Optimal sample acquisition for optimally weighted PCA from heterogeneous quality sources

David Hong, *Member, IEEE,* and Laura Balzano, *Senior Member, IEEE*

***Abstract*—Modern high-dimensional datasets are often formed by acquiring samples from multiple sources having heterogeneous quality, i.e., some sources are noisier than others. Collecting data in this manner raises the following natural question: what is the best way to collect the data (i.e., how many samples should be acquired from each source) given constraints (e.g., on time or energy)? In general, the answer depends on what analysis is to be performed. In this paper, we study the foundational signal processing task of estimating underlying low-dimensional principal components. Since the resulting dataset will be high-dimensional and will have heteroscedastic noise, we focus on the recently proposed optimally weighted PCA, which is designed specifically for this setting. We develop an efficient method for designing sample acquisitions that optimize the asymptotic performance of optimally weighted PCA given resource constraints, and we illustrate the proposed method through various case studies.**

***Index Terms*—heterogeneous quality, large-dimensional data, principal component analysis, sample acquisition design**

I. Introduction

Modern high-dimensional datasets are often formed by combining samples acquired from multiple sources, where the sources have heterogeneous quality and cost. Namely, some sources are noisier than others, and acquiring samples from higher-quality (less noisy) sources is typically more costly (whether in time or energy). For example, air quality data are currently acquired using both low-cost consumer-grade sensors and high-precision instruments that are carefully maintained by government agencies [1], [2]. In general, one can imagine deploying sensor networks with a heterogeneous mix of sensors of varying cost and corresponding quality.

One often seeks to find underlying low-dimensional structure revealed by the data. Principal component analysis (PCA) is a foundational technique for this task, and it is a workhorse method in modern signal processing. For example, PCA has been used to identify sources of air pollution from air quality data [3], [4]. However, conventional PCA is not designed for samples with heterogeneous amounts of noise; it treats all samples uniformly. A weighted PCA that gives noisier samples less weight is a more suitable method, and recent work [5] derived optimal weights for this scenario. The resulting optimally weighted PCA optimally downweights noisier samples to best recover the underlying low-dimensional components.

This paper tackles the following question: *what is the best way to acquire samples for optimally weighted PCA?* Namely, given multiple sources of data, how many samples should be acquired from each source to maximize the performance of optimally weighted PCA? Naturally, the more data, the better the performance of optimally weighted PCA. However, in practice there are typically constraints. In particular, acquiring samples often has associated costs, resulting in constraints of the following form:

$$\kappa_1 n_1 + \cdots + \kappa_L n_L \leq \tau, \tag{1}$$

where $n_1, \ldots, n_L$ are the number of samples acquired from each of L available data sources, $\kappa_1, \ldots, \kappa_L$ are the per-sample costs for each source, and τ is the corresponding budget. Each type of cost (e.g., in time or energy) adds a constraint of this form. Moreover, each source often has a finite quantity of samples it can provide, resulting in additional constraints:

$$n_\ell \leq q_\ell, \quad \text{for } \ell = 1, \ldots, L, \tag{2}$$

where $q_1, \ldots, q_L$ are the quantity of samples that each source can provide. Both (1) and (2) are linear constraints, so can be captured in general as follows:

$$\boldsymbol{A}\boldsymbol{n} \preccurlyeq \boldsymbol{b}, \tag{3}$$

where $\boldsymbol{n} := (n_1, \ldots, n_L)$ is the number of samples to acquire from each source, $\preccurlyeq$ denotes entrywise inequality, and $\boldsymbol{A}$ and $\boldsymbol{b}$ define the constraints. The goal is to choose $\boldsymbol{n}$ to maximize the performance of optimally weighted PCA subject to the constraints (3).

A number of recent works have studied the topic of PCA for high-dimensional heterogeneous-quality data [5]–[21], but they generally focus on how to use the given data rather than on how to acquire it. The question of how to best acquire data falls within the field of experimental design, for which there are numerous works (many more than can be reviewed here, see, e.g., the textbooks [22]–[27]). However, to the best of our knowledge, the question we consider here (optimal design for optimally weighted PCA) has not yet been addressed. This paper tackles this open problem.

Section II describes the data model and reviews optimally weighted PCA. Section III states the main result: a characterization of the optimal sample acquisition designs that enables us to develop a computationally efficient algorithm for finding an optimal design. Sections IV and V demonstrate the method through an illustrative example and case studies. Section VI concludes the paper with a proof of the main result.

D. Hong was supported in part by NSF Graduate Research Fellowship DGE 1256260, NSF Grant ECCS-1508943, NSF BIGDATA grant IIS 1837992, the Dean's Fund for Postdoctoral Research of the Wharton School, and NSF Mathematical Sciences Postdoctoral Research Fellowship DMS 2103353. L. Balzano was supported in part by DARPA-16-43-D3M-FP-037, NSF CAREER award CCF-1845076, NSF Grant ECCS-1508943, and ARO YIP Award W911NF1910027.

D. Hong is with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA (email: hong@udel.edu).

L. Balzano is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA.

II. Data model and optimally weighted PCA

This paper considers acquiring $n_1, \dots, n_L$ samples from L data sources with associated noise variances $v_1, \dots, v_L > 0$, where the sources share k underlying orthonormal components $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k \in \mathbb{C}^d$ with signal variances $\lambda_1 > \cdots > \lambda_k > 0$. Precisely put, we model the data block $\boldsymbol{Y}_\ell \in \mathbb{C}^{d \times n_\ell}$ obtained by acquiring n_ℓ samples from the ℓth source as in [5]:

$$\boldsymbol{Y}_\ell = \boldsymbol{F}\boldsymbol{Z}_\ell + \boldsymbol{E}_\ell \in \mathbb{C}^{d \times n_\ell}, \quad \text{for } \ell = 1, \dots, L, \tag{4}$$

where d is the number of features,

- $\boldsymbol{F} := [\sqrt{\lambda_1}\boldsymbol{u}_1, \dots, \sqrt{\lambda_k}\boldsymbol{u}_k] \in \mathbb{C}^{d \times k}$ is a deterministic factor matrix common to all the sources,
- $\boldsymbol{Z}_\ell \in \mathbb{C}^{k \times n_\ell}$ is a coefficient matrix with IID entries that have zero mean and unit variance,
- $\boldsymbol{E}_\ell \in \mathbb{C}^{d \times n_\ell}$ is a noise matrix with IID entries that have zero mean and variance $v_\ell > 0$,

and the noise further satisfies a technical condition: bounded a-th moment for some $a > 4$, i.e., $\sup_{i,j} \mathbb{E}|(\boldsymbol{E}_\ell)_{i,j}|^a < \infty$.[1]

Since the data sources have different amounts of noise, one should use a weighted PCA to account for their heterogeneous quality. Given weights $\boldsymbol{w} := (w_1, \dots, w_L)$, weighted PCA estimates the ith underlying component $\boldsymbol{u}_i$ as

$$\hat{\boldsymbol{u}}_i(\boldsymbol{w}, \boldsymbol{Y}) := i\text{th leading eigenvector of } \sum_{\ell=1}^{L} w_\ell \boldsymbol{Y}_\ell \boldsymbol{Y}_\ell^{\mathsf{H}}, \tag{5}$$

i.e., $\hat{\boldsymbol{u}}_i(\boldsymbol{w}, \boldsymbol{Y})$ is the ith leading eigenvector of the $\boldsymbol{w}$-weighted sample covariance matrix. Naturally, one wants to use weights that maximize the recovery of $\boldsymbol{u}_i$. Such weights were recently found in [5]. Namely, asymptotically optimal weights $\boldsymbol{w}_i^\star$ and their corresponding asymptotic performance $\bar{r}_i^\star$ are given by

$$\boldsymbol{w}_i^\star = \left(\frac{1}{v_1} \frac{1}{1 + v_1/\lambda_i}, \dots, \frac{1}{v_L} \frac{1}{1 + v_L/\lambda_i} \right), \tag{6}$$

$$\bar{r}_i^\star = \text{the unique solution } x \in (0,1) \tag{7}$$

$$\text{of } \sum_{\ell=1}^{L} \frac{n_\ell/d}{v_\ell/\lambda_i} \frac{1-x}{v_\ell/\lambda_i + x} = 1,$$

unless $\sum_{\ell=1}^{L} (n_\ell/d)(\lambda_i/v_\ell)^2 \leq 1$, in which case $\bar{r}_i^\star = 0$.[2]

Roughly speaking, $|\boldsymbol{u}_i^{\mathsf{H}} \hat{\boldsymbol{u}}_i(\boldsymbol{w}_i^\star, \boldsymbol{Y})|^2 \approx \bar{r}_i^\star$ when the number of features d and the numbers of samples $n_1, \dots, n_L$ are all sufficiently large. Specifically, $\bar{r}_i^\star$ is the optimal performance in the limit where d and $n_1, \dots, n_L$ all grow to infinity with fixed aspect ratios n_ℓ/d;[3] see [5] for further details. This high-dimensional regime corresponds to many big data settings.

Note that the asymptotic performance of optimally weighted PCA is component-specific, so the optimal sample acquisition strategy may differ from component to component. Note also that computing $\bar{r}_i^\star$ requires either knowing the signal and noise variances λ_i and $v_1, \dots, v_L$ a priori or estimating them from data, e.g., using the methods described in [5, Example 7.2].

III. Main result: optimal sample acquisition

The problem is to choose $\boldsymbol{n} := (n_1, \dots, n_L)$, the numbers of samples to acquire from each source, to optimize the performance (7) of optimally weighted PCA subject to linear constraints on $\boldsymbol{n}$. Precisely put,

$$\operatorname*{argmax}_{\boldsymbol{n} \in \mathbb{R}_+^L} \ \bar{r}_i^\star(\boldsymbol{n}) \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{n} \preccurlyeq \boldsymbol{b}, \tag{8}$$

where $\mathbb{R}_+$ denotes the set of nonnegative real numbers, $\preccurlyeq$ is entrywise inequality, $\boldsymbol{A}$ and $\boldsymbol{b}$ define the linear constraints, and we have made the dependence of $\bar{r}_i^\star$ on $\boldsymbol{n}$ explicit. Throughout this paper, we will assume that $\boldsymbol{A}$ and $\boldsymbol{b}$ define nontrivial constraints in the sense that $\{\boldsymbol{n} \in \mathbb{R}_+^L : \boldsymbol{A}\boldsymbol{n} \preccurlyeq \boldsymbol{b}\}$ is not only nonempty but also bounded. Otherwise, one can of course take $\boldsymbol{n} \to \infty$ and achieve perfect performance. However, one can always expect to have nontrivial constraints in practice since collecting and processing samples takes both time and space.

Solving the optimization problem (8) is nontrivial because $\bar{r}_i^\star(\boldsymbol{n})$ is not given as an explicit expression in $\boldsymbol{n}$. It is instead defined implicitly in (7) via the roots of a rational function with coefficients that depend on $\boldsymbol{n}$. Furthermore, $\bar{r}_i^\star(\boldsymbol{n})$ is a nonlinear function of $\boldsymbol{n}$. Fortunately, as our main result shows, solutions occur at extreme points of the constraint region.

Theorem 1 (Optimal sample acquisition)**.** *The optimal sample acquisition problem* (8) *has a solution (not necessarily unique) that is an extreme point of the constraint polyhedron*

$$\mathcal{P} := \{\boldsymbol{n} \in \mathbb{R}_+^L : \boldsymbol{A}\boldsymbol{n} \preccurlyeq \boldsymbol{b}\}. \tag{9}$$

Remark 1 (Non-integer values). Note that we have relaxed $\boldsymbol{n}$ to have nonnegative *real-valued* (rather than integer-valued) entries. This relaxation should have little impact in the high-dimensional regime of interest; simply round the solution of (8) to obtain integer numbers of samples to acquire. One can also avoid this technicality by instead formulating (8) in terms of the (already real-valued) asymptotic aspect ratios n_ℓ/d.

It follows from Theorem 1 that global optimization of (8) can be accomplished by simply choosing the best among the extreme points of $\mathcal{P}$ (a finite set). This leads naturally to the following simple and efficient algorithm.

Algorithm 1 Sample acquisition optimization

```
Input: noise variances $\boldsymbol{v} \in \mathbb{R}_+^L$; signal variance $\lambda_i \in \mathbb{R}_+$;
    dimension $d \in \mathbb{N}$; linear constraints $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{K \times L} \times \mathbb{R}^K$.
1: $\mathcal{P} \leftarrow \{\boldsymbol{n} \in \mathbb{R}_+^L : \boldsymbol{A}\boldsymbol{n} \preccurlyeq \boldsymbol{b}\}$    ▷ constraint polyhedron
2: $\mathcal{E} \leftarrow$ extreme points of $\mathcal{P}$    ▷ collect candidates
3: $\boldsymbol{n}^\star \leftarrow \operatorname{argmax}_{\boldsymbol{n} \in \mathcal{E}} \bar{r}_i^\star(\boldsymbol{n})$    ▷ choose best candidate
4: return $\boldsymbol{n}^\star$
```

Remark 2 (Computing extreme points). The extreme points of $\mathcal{P}$ in Line 2 of Algorithm 1 can be efficiently obtained from $(\boldsymbol{A}, \boldsymbol{b})$ using standard polyhedral software; see, e.g., [29], [30].

The runtime of Algorithm 1 can grow dramatically when the number of sources is large; developing algorithms that scale better is an interesting direction for future work. That said, we found that cases with tens to even hundreds of constraints can often complete in fractions of a second when the number of

[1] This technical condition is satisfied by numerous distributions including the sub-Gaussian and sub-Exponential families [28, Prop. 2.5.2 and 2.7.1], notably including the real-valued Gaussian setting considered in [8], [9].

[2] When $\sum_{\ell=1}^{L} (n_\ell/d)(\lambda_i/v_\ell)^2 > 1$, the solution $x \in (0,1)$ to the equation in (7) can be found simply and efficiently via bisection since it is unique.

[3] The number of components k and the number of data sources L, as well as the signal and noise variances $\lambda_1, \dots, \lambda_k$ and $v_1, \dots, v_L$, are also considered fixed with respect to n_ℓ and d.

sources is moderate (e.g., around ten or less), as is the case in many applications.

IV. Illustrative Example

We illustrate our main result (Theorem 1) and the resulting algorithm (Algorithm 1) with the following example of two sources with limited quantities and a single budget constraint.

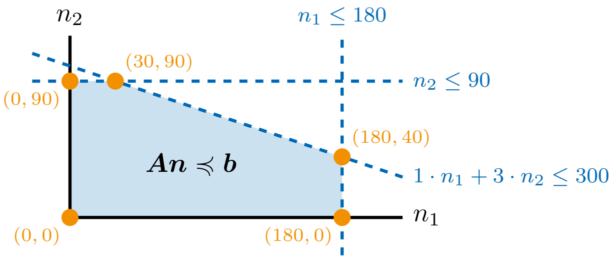
n_2
$n_1 \le 180$
$(30, 90)$
$n_2 \le 90$
$(0, 90)$
$(180, 40)$
$\boldsymbol{A}\boldsymbol{n} \preccurlyeq \boldsymbol{b}$
$1 \cdot n_1 + 3 \cdot n_2 \le 300$
n_1
$(0, 0)$
$(180, 0)$

Fig. 1. Constraint region from Example 1: $L = 2$ data sources with limited quantities of 180 and 90 samples, respectively, and corresponding per-sample costs of 1 and 3 with a total budget of 300.

Example 1. Consider acquiring $d = 100$ dimensional samples from $L = 2$ sources with an underlying component $\boldsymbol{u}_i$ having signal variance $\lambda_i = 10$, where

- source 1 samples have noise variance $v_1 = 2$, a limited quantity of 180 samples, and a per-sample cost of 1,
- source 2 samples have noise variance $v_2 = 1$, a limited quantity of 90 samples, and a per-sample cost of 3,

and the total budget is 300. This yields the linear constraints defined by

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 3 \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} 180 \\ 90 \\ 300 \end{bmatrix},$$

illustrated in Fig. 1. To find the optimal sample acquisition, we proceed according to Algorithm 1 (based on Theorem 1):

1) Find the extreme points $\mathcal{E}$ of $\mathcal{P} = \{\boldsymbol{n} \in \mathbb{R}_+^L : \boldsymbol{A}\boldsymbol{n} \preccurlyeq \boldsymbol{b}\}$:
$$\mathcal{E} = \{(0,0), (0,90), (30,90), (180,40), (180,0)\}.$$
2) Evaluate $\bar{r}_i^\star(\boldsymbol{n})$ for each $\boldsymbol{n} \in \mathcal{E}$ and choose the best:
$$(180, 40) \in \operatorname*{argmax}_{\boldsymbol{n} \in \mathcal{E}} \bar{r}_i^\star(\boldsymbol{n})$$
since
$$\bar{r}_i^\star\big((30,90)\big) \approx 0.903, \qquad \bar{r}_i^\star\big((180,40)\big) \approx 0.917.$$
Note that $(0, 0)$, $(0, 90)$, and $(180, 0)$ can all be skipped because either n_1 or n_2 can be increased for each choice, trivially yielding an improvement in performance.

The optimal sample acquisition $\boldsymbol{n}^\star = (180, 40)$ corresponds to collecting all the available samples from source 1 (i.e., all the cheaper, noisier samples that are available) then spending the rest of the budget on samples from source 2 (i.e., the more costly, higher quality samples).

V. Case studies and insights

The following case studies demonstrate Theorem 1 through a few interesting scenarios that provide some new insights into optimal sample acquisition for optimally weighted PCA.

A. Optimal sample acquisition under a single constraint

Suppose the constraint region reduces to a single constraint. For example, consider the setting of Example 1 but with over 300 source 1 samples and over 100 source 2 samples available. In this case, the constraints reduce to the single constraint

$$\boldsymbol{A} = \begin{bmatrix} 1 & 3 \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} 300 \end{bmatrix},$$

with three extreme points $\mathcal{E} = \{(0, 0), (300, 0), (0, 100)\}$. As before, $(0, 0)$ can be skipped, so either $(300, 0)$ or $(0, 100)$ is optimal. Namely, it is optimal to use only one source. In this case, $(300, 0)$ is better, i.e., it is optimal to acquire samples from only the noisier less-costly source.

Indeed, as the following corollary states, it is always optimal to use only one source when there is effectively one constraint.

Corollary 2 (Optimality for a single constraint)**.** *If the constraint region reduces to a single constraint, i.e.,*

$$\exists_{\boldsymbol{a} \in \mathbb{R}^L, b \in \mathbb{R}} \quad \mathcal{P} = \{\boldsymbol{n} \in \mathbb{R}_+^L : \boldsymbol{a}^\top \boldsymbol{n} \preccurlyeq b\},$$

then the optimal sample acquisition problem (8) *has a solution using only one source, i.e., a* 1*-sparse solution of the form*

$$\boldsymbol{n}^\star = (\boldsymbol{0}_{\ell-1}, b/a_\ell, \boldsymbol{0}_{L-\ell}) \quad \text{where} \quad \ell \in \{1, \dots, L\}. \tag{10}$$

Corollary 2 follows straightforwardly from Theorem 1 by observing that the nonzero extreme points are of the form (10).

While this result is perhaps natural given Theorem 1, note that it was not obvious a priori. It would have been natural to expect the optimal sample acquisition to require a precise mix of both high-quality high-cost samples and low-quality low-cost samples, especially given the nonlinearity of $\bar{r}_i^\star(\boldsymbol{n})$.

B. Optimal sample acquisition for multiple components

The performance of optimally weighted PCA depends nontrivially on the signal variance, so may vary from component to component. Thus, the optimal sample acquisition strategy could also differ among components. However, unlike the optimal weights (6), the optimal sample acquisition can also be the same for components that have different signal variances.

Consider the setting of Example 1 but where we sweep λ_i from 1 to 10. Fig. 2 shows the optimal sample acquisition strategies computed by Algorithm 1 (omitting the small interval $\lambda_i \in [1, 1.013]$ where the optimal performance was zero).

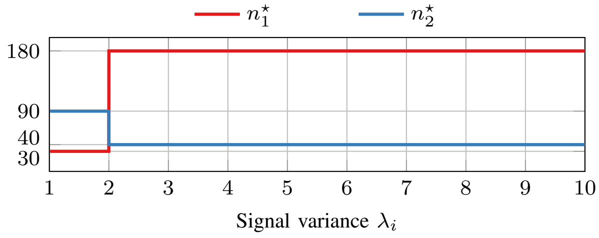

Fig. 2. Optimal sample acquisition $\boldsymbol{n}^\star = (n_1^\star, n_2^\star)$ v.s. signal variance λ_i.

Notably, the optimal sample acquisition strategy is constant over large intervals. Thus, there can be a single optimal strategy for all the components, as long as their signal variances lie in the same interval. Interestingly, these results also suggest that the optimal strategy may be somewhat robust to potential errors in estimating the signal and noise variances.

This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2025.3550280

VI. Proof of Main Result (Theorem 1)

Let $\mathcal{Q}$ be the solution set for the optimization problem (8), and let $\rho^\star$ be the associated optimal value, i.e.,

$$\mathcal{Q} := \operatorname*{argmax}_{\boldsymbol{n}\in\mathcal{P}} \bar{r}_i^\star(\boldsymbol{n}), \qquad \rho^\star := \sup_{\boldsymbol{n}\in\mathcal{P}} \bar{r}_i^\star(\boldsymbol{n}). \tag{11}$$

The goal is to show that

$$\exists\, \boldsymbol{n}^\star \in \mathcal{Q} \quad \text{s.t.} \quad \boldsymbol{n}^\star \text{ is an extreme point of } \mathcal{P}. \tag{12}$$

Note first that (12) holds trivially if $\rho^\star = 0$ because $\mathcal{Q} = \mathcal{P}$ in that case and $\mathcal{P}$ always has at least one extreme point (by [31, Theorem 2.6]) since it is a nonempty bounded polyhedron (by assumption). So it remains to show (12) for $\rho^\star > 0$.

The proof for $\rho^\star > 0$ proceeds in two stages: (i) show that $\mathcal{Q}$ has extreme points, and (ii) show that extreme points of $\mathcal{Q}$ are also extreme points of $\mathcal{P}$. (12) then follows immediately.

Stage 1. This stage shows that $\mathcal{Q}$ has extreme points when $\rho^\star > 0$. To begin, we establish the following properties of $\mathcal{Q}$:

(nonempty) $\bar{r}_i^\star(\boldsymbol{n})$ is a continuous function on the domain $\mathcal{P}$ and $\mathcal{P}$ is both compact (it is a bounded polyhedron) and nonempty. Thus, $\bar{r}_i^\star(\boldsymbol{n})$ attains its maximum $\rho^\star$ on $\mathcal{P}$, and $\mathcal{Q}$ is nonempty.

(bounded) $\mathcal{Q} \subseteq \mathcal{P}$ and $\mathcal{P}$ is bounded (by assumption), so $\mathcal{Q}$ is also bounded.

(polyhedron) Note that

$$\forall_{\boldsymbol{n}\in\mathbb{R}^L_+,\rho>0} \quad \left[\bar{r}_i^\star(\boldsymbol{n}) = \rho \iff R_{i,\boldsymbol{n}}(\rho) = 0\right], \tag{13}$$

where

$$R_{i,\boldsymbol{n}}(x) := 1 - \sum_{\ell=1}^{L} \frac{n_\ell/d}{v_\ell/\lambda_i}\,\frac{1-x}{v_\ell/\lambda_i + x}, \tag{14}$$

because $R_{i,\boldsymbol{n}}(x)$ has exactly one root in the interval $(0,1)$ unless $\sum_{\ell=1}^{L}(n_\ell/d)(\lambda_i/v_\ell)^2 \leq 1$, in which case it has no roots in the interval. Thus, $\mathcal{Q}$ can be expressed as follows:

$$\begin{aligned}\mathcal{Q} &= \{\boldsymbol{n}\in\mathcal{P} : \bar{r}_i^\star(\boldsymbol{n}) = \rho^\star\} \\ &= \{\boldsymbol{n}\in\mathcal{P} : R_{i,\boldsymbol{n}}(\rho^\star) = 0\} \\ &= \left\{\boldsymbol{n}\in\mathcal{P} : 1 - \sum_{\ell=1}^{L} \frac{n_\ell/d}{v_\ell/\lambda_i}\,\frac{1-\rho^\star}{v_\ell/\lambda_i + \rho^\star} = 0\right\} \\ &= \mathcal{P} \cap \left\{\boldsymbol{n}\in\mathbb{R}^L : \sum_{\ell=1}^{L} n_\ell \cdot \left[\frac{\lambda_i}{v_\ell}\,\frac{1-\rho^\star}{v_\ell/\lambda_i + \rho^\star}\right] = d\right\},\end{aligned} \tag{15}$$

where the first line follows from the definition of $\mathcal{Q}$, the second line follows from (13), the third line follows by substituting (14), and the fourth line is rearranging. It follows from the form of (15) that $\mathcal{Q}$ is a polyhedron.

Hence, $\mathcal{Q}$ is a nonempty bounded polyhedron, and it follows from [31, Theorem 2.6] that $\mathcal{Q}$ has at least one extreme point.

Stage 2. This stage shows that any extreme point of $\mathcal{Q}$ must also be an extreme point of $\mathcal{P}$ when $\rho^\star > 0$. We proceed by proving the contrapositive, i.e., that any non-extreme point of $\mathcal{P}$ cannot be an extreme point of $\mathcal{Q}$.

Note that the statement trivially holds for any non-extreme point of $\mathcal{P}$ that is not in $\mathcal{Q}$. So, it remains to consider the case where the non-extreme point of $\mathcal{P}$ is in $\mathcal{Q}$. Let $\boldsymbol{n}^\star \in \mathcal{Q}$ be a non-extreme point of $\mathcal{P}$. Then, there must exist points $\boldsymbol{s}, \boldsymbol{t} \in \mathcal{P} \setminus \{\boldsymbol{n}^\star\}$ and $\gamma \in (0,1)$ so that

$$\boldsymbol{n}^\star = \gamma \cdot \boldsymbol{s} + (1-\gamma)\cdot\boldsymbol{t}. \tag{16}$$

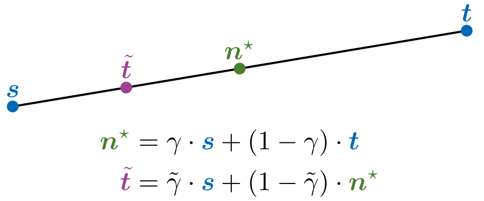
$\boldsymbol{s}$
$\tilde{\boldsymbol{t}}$
$\boldsymbol{n}^\star$
$\boldsymbol{t}$
$\boldsymbol{n}^\star = \gamma\cdot\boldsymbol{s} + (1-\gamma)\cdot\boldsymbol{t}$
$\tilde{\boldsymbol{t}} = \tilde{\gamma}\cdot\boldsymbol{s} + (1-\tilde{\gamma})\cdot\boldsymbol{n}^\star$

Fig. 3. Diagram illustrating the relative positions of $\boldsymbol{s}$, $\tilde{\boldsymbol{t}}$, $\boldsymbol{n}^\star$, and $\boldsymbol{t}$. Namely, $\boldsymbol{n}^\star$ is between $\boldsymbol{s}$ and $\boldsymbol{t}$, and $\tilde{\boldsymbol{t}}$ is between $\boldsymbol{s}$ and $\boldsymbol{n}^\star$.

Moreover, since $\bar{r}_i^\star(\boldsymbol{n})$ is continuous and $\mathcal{P}$ is convex, we can always choose the points $\boldsymbol{s}$ and $\boldsymbol{t}$ to be close enough to $\boldsymbol{n}^\star$ so that $\bar{r}_i^\star(\boldsymbol{s}), \bar{r}_i^\star(\boldsymbol{t}) > \rho^\star/2$. Without loss of generality, suppose that $\bar{r}_i^\star(\boldsymbol{s}) \leq \bar{r}_i^\star(\boldsymbol{t})$. Since $\boldsymbol{n}^\star$ maximizes $\bar{r}_i^\star(\boldsymbol{n})$, this yields

$$\bar{r}_i^\star(\boldsymbol{s}) \leq \bar{r}_i^\star(\boldsymbol{t}) \leq \bar{r}_i^\star(\boldsymbol{n}^\star). \tag{17}$$

Thus, it follows by the intermediate value theorem that $\bar{r}_i^\star(\boldsymbol{n})$ takes on the value $\tilde{\rho} := \bar{r}_i^\star(\boldsymbol{t})$ at a point between $\boldsymbol{s}$ and $\boldsymbol{n}^\star$, i.e., there exists some $\tilde{\gamma} \in [0,1]$ so that

$$\bar{r}_i^\star(\tilde{\boldsymbol{t}}) = \tilde{\rho}, \quad \text{where} \quad \tilde{\boldsymbol{t}} := \tilde{\gamma}\cdot\boldsymbol{s} + (1-\tilde{\gamma})\cdot\boldsymbol{n}^\star. \tag{18}$$

Note next that $\boldsymbol{s}$, $\tilde{\boldsymbol{t}}$, $\boldsymbol{n}^\star$, and $\boldsymbol{t}$ are collinear (as shown in Fig. 3), so it follows that $\boldsymbol{n}^\star$ and $\boldsymbol{s}$ are both affine combinations of $\tilde{\boldsymbol{t}}$ and $\boldsymbol{t}$. Specifically, solving (16) and (18) for $\boldsymbol{n}^\star$ and $\boldsymbol{s}$ yields

$$\boldsymbol{n}^\star = \mu\cdot\tilde{\boldsymbol{t}} + (1-\mu)\cdot\boldsymbol{t}, \qquad \boldsymbol{s} = \tilde{\mu}\cdot\tilde{\boldsymbol{t}} + (1-\tilde{\mu})\cdot\boldsymbol{t}, \tag{19}$$

where $\mu = \gamma/(\gamma + \tilde{\gamma} - \gamma\tilde{\gamma})$ and $\tilde{\mu} = 1/(\gamma + \tilde{\gamma} - \gamma\tilde{\gamma})$.

Substituting the first equation of (19) into (14) yields

$$\begin{aligned}R_{i,\boldsymbol{n}^\star}(x) &= 1 - \sum_{\ell=1}^{L} \frac{n_\ell^\star/d}{v_\ell/\lambda_i}\,\frac{1-x}{v_\ell/\lambda_i + x} \\ &= \mu\cdot\left(1 - \sum_{\ell=1}^{L} \frac{\tilde{t}_\ell/d}{v_\ell/\lambda_i}\,\frac{1-x}{v_\ell/\lambda_i + x}\right) \\ &\quad + (1-\mu)\cdot\left(1 - \sum_{\ell=1}^{L} \frac{t_\ell/d}{v_\ell/\lambda_i}\,\frac{1-x}{v_\ell/\lambda_i + x}\right) \\ &= \mu\cdot R_{i,\tilde{\boldsymbol{t}}}(x) + (1-\mu)\cdot R_{i,\boldsymbol{t}}(x).\end{aligned} \tag{20}$$

Likewise, substituting the second equation of (19) into (14) yields

$$R_{i,\boldsymbol{s}}(x) = \tilde{\mu}\cdot R_{i,\tilde{\boldsymbol{t}}}(x) + (1-\tilde{\mu})\cdot R_{i,\boldsymbol{t}}(x). \tag{21}$$

Since $\bar{r}_i^\star(\tilde{\boldsymbol{t}}) = \bar{r}_i^\star(\boldsymbol{t}) = \tilde{\rho}$, it follows that $R_{i,\tilde{\boldsymbol{t}}}(\tilde{\rho}) = R_{i,\boldsymbol{t}}(\tilde{\rho}) = 0$ and hence $R_{i,\boldsymbol{n}^\star}(\tilde{\rho}) = R_{i,\boldsymbol{s}}(\tilde{\rho}) = 0$. Noting that $\tilde{\rho} > \rho^\star/2 > 0$ and applying (13) then yields that $\bar{r}_i^\star(\boldsymbol{n}^\star) = \bar{r}_i^\star(\boldsymbol{s}) = \tilde{\rho}$, and as a result

$$\bar{r}_i^\star(\boldsymbol{s}) = \bar{r}_i^\star(\tilde{\boldsymbol{t}}) = \bar{r}_i^\star(\boldsymbol{n}^\star) = \bar{r}_i^\star(\boldsymbol{t}), \tag{22}$$

i.e., $\bar{r}_i^\star(\boldsymbol{s}) = \bar{r}_i^\star(\boldsymbol{t}) = \rho^\star$ so $\boldsymbol{s}, \boldsymbol{t} \in \mathcal{Q} \setminus \{\boldsymbol{n}^\star\}$, implying that $\boldsymbol{n}^\star$ is not an extreme point of $\mathcal{Q}$ and completing the proof. □

Acknowledgment

The authors thank Jeffrey A. Fessler and Claire Y. Hong for helpful discussions during the development of these ideas.

This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2025.3550280

References

[1] US Environmental Protection Agency, "Air data: Air quality data collected at outdoor monitors across the us," accessed on November 10, 2022. [Online]. Available: https://www.epa.gov/outdoor-air-quality-data

[2] PurpleAir, "Real time air quality monitoring," accessed on November 10, 2022. [Online]. Available: https://www2.purpleair.com

[3] M. Chavent, H. Guegan, V. Kuentz, B. Patouille, and J. Saracco, "PCA- and PMF-based methodology for air pollution sources identification and apportionment," *Environmetrics: The official journal of the International Environmetrics Society*, vol. 20, no. 8, pp. 928–942, 2009.

[4] A. Azid, H. Juahir, M. E. Toriman, M. K. A. Kamarudin, A. S. M. Saudi, C. N. C. Hasnam, N. A. A. Aziz, F. Azaman, M. T. Latif, S. F. M. Zainuddin *et al.*, "Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia," *Water, Air, & Soil Pollution*, vol. 225, no. 8, pp. 1–14, 2014.

[5] D. Hong, F. Yang, J. A. Fessler, and L. Balzano, "Optimally weighted PCA for high-dimensional heteroscedastic data," *SIAM Journal on Mathematics of Data Science*, vol. 5, no. 1, pp. 222–250, 2023.

[6] D. Hong, L. Balzano, and J. A. Fessler, "Towards a theoretical analysis of PCA for heteroscedastic data," in *54th Allerton Conference on Communication, Control, and Computing*, Sep. 2016, pp. 496–503.

[7] ——, "Asymptotic performance of PCA for high-dimensional heteroscedastic data," *Journal of Multivariate Analysis*, vol. 167, pp. 435–452, 2018.

[8] ——, "Probabilistic PCA for heteroscedastic data," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Dec. 2019.

[9] D. Hong, K. Gilman, L. Balzano, and J. A. Fessler, "HePPCAT: Probabilistic PCA for data with heteroscedastic noise," *IEEE Transactions on Signal Processing*, vol. 69, pp. 4819–4834, 2021.

[10] A. Breloy, G. Ginolhac, F. Pascal, and P. Forster, "Clutter subspace estimation in low rank heterogeneous noise context," *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2173–2182, 2015.

[11] ——, "Robust covariance matrix estimation in heterogeneous low rank context," *IEEE Transactions on Signal Processing*, vol. 64, no. 22, pp. 5794–5806, 2016.

[12] Y. Sun, A. Breloy, P. Babu, D. P. Palomar, F. Pascal, and G. Ginolhac, "Low-complexity algorithms for low rank clutter parameters estimation in Radar systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 8, pp. 1986–1998, 2016.

[13] A. Collas, F. Bouchard, A. Breloy, G. Ginolhac, C. Ren, and J.-P. Ovarlez, "Probabilistic PCA from heteroscedastic signals: Geometric framework and application to clustering," *IEEE Transactions on Signal Processing*, vol. 69, pp. 6546–6560, 2021.

[14] O. Besson, "Bounds for a mixture of low-rank compound-Gaussian and white Gaussian noises," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5723–5732, 2016.

[15] R. B. Abdallah, A. Breloy, M. N. E. Korso, and D. Lautru, "Bayesian signal subspace estimation with compound Gaussian sources," *Signal Processing*, vol. 167, pp. 1–15, 2020.

[16] W. Leeb and E. Romanov, "Optimal spectral shrinkage and PCA with heteroscedastic noise," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 3009–3037, 2021.

[17] W. E. Leeb, "Matrix denoising for weighted loss functions and heterogeneous signals," *SIAM Journal on Mathematics of Data Science*, vol. 3, no. 3, pp. 987–1012, 2021.

[18] A. R. Zhang, T. T. Cai, and Y. Wu, "Heteroskedastic PCA: Algorithm, optimality, and applications," *The Annals of Statistics*, vol. 50, no. 1, pp. 53–80, 2022.

[19] D. Hong, Y. Sheng, and E. Dobriban, "Selecting the number of components in PCA via random signflips," 2024. [Online]. Available: http://arxiv.org/abs/2012.02985v3

[20] Z. T. Ke, Y. Ma, and X. Lin, "Estimation of the number of spiked eigenvalues in a covariance matrix by bulk eigenvalue matching analysis," *Journal of the American Statistical Association*, pp. 1–19, 2021.

[21] B. Landa, T. T. C. K. Zhang, and Y. Kluger, "Biwhitening reveals the rank of a count matrix," *SIAM Journal on Mathematics of Data Science*, vol. 4, no. 4, pp. 1420–1446, 2022.

[22] R. L. Mason, R. F. Gunst, and J. L. Hess, *Statistical Design and Analysis of Experiments: With Applications to Engineering and Science*. Wiley, 2003.

[23] F. Pukelsheim, *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006.

[24] P. Goos and B. Jones, *Optimal Design of Experiments: A Case Study Approach*. Wiley, 2011.

[25] D. C. Montgomery, *Design and Analysis of Experiments*, 10th ed. Wiley, 2019.

[26] C. F. J. Wu and M. Hamada, *Experiments: Planning, Analysis, and Optimization*. Wiley, 2021.

[27] H.-M. Kaltenbach, *Statistical Design and Analysis of Biological Experiments*. Springer International Publishing, 2021.

[28] R. Vershynin, *High-Dimensional Probability*. Cambridge University Press, 2018.

[29] M. Forets and C. Schilling, "LazySets.jl: Scalable Symbolic-Numeric Set Computations," *Proceedings of the JuliaCon Conferences*, vol. 1, no. 1, p. 11, 2021.

[30] B. Legat, "Polyhedral computation," in *JuliaCon*, 2023. [Online]. Available: https://pretalx.com/juliacon2023/talk/JP3SPX/

[31] D. Bertsimas and J. Tsitsiklis, *Introduction to Linear Optimization*. Athena Scientific, 1997.