \in A_n \) for \( n \geq m \) such that \( y_n \to y \), and;
(ii) for all \( y_n \in A_n \), \( n \geq 1 \), and for any accumulation point \( y \) of \( (y_n) \), \( y \in A \).

**Lemma C.5.** [6, Remark A.2] Convergence with respect to the Hausdorff distance implies convergence in the sense of Kuratowski. Moreover, if \( Y \) is compact both notions of convergence coincide.

**Lemma C.6.** [3, Proposition 4.15] If \( A_n \), \( A \subset Y \) for \( n \geq 1 \) then \( \text{K- lim}_{n \to \infty} A_n = A \) if and only if \( \Gamma- \lim_{n \to \infty} \chi_{A_n} = \chi_A \).

**C.2. Correspondence and Berge’s Maximum Theorem.** The following definitions and results, which can be found in [11], are used to prove Theorem 3.6. Recall that a correspondence \( \Phi : X \rightrightarrows Y \) maps each \( x \in X \) to a subset of \( Y \). Here we assume that \( X \) and \( Y \) are topological spaces. The graph of \( \Phi \) is defined by

\[
\text{Gr}(\Phi) = \{(x, y) \in X \times Y : y \in \Phi(x)\}.
\]

**Definition C.7.** Let \( \Phi : X \rightrightarrows Y \) be a correspondence and \( x \in X \).

(i) We say that \( \Phi \) is upper hemicontinuous at \( x \) if for every neighborhood \( U \) of \( \Phi(x) \), there is a neighborhood \( V \) of \( x \) such that \( z \in V \) implies \( \Phi(z) \subset U \).

(ii) We say that \( \Phi \) lower hemicontinuous at the point \( x \) if for every open set \( U \) that intersects \( \Phi(x) \), there is a neighborhood \( V \) of \( x \) such that \( z \in V \) implies \( \Phi(z) \cap U \neq \emptyset \).

If \( \Phi \) is both upper and lower hemicontinuous at \( x \) we say it is continuous at \( x \).

**Lemma C.8.** A singleton-valued correspondence is upper hemicontinuous if and only if it is lower hemicontinuous, in which case it is continuous as a function.

**Theorem C.9.** Given \( f_1, \ldots, f_k : X \to Y \), define a correspondence \( \Phi : X \to Y \) by \( \Phi(x) = \{f_1(x), \ldots, f_k(x)\} \). If each \( f_i \) is continuous at some point \( x_0 \) then:

(i) \( \Phi \) is continuous at \( x_0 \).

(ii) If \( Y \) is a locally convex topological vector space, then the correspondence \( x \mapsto \text{conv}(\Phi(x)) \) is also continuous at \( x_0 \).

**Theorem C.10** (Berge’s Maximum Theorem). Let \( \Phi : X \to Y \) be a continuous correspondence such that each \( \Phi(x) \) is a nonempty compact subset of \( Y \), and let \( f : \text{Gr}(\Phi) \to \mathbb{R} \) be continuous. Define the value function \( m : X \to \mathbb{R} \) by

\[
m(x) = \max_{y \in \Phi(x)} f(x, y),
\]

and the correspondence \( \mu : X \to Y \) of maximizers by

\[
\mu(x) = \{y \in \Phi(x) : f(x, y) = m(x)\}.
\]

Then: \( m \) is continuous and \( \mu \) has non-empty compact values. Moreover, if either \( f \) has a continuous extension to all of \( X \times Y \) or \( Y \) is Hausdorff, then \( \mu \) is upper hemicontinuous.

**C.3. Generalized envelope theorem.** The following result, taken from [26, Theorem 1] is used in the proof of Theorem 3.7. For convenience, we state the result using notations that are compatible with ours.
Theorem C.11 (Generalized envelope theorem). Consider the problem

\[ V(p) = \sup \{ J(a, p) : a \in A \text{ and } g_j(a, p) \geq 0 \text{ for } j = 1, \ldots, k \}, \]

where \( A \subset \mathbb{R}^m, p \in \mathcal{P} \subset \mathbb{R}^n \) and \( J \) and \( g_j \) are real-valued functions on \( \mathbb{R}^m \times \mathbb{R}^n \). For \( \lambda = (\lambda_1, \ldots, \lambda_k) \in \mathbb{R}_+^k \) consider the Lagrangian

\[ \mathcal{L}(p, a, \lambda) = J(a, p) + \sum_{j=1}^{k} \lambda_j g(a, p), \]

and assume that the following regularity conditions hold.

(i) The set \( A \) is convex.
(ii) The functions \( J \) and \( g_j, j = 1, \ldots, k \), are continuous on \( \mathbb{R}^m \times \mathbb{R}^n \).
(iii) The constraint set \( \Phi(p) = \{ a \in A : g_j(a, p) \geq 0, j = 1, \ldots, k \} \) is compact for every \( p \in \mathcal{P} \).
(iv) The correspondence \( \Phi(p) \) is continuous.
(v) For every \( p \in \mathcal{P} \) there exist \( j \) and \( \hat{a}_j \in A \) such that \( g_j(\hat{a}_j, p) > 0 \) and \( g_j(\hat{a}_j, p) \geq 0 \) for \( \ell \neq j \).
(vi) The set of saddle-points of \( \mathcal{L} \), which has the form \( A^*(p) \times \Lambda^*(p) \), is non-empty for every \( p \in \mathcal{P} \).
(vii) The gradients with respect to \( p, \nabla_p J, \nabla_p g_j \), exist and are jointly continuous in \( (a, p) \).

Then the directional derivatives of \( V \) exist for \( p \in \text{int}(\mathcal{P}) \) and are given by

\[ V''(p; \hat{p}) = \max_{a \in A^*(p)} \min_{\lambda \in \Lambda^*(p)} \left( \nabla_p J(a, p) + \sum_{j=1}^{k} \lambda_j \nabla_p g_j(a, p) \right) \hat{p}, \]

and the order of maximum and minimum does not matter.

Corollary C.12. Suppose that in addition to the assumptions of Theorem C.11 the following conditions hold at a given \( p_0 \in \text{int}(\mathcal{P}) \):

(i) \( J \) is strictly concave in \( a \) and \( g_j \) is concave in \( a \), \( j = 1, \ldots, k \).
(ii) \( J \) and \( g_j \) are continuously differentiable in \( a \).
(iii) The Linear Independence Constraint Qualification (LICQ) holds. That is, the vectors \( \nabla_a g_j(a^*, p_0) \) are linearly independent for binding constraints \( i \in I(a^*, p_0) = \{ j : g_j(a^*, p_0) = 0 \} \).

Then the saddle-point \((a^*(p_0), \lambda^*(p_0))\) is unique and the gradient of \( V \) exists at \( p_0 \) and is given by

\[ \nabla_p V(p_0) = \nabla_p J(a^*(p_0), p_0) + \sum_{j=1}^{k} \lambda^*_j(p_0) \nabla_p g_j(a^*(p_0), p_0). \]

Proof. Condition (i) implies the uniqueness of the saddle point solution \( a^*(p_0) \in A^*(p_0) \) and conditions (ii) and (iii) imply the uniqueness of the multiplier \( \lambda^*(p_0) \in \Lambda^*(p_0) \) (see the discussion in [26]). Consequently, the saddle point corresponding to \( p_0 \) is unique and the representation for the gradient follows from Theorem C.11. \( \square \)

\[ \text{For the definition of saddle point see [26] Section 2}. \]
C.4. **Miscellaneous results.** The following lemma summarizes some properties of level sets of convex functions. Details can be found in [30].

**Lemma C.13.** Let \( g : \mathbb{R}^n \to \mathbb{R} \cup \{ \infty \} \) be convex. Then \( L = \{ x \in \mathbb{R}^n : g(x) < 0 \} \) is open and convex. If \( L \) is non-empty then the closure of \( L \) is given by \( \overline{L} = \{ x \in \mathbb{R}^n : g(x) \leq 0 \} \) and \( \overline{L} = \text{int}(\overline{L}) \). In particular, we have \( \partial L = \partial \overline{L} = \{ x \in \mathbb{R}^n : g(x) = 0 \} \).

Now let \( g_1, \ldots, g_k \) be convex, and let \( L_j = \{ x \in \mathbb{R}^n : g_j(x) < 0 \} \). Suppose \( C = \bigcap_{j=1}^k L_j \) is nonempty. Then \( \partial C \subseteq \bigcup_{j=1}^k \partial L_j \).

**Acknowledgment**

This work is partially supported by NSERC Grant RGPIN-2019-04419, a Seed Funding for Methodologists Grant from the Data Sciences Institute (DSI) at the University of Toronto, and an NSERC Alexander Graham Bell Canada Graduate Scholarship (Application No. CGSD3-535625-2019).

**References**

[1] H. Abdi and L. J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[2] A. Agapova, R. Ferguson, and J. Greene. Market diversity and the performance of actively-managed portfolios. *The Journal of Portfolio Management*, 38(1):48–59, 2011.

[3] H. Attouch. *Variational Convergence for Functions and Operators*, volume 1. Pitman Advanced Publishing Program, 1984.

[4] J. Bai and P. Wang. Econometric analysis of large factor models. *Annual Review of Economics*, 8:53–80, 2016.

[5] M. Beraha and M. Pegoraro. Wasserstein principal component analysis for circular measures. *arXiv preprint arXiv:2304.02402*, 2023.

[6] J. Bigot, R. Gouet, T. Klein, and A. López. Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 53(1):1–26, 2017.

[7] E. M. Bronshtein and L. Ivanov. The approximation of convex sets by polyhedra. *Sibirskii matematicheskii zhurnal*, 16(5):1110–1112, 1975.

[8] S. Campbell and T.-K. L. Wong. Functional portfolio optimization in stochastic portfolio theory. *SIAM Journal on Financial Mathematics*, 13(2):576–618, 2022.

[9] E. Cazelles, V. Seguy, J. Bigot, M. Cuturi, and N. Papadakis. Geodesic pca versus log-pca of histograms in the wasserstein space. *SIAM Journal on Scientific Computing*, 40(2):B429–B456, 2018.

[10] Center for Research in Security Prices (CRSP). CRSP US Stock Database. [http://www.crsp.org/products/research-products/crsp-us-stock-databases](http://www.crsp.org/products/research-products/crsp-us-stock-databases), 2021.

[11] D. Charalambos and B. Aliprantis. *Infinite Dimensional Analysis: A Hitchhiker’s Guide*. Springer, 2006.

[12] Y. Chen, Z. Lín, and H.-G. Müller. Wasserstein regression. *Journal of the American Statistical Association*, pages 1–14, 2021.

[13] M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal components analysis to the exponential family. *Advances in Neural Information Processing Systems*, 14, 2001.

[14] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223, 2001.

[15] G. Dal Maso. *An Introduction to $\Gamma$-Convergence*, volume 8. Springer Science & Business Media, 2012.

[16] D. Eddelbuettel and R. Françoise. Repp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

[17] J. J. Egozcue, J. L. Díaz-Barrero, and V. Pawlowsky-Glahn. Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica*, 22(4):1175–1182, 2006.
[18] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

[19] E. R. Fernholz. *Stochastic Portfolio Theory*. Springer, 2002.

[20] R. Fernholz, R. Garvy, and J. Hannon. Diversity-weighted indexing. *Journal of Portfolio Management*, 24(2):74, 1998.

[21] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005, 2004.

[22] D. Itkin and M. Larsson. Open markets and hybrid Jacobi processes. *arXiv preprint arXiv:2110.14046*, 2021.

[23] S. G. Johnson. The NLopt nonlinear-optimization package, 2010.

[24] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[25] J. Maas. Gradient flows of the entropy for finite markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011.

[26] R. Marimon and J. Werner. The envelope theorem, Euler and Bellman equations, without differentiability. *Journal of Economic Theory*, 196:105309, 2021.

[27] R. J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

[28] M. Pegoraro and M. Beraha. Projected statistical methods for distributional data on the real line with the Wasserstein metric. *Journal of Machine Learning Research*, 23(37):1–59, 2022.

[29] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[30] R. T. Rockafellar. *Convex Analysis*. Number 28. Princeton university press, 1970.

[31] J. Ruf and K. Xie. The impact of proportional transaction costs on systematically generated portfolios. *SIAM Journal on Financial Mathematics*, 11(3):881–896, 2020.

[32] R. Schneider. *Convex Bodies: The Brunn–Minkowski Theory*. Cambridge University Press, 2014.

[33] R. Schneider and J. A. Wieacker. Approximation of convex bodies by polytopes. *Bulletin of the London Mathematical Society*, 13(2):149–156, 1981.

[34] H. L. Shang. A survey of functional principal component analysis. *ASA Advances in Statistical Analysis*, 98(2):121–142, 2014.

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

[36] C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

[37] C. Zhang, P. Kokoszka, and A. Petersen. Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis*, 43(1):30–52, 2022.

[38] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada.

*Email address: steven.campbell@mail.utoronto.ca*

Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada.

*Email address: tkl.wong@utoronto.ca*