On the Statistical Properties of Entromin as an Orthogonal Rotation Criterion

by

Kenny Chiu

B.Sc., The University of British Columbia, 2018

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Statistics)

The University of British Columbia
(Vancouver)

August 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

On the Statistical Properties of Entromin as an Orthogonal Rotation Criterion

submitted by **Kenny Chiu** in partial fulfillment of the requirements for the degree of **Master of Science** in **Statistics**.

Examining Committee:

Benjamin Bloem-Reddy, Statistics
Supervisor

Daniel J. McDonald, Statistics
Supervisory Committee Member

Abstract

The primary goal of factor analysis is to uncover a set of latent factors that can explain the variation in the data. Principal Component Analysis is one approach that estimates the factors by a set of orthogonal vectors. However, it may be difficult to interpret the factors as-is, and so it is common to rotate the estimated factors to make their coefficients as sparse as possible to improve interpretability. Varimax is the most popular method for factor rotations, and its statistical properties have been studied in recent literature. Entromin is another factor rotation method that is less commonly used and not as well-studied, but there exists conventional wisdom that Entromin generally finds sparser rotations compared to Varimax.

In this thesis, we aim to explain the sparsity claim for Entromin by studying its statistical properties. Our main contributions include several theoretical results that take steps towards this aim. We show that Varimax is a first-order approximation of Entromin, and that generalizing this connection leads to a family of Entromin approximations. We derive the conditions under which the second-order approximation can be viewed as performing statistical inference in a latent factor model. We then make the connection between optimizing the Entromin criterion and recovering sparsity in the factors. Other contributions of this thesis include novel connections to statistical concepts that have not been made in the literature to our knowledge, and an empirical study of Entromin on a dataset of New York Times articles.

Lay Summary

The primary goal of factor analysis is to find a set of unobserved factors that can explain the observed data. A factor is typically represented by a vector of numbers. It is common to rotate the factors to make as many numbers zero as possible to improve interpretability. Varimax is the most popular method for factor rotations and has been studied in the statistical literature. Entromin is another factor rotation method that is less popular and not as well-studied, but it is said that Entromin generally returns factors that have more zeros compared to Varimax. In this thesis, we aim to statistically explain this claim for Entromin. Our main contributions consist of several theoretical results that take steps towards this aim, including the mathematical connection between Entromin and Varimax. Other contributions include an empirical study of Entromin on a dataset of New York Times articles.

Preface

This thesis is original and unpublished work by the author, Kenny Chiu, under the supervision of Professor Benjamin Bloem-Reddy. The statement and proof techniques of Theorem 4, Proposition 5, Corollary 6 and Proposition 7 are based off of similar results by Rohe and Zeng (2020). The empirical analyses in Chapter 5 make use of the code by Rohe and Zeng (2020) available on their public GitHub repository (accessed June 7, 2021). Proposition 1 is a well-known result that is included only for convenience. All other results, code and analyses were primarily contributed by the author.

Table of Contents

List of Tables

List of Figures

List of Results

List of Assumptions

List of Definitions

List of Algorithms

Acknowledgments

I must express my sincere appreciation to my supervisor, Professor Benjamin Bloem-Reddy, without whom this thesis would not be possible. Ben has provided me invaluable guidance over the course of my M.Sc., extended his empathy and offered me words of wisdom during challenging times, and motivated me to continue in academia and to pursue a Ph.D. I am excited for what the future has to bring as we continue our research on various interesting statistical topics.

I would like to thank my second reader, Professor Daniel J. McDonald, who has taken the time to thoroughly review this thesis. I would also like to acknowledge the various members of the UBC Department of Statistics who have played a key role during my M.Sc. To the faculty and ASDa consultants whom I have learned much from, to the staff whom I could always count on to take my matters into their own hands, and to my fellow peers whom I have bonded and made memories with—thank you all.

I would like to acknowledge the UBC Faculty of Science STAIR grant program and NSERC for funding my research assistantship with Ben. I am also grateful to my director, Seyed Ali Mussavi Rizi, at the Provincial Health Services Authority as well as to the rest of the Data Analytics Research and Evaluation team who have supported my return to academic studies and provided the opportunity to work throughout my M.Sc. and onwards.

Finally, I would like to give a shoutout to my friends and family for their continual support. This thesis is dedicated to my friends and their pets who have kept me sane—and to my mom, my dad, and my sister for their unconditional love.

Chapter 1

Introduction

In factor analysis, it is assumed that a set of latent factors generates the observed data. The goal is to then infer and estimate some representation of the latent factors that can explain the observed variation in the data. As the inferred factors need to be interpretable for them to be useful to the analyst, the representation is generally restricted to be low-dimensional in order to ease interpretation.

Principal Component Analysis (PCA) is one common method for performing factor analysis where the factors are estimated by a small set of orthogonal vectors. Given a data matrix $A \in \mathbb{R}^{n \times d}$, PCA finds the transformation

$$S = AV \tag{1}$$

where the columns of $V \in \mathbb{R}^{d \times d}$ (the eigenvectors of $A^T A$) are the scaled *loadings* and the columns of $S \in \mathbb{R}^{n \times d}$ are the *principal components*. The principal components are chosen sequentially such that each successive component explains the most variation that remains unexplained by the previously chosen components. For example, the loadings for the first two principal components are

$$V_{(1)} = \underset{\|v\|=1}{\arg\max}\ v^T A^T A v\ ,$$
$$V_{(2)} = \underset{\|v\|=1, v^T V_{(1)}=0}{\arg\max}\ v^T A^T A v\ .$$

The factors are represented by the leading k principal components. However, interpreting each factor without further adjustments may still be a challenge if all of its coefficients are non-trivial. For this reason, a natural post-processing step after performing PCA is to rotate the principal components to make as many coefficients zero or as small as possible. The representation of the data can be made invariant to rotations of the principal components by rotating the loadings accordingly, but a change of basis may make the individual factors sparser and thus easier to interpret (Thurstone, 1935).

Finding the rotation that leads to the sparsest factors is typically done through a procedure in which some criterion is optimized. One popular method for factor rotation is *Varimax* (Kaiser, 1958). Varimax finds the rotation of the principal component matrix that maximizes the sum of column variances, and it can be shown that maximizing this criterion correlates with promoting sparsity in the factors (Kaiser, 1958). Figure 1 shows an example of eight unrotated principal components from the New York Times articles dataset examined in Section 5.2 and the same principal components rotated by Varimax. Visually, Varimax attempts to align the observed "radial streaks" with the axes which corresponds to making a coefficient of one principal component (close to) zero when a coefficient of another is large. While Varimax is simply finding the rotation that maximizes its particular criterion, Rohe and Zeng (2020) showed that under certain conditions, Varimax can be viewed as performing statistical inference on the assumed latent factors.

In this thesis, we study the *Entromin* factor rotation method (McCammon, 1970; Inagaki, 1993), which uses a form of entropy from information theory as a criterion. Entromin is seemingly less popular than Varimax, with fewer dedicated works in the literature and no base R implementation in contrast. However, there appears to exist conventional wisdom that Entromin tends to recover sparser factors compared to Varimax (McCammon, 1970). The goal of our work is to explain this claim and to take steps towards a statistical understanding of Entromin. Our theoretical results suggest that there is indeed a connection between Entromin and Varimax, and that there may be some truth to the claim that Entromin recovers sparser factors from a statistical perspective. The results from our empirical anal-

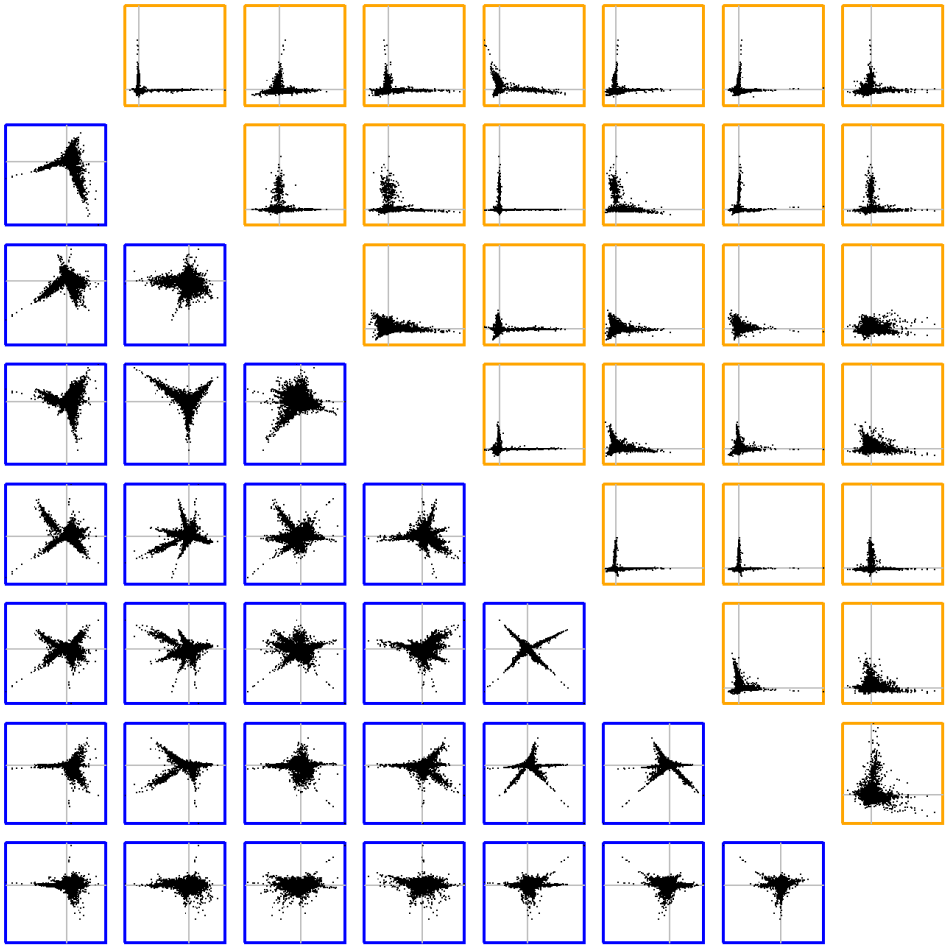

Figure 1: Pairwise plots of eight unrotated principal components (blue) and Varimax-rotated components (orange) for the New York Times articles dataset. See Section 5.2 for more details about the dataset. (Best viewed in colour.)

yses also suggest that the rotated factors obtained from Entromin are generally as sparse as the factors obtained from Varimax, if not sparser.

The main contributions of this thesis include

1. Corollary 3, which states that Varimax is a first-order approximation of En-

tromin under certain conditions;

2. a family of factor rotation methods that approximate Entromin;

3. Theorem 4, which states the exact conditions under which a second-order approximation of Entromin is expected to recover the true latent factors in the limit of an infinite number of samples;

4. Theorem 9, which makes the formal connection between minimizing a particular Entromin criterion and recovering sparsity in the factors;

5. novel connections between properties of Entromin and statistical concepts such as cumulants which have not been made in the literature to the best of our knowledge; and

6. an analysis of the algorithms used to implement orthogonal rotation methods and an empirical evaluation of Varimax, Entromin and its second-order approximation.

The organization of this thesis is as follows: Chapter 2 provides more context to factor analysis and related work, outlines the Varimax and Entromin rotation procedure, and introduces the mathematical notation used throughout this thesis; Chapter 3 presents the main theoretical results and discusses the connections to various statistical concepts; Chapter 4 examines the algorithms used to implement factor rotation methods and discusses the key computational considerations; Chapter 5 details the results of our empirical analyses of the factor rotation methods on synthetic and real datasets; and Chapter 6 summarizes our conclusions and highlights relevant ongoing and future work. Appendix A includes proofs for the main theoretical results; Appendix D includes additional details of the algorithms examined in Chapter 4; and Appendix E contains additional plots from the empirical analyses in Chapter 5.

Chapter 2

Background

In this chapter, we discuss the history of factor analysis and motivate why sparsity is desirable. We also introduce the latent factor model under which the statistical properties of the factor rotation methods are studied. We outline the general factor rotation procedure and provide a detailed overview of the Varimax and Entromin rotation criteria. We end the chapter with a brief overview of related work. The mathematical notation introduced in this chapter is used throughout this thesis.

2.1 Multiple factor analysis

Early forms of factor analysis developed in the field of psychology under key figures such as Louis Leon Thurstone, who authored the following quote in his explanation of the "factor problem":

> *It is the faith of all science that an unlimited number of phenomena can be comprehended in terms of a limited number of concepts or ideal constructs... To discover a scientific law is merely to discover that a man-made scheme serves to unify, and thereby to simplify, comprehension of a certain class of natural phenomena.*
> — Thurstone (1935, p. 44)

The first line in his explanation provides motivation for uncovering the presumed limited factors that explain a phenomenon of interest. The second line in his explanation can be interpreted as saying that any inferred factors are ultimately man-

made constructs intended for understanding and simplifying the unobserved underlying concepts. If the primary objective is to estimate factors that well-approximate the concepts, then the secondary objective is to make the factors as easy to comprehend as possible. To make progress towards the secondary objective, Thurstone (1935) proposed rotating the inferred factors to have, as much as possible, his *simple structure* where each factor is contained in one or more orthogonal hyperplanes. Several guidelines for diagnosing simple structure were introduced, such as every row of the rotated factor matrix having at least one zero and every pair of columns being close to orthogonal (Thurstone, 1947).

Thurstone (1954) also advocated for analytical methods over graphical methods for achieving simple structure. These analytical methods involve optimizing some criterion function, with different criteria having differing properties and returning potentially different representations of the factors. Several of these criteria were introduced shortly after Thurstone's initial proposition and remain in use today. Modern factor rotation methods generally all involve optimizing some criterion.

Thurstone imposed sparsity on the inferred factors to improve their interpretability. An additional useful consequence of such a constraint is that under certain conditions, the constraint also breaks the symmetry in the set of otherwise equivalent representations of the factors. It is this property that allows these rotation methods to be viewed through the lens of statistical inference.

2.2 Semi-parametric factor model

We examine the factor rotation methods under the same latent factor model described by Rohe and Zeng (2020).

Definition 1. Let $Z \in \mathbb{R}^{n\times k}$ and $Y \in \mathbb{R}^{d\times k}$ be latent factor matrices, and let $B \in \mathbb{R}^{k\times k}$ be an arbitrary latent matrix. Under the *semi-parametric factor model*, the observed data matrix $A \in \mathbb{R}^{n\times d}$ is assumed to have independent elements with expectation

$$\mathbb{E}[A|Z,Y] = ZBY^T \ . \tag{2}$$

Note that no assumptions are made about the structures of the matrices Z, Y and B under the semi-parametric factor model. The model suggests a two-step generative procedure where Z and Y are generated before the data matrix A is generated conditional on Z and Y. Working with the population matrix defined in Equation 2 allows us to remove the randomness from the second generative step. When studying the properties of factor rotation methods under the model, randomness from the first step can be removed by taking the expectation of the method output with respect to Z or Y.

In this thesis, we specifically focus on making inference on Z. However, assuming that Y satisfies all of the specified conditions, any results also hold for making inference on Y.

2.3 General factor rotation procedure

The general factor rotation procedure that we work with is a generalization of the Vintage Sparse PCA algorithm described by Rohe and Zeng (2020) that allows for any criterion. Note that the procedure uses the *Singular Value Decomposition* (SVD) in place of PCA. The mathematical relationship between PCA and SVD has been well-studied (Wall et al., 2003). Given a data matrix $A \in \mathbb{R}^{n \times d}$, SVD finds the decomposition

$$A = UDV^T$$

where $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{d \times d}$ are orthogonal matrices that contain the left and right singular vectors of A, respectively, and $D \in \mathbb{R}^{n \times d}$ is a rectangular diagonal matrix that contains the singular values. Rewriting A in Equation 1 in terms of its SVD gives

$$S = UD \ .$$

In other words, the principal component matrix S in PCA can also be obtained through the products of SVD. However, it will be more convenient to directly represent and estimate the factors using the (leading) singular vectors in U. From this

point onwards, we implicitly mean the singular vectors when we refer to the principal components.

Also, note that Entromin is only appropriate for finding orthogonal rotations (the orthogonal case) (Browne, 2001; Bernaards and Jennrich, 2005). Therefore, we do not consider rotations where the columns of the rotation matrix have norm one but are not orthogonal (the oblique case). Define the set of $k \times k$ orthogonal matrices as

$$\mathcal{O}(k) = \left\{ R \in \mathbb{R}^{k \times k} : R^T R = R R^T = I_k \right\} .$$

The procedure is then as follows:

Inputs: data matrix $A \in \mathbb{R}^{n \times d}$, desired number of dimensions k.

1. **Normalize** (optional). Define the row sums, row regularization parameter, and row normalization matrix as

$$\begin{aligned} A_{r\cdot} &= A\mathbf{1}_d \in \mathbb{R}^n , \\ \tau_r &= \frac{1}{n}\mathbf{1}_n^T A_{r\cdot} \in \mathbb{R} , \\ D_r &= \operatorname{diag}(A_{r\cdot} + \tau_r \mathbf{1}_n) \in \mathbb{R}^{n \times n} \end{aligned}$$

   and define corresponding column quantities $A_{\cdot c}$, τ_c, and D_c. Normalize the rows and columns of the data matrix A by computing

$$A' = D_r^{-\frac{1}{2}} A D_c^{-\frac{1}{2}} .$$

2. **Center** (optional). Define the row, column, and overall mean as

$$\begin{aligned} \widehat{\mu}_r &= \frac{1}{d} A\mathbf{1}_d \in \mathbb{R}^n , \\ \widehat{\mu}_c &= \frac{1}{n}\mathbf{1}_n^T A \in \mathbb{R}^d , \\ \widehat{\mu} &= \frac{1}{nd}\mathbf{1}_n^T A \mathbf{1}_d \in \mathbb{R} . \end{aligned}$$

Center the data matrix A (or A') by computing

$$A'' = A - \widehat{\mu}_r \mathbf{1}_d^T - \mathbf{1}_n \widehat{\mu}_c + \widehat{\mu} \mathbf{1}_n \mathbf{1}_d^T \ .$$

3. **SVD**. Apply SVD to A (A' if Step 1 is performed, A'' if Step 2 is performed) to obtain

$$A \approx \widehat{U} \widehat{D} \widehat{V}^T$$

where $\widehat{U} \in \mathbb{R}^{n \times k}$ and $\widehat{V} \in \mathbb{R}^{d \times k}$ contain the first k left and right singular vectors of A and $\widehat{D}$ contains the first k singular values.

4. **Maximize**. Given a matrix X to rotate, let $f_X(R)$ be the objective corresponding to some criterion for $R \in \mathcal{O}(k)$. Compute the optimal rotation matrices

$$R_{\widehat{U}} = \underset{R \in \mathcal{O}(k)}{\arg\max} \ f_{\widehat{U}}(R) \ ,$$
$$R_{\widehat{V}} = \underset{R \in \mathcal{O}(k)}{\arg\max} \ f_{\widehat{V}}(R) \ .$$

Note that f_X is defined to be maximized. For criteria such as Entromin that are intended to be minimized, f_X corresponds to the negative of the criterion.

5. **Estimate**. Compute the estimates of the latent matrices

$$\widehat{Z} = \sqrt{n} \widehat{U} R_{\widehat{U}} \ ,$$
$$\widehat{Y} = \sqrt{d} \widehat{V} R_{\widehat{V}} \ ,$$
$$\widehat{B} = \frac{1}{\sqrt{nd}} R_{\widehat{U}}^T \widehat{D} R_{\widehat{V}} \ .$$

6. **Recenter** (optional). Recenter the estimates by computing

$$\widehat{Z}' = \widehat{Z} + \sqrt{n} \mathbf{1}_n \widehat{\mu}_c \widehat{V} \widehat{D}^{-1} R_{\widehat{U}} \ ,$$
$$\widehat{Y}' = \widehat{Y} + \sqrt{d} \mathbf{1}_d \widehat{\mu}_r^T \widehat{U} \widehat{D}^{-1} R_{\widehat{V}} \ .$$

7. **Unnormalize** (optional). Unnormalize the estimates $\widehat{Z}$ and $\widehat{Y}$ (or $\widehat{Z}'$ and $\widehat{Y}'$

if the recentering in Step 6 is done) by computing

$$\widehat{Z}'' = D_r^{\frac{1}{2}} \widehat{Z} \ ,$$
$$\widehat{Y}'' = D_c^{\frac{1}{2}} \widehat{Y} \ .$$

Outputs: estimates of the latent matrices $\widehat{Z}$ ($\widehat{Z}'$, $\widehat{Z}''$), $\widehat{Y}$ ($\widehat{Y}'$, $\widehat{Y}''$) and $\widehat{B}$.

The centering and recentering steps (Step 2 and Step 6) determine whether $\widehat{Z}$ estimates the original factor matrix Z or the centered matrix $Z - \mathbb{E}[Z]$ (Rohe and Zeng, 2020). Also note that following Rohe and Zeng (2020), the unnormalization step (Step 7) is never performed in our empirical analyses even when the normalization step (Step 1) is performed.

The unrotated principal components in Figure 1 are obtained from the SVD in Step 3 and the rotated components are the outputs of Step 5.

2.4 Varimax

The Varimax criterion was first introduced by Kaiser (1958) and has since become one of the most commonly used criterion for factor rotations (as evident by being one of the few criteria implemented in base R). See the work by Rohe and Zeng (2020) for an extended discussion of Varimax.

Definition 2. The *Varimax* criterion has the form

$$v(U, R) = \sum_{j=1}^{k} \left(\frac{1}{n} \sum_{i=1}^{n} [UR]_{ij}^4 - \left(\frac{1}{n} \sum_{i=1}^{n} [UR]_{ij}^2 \right)^2 \right) \tag{3}$$

where $U \in \mathbb{R}^{n \times k}$ is the matrix to be rotated and $R \in \mathcal{O}(k)$. Its corresponding objective is

$$\underset{R \in \mathcal{O}(k)}{\arg\max} \ f_U^v(R) = \underset{R \in \mathcal{O}(k)}{\arg\max} \ v(U, R) \ . \tag{4}$$

⌟

For a matrix X, denote $X^{(2)}$ as the matrix obtained by squaring the elements of X,

i.e., $X_{ij}^{(2)} = X_{ij}^2$. The Varimax objective can be interpreted as maximizing the sum of sample column variances of the matrix $[UR]^{(2)}$. Intuitively, if UR is an orthonormal matrix where the entries are in the interval $[-1, 1]$, then the sample variance of a column is larger when the squared elements consist of values that are close to either zero or one as opposed to all being somewhere in between. Note that when U has orthonormal columns, the second term in Equation 3 is a constant and so only the first term is relevant for optimization as a function of R.

Rohe and Zeng (2020) showed that under certain assumptions, Varimax can be viewed as performing statistical inference on the latent factors Z and is expected to recover the matrix up to column permutations and sign-flips. The following definitions will be useful for interpreting the assumptions.

Definition 3. Let $X \in \mathbb{R}$ be a random variable. The n^{th} *central moment* of X is defined as

$$\mu_n = \mathbb{E}\left[(X - \mathbb{E}[X])^n\right] .$$

If X has n finite moments, the n^{th} *standardized moment* of X is defined as

$$\widetilde{\mu}_n = \frac{\mu_n}{(\sqrt{\mu_2})^n} .$$

If $\mathrm{Var}(X) = \mu_2 = 1$, then the n^{th} central moment is also the n^{th} standardized moment. For random variables $X_1, \ldots, X_k$, let $(\mu_n)_i$ and $(\widetilde{\mu}_n)_i$ denote the corresponding n^{th} moments of X_i for $i \in \{1, \ldots, k\}$. ⌟

In particular, the fourth standardized moment has a name and plays an important role in the assumptions for Varimax.

Definition 4. Let $X \in \mathbb{R}$ be a random variable with four finite moments. The fourth standardized moment $\widetilde{\mu}_4$ is the *kurtosis* of X. The random variable X and its distribution is said to be *leptokurtic* if $\widetilde{\mu}_4 > 3$, *mesokurtic* if $\widetilde{\mu}_4 = 3$, and *platykurtic* if $\widetilde{\mu}_4 < 3$. ⌟

The kurtosis of a distribution has interpretations in terms of the tails of the dis-

tribution (Westfall, 2014). A leptokurtic distribution has relatively heavier tails compared to a normal distribution while a platykurtic distribution has relatively lighter tails. Note that in some literature, kurtosis refers to the quantity $\widetilde{\mu}_4 - 3$. We refer to this quantity as the *excess kurtosis* and follow the definition given in Definition 4 for kurtosis.

The idenfication assumptions for Varimax (Rohe and Zeng, 2020) are given in Assumption 1.

Assumption 1 (Identification assumptions for Varimax)**.** The rows $Z_i \in \mathbb{R}^k$, $i \in \{1, \ldots, n\}$, of the latent matrix $Z \in \mathbb{R}^{n \times k}$ satisfy all of the following conditions:

(i) the rows $Z_1, \ldots, Z_n$ are i.i.d.,

(ii) each row Z_i consists of k independent random variables,

(iii) $\mathrm{Var}(Z_{ij}) = 1$ for all j, and

(iv) $(\widetilde{\mu}_4)_{ij} > 3$ for all j where $(\widetilde{\mu}_4)_{ij}$ is the kurtosis of Z_{ij}.

⌟

Denote the set of matrices that includes column permutations and sign-flips as

$$\mathcal{P}(k) = \{P \in \mathcal{O}(k) : P_{ij} \in \{-1, 0, 1\}\} \ .$$

Define Z_1 to be the first row of Z and $Z^o = Z_1 - \mathbb{E}[Z_1]$. Rohe and Zeng (2020) showed that under Assumption 1, for any nuisance rotation matrix $\widetilde{R} \in \mathcal{O}(k)$,

$$\underset{R \in \mathcal{O}(k)}{\arg\max} \ \mathbb{E}_{Z_1} \left[v(Z^o \widetilde{R}^T, R) \right] = \left\{ \widetilde{R} P : P \in \mathcal{P}(k) \right\} \ . \tag{5}$$

Equation 5 says that the under the specified assumptions, Varimax is expected to recover the true latent factors Z. Note that Condition (iii) of Assumption 1 is not restrictive as the columns of Z can be scaled without changing the definition of the population matrix in Equation 2 by correspondingly scaling B. The main condition of interest is Condition (iv) which says that the distribution of each factor

must be leptokurtic. The intuition here comes from Maxwell's theorem which says that multivariate normal distributions with independent components are the only distributions that are rotationally invariant (Maxwell, 1860; Feller, 1971). Every normal distribution is mesokurtic (Smith, 2021), and so any distribution that is leptokurtic must not be Gaussian. Therefore, any leptokurtic distribution must not be rotationally invariant and should be identifiable in theory.

2.5 Entromin criteria

Various criteria with similar forms have been introduced and referred to as Entromin (McCammon, 1970; Inagaki, 1993, 1994; Jennrich, 2004). While these criteria have existed for at least a few decades, there does not appear to have been much work in the literature studying their statistical properties. The various Entromin criteria are not necessarily equivalent in all contexts and so we introduce the few that we study in this thesis in the following sections. All Entromin criteria are formulated to be minimized as an objective.

2.5.1 Entromin

In this thesis, we mainly focus on the criterion used by Jennrich (2004) which we refer to as just the Entromin criterion. This particular criterion is generally easier to analyze compared to the other related criteria.

Definition 5. The *Entromin* criterion has the form

$$h(U,R) = -\sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^2 \log[UR]_{ij}^2 \ . \tag{6}$$

Its corresponding objective is

$$\underset{R\in\mathcal{O}(k)}{\arg\max}\ f_U^h(R) = \underset{R\in\mathcal{O}(k)}{\arg\max}\ -h(U,R) \ .$$

⌟

The criterion is closely related to the *Shannon entropy* (Shannon, 1948) which,

given a discrete random variable X with n possible outcomes, is defined as

$$H(X) = -\sum_{i=1}^{n} \mathbb{P}(X = i) \log \mathbb{P}(X = i) \ . \tag{7}$$

The Shannon entropy is maximized at $H(X) = \log n$ when $\mathbb{P}(X = i) = \frac{1}{n}$ for all i and minimized at $H(X) = 0$ when $\mathbb{P}(X = i) = 1$ and $\mathbb{P}(X = j) = 0$ for all $j \neq i$.[1] The name Entromin comes from viewing each column of the matrix $[UR]^{(2)}$ as the probabilities of a discrete distribution and minimizing the sum of column Shannon entropies, which should intuitively promote sparsity. While Entromin as a factor rotation method has not been well-studied, there has been some work formalizing the ties between the Shannon entropy and sparsity (e.g., Pastor et al., 2020).

2.5.2 McCammon entropy

One of the earlier forms of Entromin was introduced by McCammon (1970). The *McCammon entropy* criterion has the form

$$h_{\text{MC}}(U, R) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} \frac{[UR]_{ij}^2}{c_j} \log\left(\frac{[UR]_{ij}^2}{c_j}\right)}{\sum_{j=1}^{k} \frac{c_j}{c} \log\left(\frac{c_j}{c}\right)}$$

where the quantities

$$c_j = \sum_{i=1}^{n} [UR]_{ij}^2 \ , \qquad\qquad c = \sum_{j=1}^{k} c_j$$

are the squared column norms and the squared Frobenius norm of the rotated matrix UR, respectively. This criterion can be interpreted as a version of the Entromin criterion that adjusts for imbalanced columns when the norms are not identical across columns. Notice that in the orthogonal case, the rotation matrix R is norm-preserving and so the norms are constant. Hence, when the matrix U has orthonormal columns, the McCammon entropy criterion reduces to the Entromin criterion

[1] We follow the convention that $0 \log 0 = 0$.

up to a constant scaling factor and is equivalent as a factor rotation objective.

2.5.3 Inagaki entropy

Similar to how the McCammon entropy criterion adjusts for imbalanced columns, the criterion introduced by Inagaki (1993) can be viewed as a version of the Entromin criterion that adjusts for imbalanced rows. We refer to this criterion as the *Inagaki entropy* criterion.

Definition 6. The *Inagaki entropy* criterion has the form

$$h_{\text{IG}}(U, R) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{[UR]_{ij}^2}{r_i} \log \left(\frac{[UR]_{ij}^2}{r_i} \right) \tag{8}$$

where

$$r_i = \sum_{j=1}^{k} [UR]_{ij}^2$$

are the squared row norms of the rotated matrix UR. Its corresponding objective is defined as

$$\underset{R \in \mathcal{O}(k)}{\arg\max} \; f_U^{h_{\text{IG}}}(R) = \underset{R \in \mathcal{O}(k)}{\arg\max} \; -h_{\text{IG}}(U, R) \; . \tag{9}$$

⌟

Note that while the row norms are constant in the orthogonal rotation case, the Inagaki entropy criterion is not equivalent to the Entromin criterion when the rows of U are not constrained to have norm one. The Inagaki entropy criterion is also not equivalent to the Entromin criterion when the optional row normalization step in the procedure described in Section 2.3 is performed. The optional step normalizes the rows of the input data matrix A prior to taking its SVD, and only its leading k left singular vectors are retained to form the matrix U.

Although we focus on the properties of the Entromin criterion in this thesis, we will show that the Inagaki entropy criterion allows for a formal result that makes the connection between minimizing an Entromin criterion and recovering sparsity in the factors.

2.6 Related work

We end this chapter with a brief overview of related work in the literature.

This thesis is inspired by and therefore most closely related to the work by Rohe and Zeng (2020). Their work draws on classical results such as Maxwell's theorem to show that Varimax can be seen as performing statistical inference on the latent factors, and their results relate the conditions under which Varimax can identify the factors to the general goal of sparsity in factor analysis. We aim to provide similar results for Entromin in this thesis and hence the similarities between our work and theirs.

Rohe and Zeng (2020) also discuss the connection between the semi-parametric factor model and several modern factor models including Independent Component Analysis (Comon, 1994), various Stochastic Blockmodels (Holland et al., 1983) and Latent Dirichlet Allocation (Blei et al., 2003), all of which can be seen as involving some form of the semi-parametric factor model. While we examine the statistical properties of Entromin under the semi-parametric factor model, the model itself is not a focus of our work.

The idea of finding sparse representations of data via rotations exist more generally beyond the factor analysis literature where several rotation-based sparse PCA techniques have been proposed and studied (e.g., Zou et al., 2006; Hu et al., 2016; Chen and Rohe, 2020). However, the majority of these references do not specifically study rotations based on entropy. On the other hand, achieving sparsity through entropy-related measures via non-rotational techniques are common outside of factor analysis. For example, the *minimum description length* principle (Rissanen, 1978)—which has origins in information theory and has close ties with entropy—has been employed as an objective in statistical learning for obtaining sparse and compressed representations of data (e.g., Ramírez and Sapiro, 2011; Squires et al., 2019). In this thesis, we investigate the intersection of the two described types of literature by studying how Entromin finds sparse representations of data from a statistical perspective.

Chapter 3

Entromin from a statistical perspective

The main goals of our analysis of Entromin are

1. to provide a statistical explanation for the observation that Entromin tends to uncover sparser structure compared to Varimax, and
2. to define the conditions under which the observation holds generally.

Results analogous to those for Varimax given by Rohe and Zeng (2020) would be desirable. However, the form of the Entromin criterion given in Definition 5 does not easily allow itself to be studied using the same approach that Rohe and Zeng use to analyze Varimax. For example, the reasoning used to obtain Equation 5 involves expressing the inner matrix product summations in terms of moments of the latent factor distributions, but applying the same reasoning to the expectation of the Entromin criterion fails to reproduce similar quantities due to the logarithm in the criterion. Instead, we draw on the Varimax results for inspiration but otherwise take a different approach to study the statistical properties of Entromin.

In this chapter, we show that Varimax can be viewed as a first-order approximation of Entromin under certain conditions. By generalizing this connection to a family of Entromin approximations, we are able to take steps towards our goals above by

studying other members of this family. We derive the identifiability conditions for a second-order approximation of Entromin and explain why it is more natural to express the conditions in terms of the cumulants rather than the moments of the latent factor distributions. We discuss the ties between the identifiability conditions and sparsity, and show that a more meaningful connection between Entromin and sparsity can be made through the Inagaki entropy criterion. We end the chapter with a comment on how our results for the second-order approximation may generalize to the family of approximations as well as to Entromin itself.

3.1 Varimax as a first-order approximation of Entromin

The logarithm in the Entromin criterion is mathematically challenging to work with. However, by rewriting the logarithm in terms of an infinite sum, we obtain a form of the Entromin criterion that is statistically more convenient to study. Taking a Taylor expansion of the logarithm in Equation 6 around one leads to the form

$$h(U,R) = \sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^2 \left(\sum_{q=1}^{\infty} \frac{(-1)^q}{q} \left([UR]_{ij}^2 - 1\right)^q \right) .$$

Note that this expansion is only valid when $[UR]_{ij}^2 \in [0,2]$ for all $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,k\}$, as in the case when U has orthonormal columns and $R \in \mathcal{O}(k)$. This form of the Entromin criterion also produces approximations that can be obtained by truncating the infinite sum.

Definition 7. For $N \in \mathbb{N}$, an *N^{th}-order approximation* of the Entromin criterion is given by

$$h_N(U,R) = \sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^2 \left(\sum_{q=1}^{N} \frac{(-1)^q}{q} \left([UR]_{ij}^2 - 1\right)^q \right) \tag{10}$$

and its corresponding objective is defined as

$$\underset{R\in\mathcal{O}(k)}{\arg\max}\ f_U^{h_N}(R) = \underset{R\in\mathcal{O}(k)}{\arg\max}\ -h_N(U,R)\ . \tag{11}$$

⌟

The approximations h_N for $N \in \mathbb{N}$ form a family of finite-order approximations of Entromin with $h_N \to h$ as $N \to \infty$. The first-order member of the family is particularly notable. We show that the first-order Entromin objective is equivalent to the Varimax objective, and hence Varimax can be viewed as a first-order approximation of Entromin. To see this, we first introduce another related criterion that will bridge the two criteria.

Definition 8. The *Quartimax* criterion introduced by Neuhaus and Wrigley (1954) has the form

$$q(U, R) = \sum_{i=1}^{n} \sum_{j=1}^{k} [UR]_{ij}^4$$

and its corresponding objective is

$$\underset{R \in \mathcal{O}(k)}{\arg\max} \; f_U^q(R) = \underset{R \in \mathcal{O}(k)}{\arg\max} \; q(U, R) \; . \tag{12}$$

⌟

The following proposition is a well-known result that states that the Varimax objective is equivalent to the Quartimax objective when the columns of U are orthonormal. The result is useful and so we restate it here for convenience.

Proposition 1 (Kaiser, 1958)**.** *Let the Varimax objective and the Quartimax objective be defined as in Equation 4 and Equation 12, respectively. If the matrix U has orthonormal columns, then*

$$\underset{R \in \mathcal{O}(k)}{\arg\max} \; f_U^v(R) = \underset{R \in \mathcal{O}(k)}{\arg\max} \; f_U^q(R) \; .$$

The proof of Proposition 1 is given in Section A.1 in the Appendix.

It can also be shown that minimizing the criterion of the first-order member in the approximating family is equivalent to maximizing the Quartimax criterion, which is what the following result states.

Proposition 2. *Let the Quartimax objective and the objective of a first-order approximation ($N = 1$) of Entromin be defined as in Equation 12 and Equation 11, respectively. If the matrix U has orthonormal columns, then*

$$\underset{R \in \mathcal{O}(k)}{\arg\max}\ f_U^{h_1}(R) = \underset{R \in \mathcal{O}(k)}{\arg\max}\ f_U^{q}(R)\ .$$

The proof of Proposition 2 is given in Section A.2 in the Appendix.

The connection between Varimax and Entromin then immediately follows from Proposition 1 and Proposition 2, and is formally stated in the following corollary.

Corollary 3. *Let the Varimax objective and the objective of a first-order approximation $(N = 1)$ of Entromin be defined as in Equation 4 and Equation 11, respectively. If the matrix U has orthonormal columns, then*

$$\underset{R \in \mathcal{O}(k)}{\arg\max}\ f_U^{v}(R) = \underset{R \in \mathcal{O}(k)}{\arg\max}\ f_U^{h_1}(R)\ .$$

3.2 Second-order approximation of Entromin

Corollary 3 says that when U is orthonormal, Varimax is equivalent to a first-order approximation of Entromin. It is then natural to consider whether other members of the approximation family share similar statistical properties to Varimax. If the properties of a finite-order member can be identified and generalized, we may be able to derive the corresponding properties of Entromin by taking the order to the limit. As a first step in this process, we start with the next simplest member of the family after Varimax: the second-order approximation.

The criterion of the second-order member of the Entromin approximating family—which from this point onwards we refer to as *Entromin*$_2$—is given by

$$h_2(U, R) = \sum_{i=1}^{n} \sum_{j=1}^{k} \left(\frac{1}{2}[UR]_{ij}^6 - 2[UR]_{ij}^4 + \frac{3}{2}[UR]_{ij}^2 \right)\ . \tag{13}$$

The derivation of Equation 13 is given in Section B.1 in the Appendix. The above form is already informative in that it suggests a few properties for the members of the approximating family. These properties include that

1. expectations of higher-order members will involve higher-order moments of the latent factor distributions which may not be as interpretable, and
2. the alternating signs of the terms may complicate mathematical arguments for the optimal rotation R.

Nonetheless, the Entromin$_2$ criterion is still simple enough that an identification result analogous to that of Equation 5 can be derived and interpreted. Assumption 2 lists the conditions under which the latent factor matrix Z can be recovered (up to column permutations and sign-flips) by minimizing the Entromin$_2$ criterion.

Assumption 2 (Identification assumptions for second-order Entromin approximation)**.** The rows $Z_i \in \mathbb{R}^k$, $i \in \{1, \dots, n\}$, of the matrix $Z \in \mathbb{R}^{n \times k}$ satisfy all of the following conditions:

(i) the rows $Z_1, \dots, Z_n$ are i.i.d.,

(ii) each row Z_i consists of k independent random variables,

(iii) $\mathrm{Var}(Z_{ij}) = 1$ for all j,

(iv) $\mathbb{E}[Z_{ij}^3] = 0$ for all j, and

(v) if $\frac{31}{15} < (\widetilde{\mu}_4)_{ij} \leq 3$, the inequality

$$(\widetilde{\mu}_6)_{ij} < 15(\widetilde{\mu}_4)_{ij} - 30 \tag{14}$$

is satisfied for all j. Otherwise, if $(\widetilde{\mu}_4)_{ij} > 3$, the inequality

$$(\widetilde{\mu}_6)_{ij} < 4(\widetilde{\mu}_4)_{ij} + 3 \tag{15}$$

is satisfied for all j.

⌟

Conditions (i)–(iii) in Assumption 2 are identical to those in Assumption 1 for Varimax. Condition (iv) says that the latent factor distributions are all symmetric, and is only necessary for mathematical convenience in the proof of the identifiability result. If the distribution of any factor is skewed, the distribution must be non-Gaussian and so the latent factors should be identifiable by Maxwell's theorem.

Condition (v) is analogous to the leptokurtic condition in Assumption 1. In both the leptokurtic and non-leptokurtic cases, Condition (v) is most intuitive when viewed as a test of whether the factor distributions are consistent with that of a normal distribution. The kurtosis and sixth standardized moment of any normal distribution are $\widetilde{\mu}_4 = 3$ and $\widetilde{\mu}_6 = 15$, respectively (Smith, 2021). Thus, Condition (v) does not hold for any normal distribution. If the distribution of any factor is assumed to be mesokurtic or platykurtic, then Equation 14 says that Entromin$_2$ may still identify the latent factors as long as the sixth standardized moment is sufficiently small. Conversely, if a factor distribution is assumed to have $\widetilde{\mu}_6 \geq 15$, then Equation 15 says that the latent factors may still be identified as long as the distribution is "sufficiently" leptokurtic. Note that the lower bound in the non-leptokurtic case comes from standardized moments being bounded below by one (Westfall, 2014), and so Equation 14 cannot be satisfied if $\widetilde{\mu}_4$ is too small. The conditions on the moments of the factor distributions imposed by Condition (v) are organized in Table 1.

	$\mathbb{E}\left[(Z_\ell^o)^6\right]$		
$\mathbb{E}\left[(Z_\ell^o)^4\right]$	< 15	$= 15$	> 15
$\leq \frac{31}{15}$	No	No	No
$\in \left(\frac{31}{15}, 3\right)$	Equation 14	No	No
$= 3$	Yes	No	No
> 3	Yes	Yes	Equation 15

Table 1: Conditions on the moments of the latent factors Z_ℓ^o, $\ell \in \{1, \ldots, k\}$, under which Entromin$_2$ is able to recover them. **Yes** (**No**) denotes a condition under which the factors are always (never) identifiable. An equation denotes an inequality that must be satisfied in order to have identifiability.

Table 1 makes it clear that Condition (v) is neither a direct extension nor a complete relaxation of the leptokurtic condition for Varimax. Whereas Varimax is expected to be able to identify the latent factors when the factor distributions are leptokurtic, Entromin$_2$ is only expected to be able to do so when the sixth standardized moment is sufficiently small relative to the kurtosis. However, in the case that a distribution is not leptokurtic and the kurtosis is not too small, then Entromin$_2$ may still show promise.

Under Assumption 2, an identification result analogous to that for Varimax can be derived for Entromin$_2$.

Theorem 4. *Consider the criterion for the second-order approximation of Entromin defined in Equation 13. Suppose that the latent matrix $Z \in \mathbb{R}^{n\times k}$ satisfies the conditions in Assumption 2. Let $Z_1 \in \mathbb{R}^k$ be the first row of Z, and define $Z^o = Z_1 - \mathbb{E}[Z_1]$. Then for any nuisance rotation matrix $\tilde{R} \in \mathcal{O}(k)$,*

$$\underset{R\in\mathcal{O}(k)}{\arg\min}\ \mathbb{E}_{Z_1}\left[h_2(Z^o\tilde{R}^T, R)\right] = \left\{\tilde{R}P : P \in \mathcal{P}(k)\right\}\ .$$

The proof of Theorem 4 is given in Section A.3 in the Appendix. Theorem 4 says that under the specified assumptions, we can expect Entromin$_2$ to recover the latent factor matrix up to column permutations and sign-flips.

We comment that so far, we have not been able to find nor construct a symmetric, non-leptokurtic distribution that satisfies Equation 14. In particular, members of the non-leptokurtic family of generalized normal distributions (Dytso et al., 2018)—which includes the normal and uniform distributions—do not satisfy the condition. It may be the case that the condition is theoretically impossible under the given assumptions. Due to time constraints on this thesis, we leave further investigation of this case for future work.

3.3 Expressing conditions in terms of cumulants

As seen in Assumption 2, the conditions on the moments of the factor distributions become increasingly complex and start to lose interpretability when higher-order

moments are introduced. However, it appears that there is a more natural expression of these conditions in terms of the cumulants of the factor distributions.

Definition 9. Let X be a random variable with a moment-generating function. The *cumulant-generating function* (CGF) of X is defined as

$$K_X(t) = \log \mathbb{E}_X \left[e^{tX} \right] \ . \tag{16}$$

The CGF can be represented in the form of a series

$$K_X(t) = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!} \tag{17}$$

where κ_n is the n^{th} *cumulant* of X. ⌟

The series representation of the CGF in Equation 17 is obtained from Equation 16 by rewriting the inside of the expectation as a power series, rewriting the outer logarithm as a Taylor series and then collapsing the two summations into one by collecting the coefficients (the cumulant) for each power of t. The first six cumulants of any centered random variable are

$$\begin{aligned}
\kappa_1 &= 0 \ , \\
\kappa_2 &= \mu_2 \ , \\
\kappa_3 &= \mu_3 \ , \\
\kappa_4 &= \mu_4 - 3\mu_2^2 \ , \\
\kappa_5 &= \mu_5 - 10\mu_3\mu_2 \ , \\
\kappa_6 &= \mu_6 - 15\mu_4\mu_2 - 10\mu_3^2 + 30\mu_2^3
\end{aligned}$$

where μ_n is the n^{th} central moment as defined in Definition 3. It can then be seen that under the conditions specified in Assumption 1 for Varimax, for any given

row i with k independent elements, $(\mu_2)_{ij} = 1$ for all $j \in \{1, \ldots, k\}$ and so

$$\begin{aligned}(\kappa_4)_{ij} &= (\mu_4)_{ij} - 3(\mu_2)_{ij}^2 \\ &= (\mu_4)_{ij} - 3 \\ &= (\widetilde{\mu}_4)_{ij} - 3\end{aligned}$$

for all j. Hence, the leptokurtic condition for Varimax is equivalent to the condition $(\kappa_4)_{ij} > 0$ for all j. Similarly, under the conditions specified in Assumption 2 for Entromin$_2$, $(\mu_2)_{ij} = 1$ and $(\mu_3)_{ij} = 0$ for all $j \in \{1, \ldots, k\}$ and so

$$\begin{aligned}(\kappa_6)_{ij} &= (\mu_6)_{ij} - 15(\mu_4)_{ij}(\mu_2)_{ij} - 10(\mu_3)_{ij}^2 + 30(\mu_2)_{ij}^3 \\ &= (\mu_6)_{ij} - 15(\mu_4)_{ij} + 30 \\ &= (\widetilde{\mu}_6)_{ij} - 15(\widetilde{\mu}_4)_{ij} + 30 \ .\end{aligned}$$

for all j. Thus, the condition

$$(\widetilde{\mu}_6)_{ij} < 15(\widetilde{\mu}_4)_{ij} - 30$$

for when $(\widetilde{\mu}_4)_{ij} \leq 3$ is equivalent to the condition $(\kappa_6)_{ij} < 0$. Although not as easy to interpret, the condition

$$(\widetilde{\mu}_6)_{ij} < 4(\widetilde{\mu}_4)_{ij} + 3$$

in the leptokurtic case can also be written as a linear combination of cumulants given by

$$(\kappa_6)_{ij} + 11(\kappa_4)_{ij} < 0 \ .$$

The view of the conditions being a test for normality becomes more intuitive when the conditions are expressed in terms of cumulants. A random variable is normally distributed if and only if its third and higher-order cumulants are zero (Willink, 2008). The identifiability conditions for Entromin$_2$ state that if the fourth or sixth cumulants are non-zero (in a certain direction), then the latent factors are likely not Gaussian and so it is expected that the nuisance rotation can be identified.

3.4 Identifiability of sparse factors

The identifiability results state the conditions under which the latent factors can be recovered, but it may not be obvious how these results tie into sparsity in the factors. We make the connection by examining the effect of sparsity on the identifiability conditions.

Rohe and Zeng (2020) showed that the assumptions under which Varimax is expected to recover the factors are satisfied when the factor distributions are sufficiently sparse. Specifically, they showed that for any random variable X with four finite moments, if $\frac{5}{6} < \mathbb{P}(X = 0) < 1$ then X is leptokurtic. This result suggests that sparse distributions are more likely to satisfy the identifiability conditions, and hence it is likely that there exists a sparse configuration of the factors that Varimax is expected to recover. We build on their result and show that a weaker sparsity requirement is possible by further assuming that X has expectation zero. In the context of factor rotations, this additional assumption is unrestrictive as it can be achieved by centering the data matrix.

We first give the following proposition that makes use of the fact that any random variable X can be rewritten as the product of its independent zero and non-zero parts, i.e., $X \overset{\text{d}}{=} BY$ where $B \sim \text{Bernoulli}(p)$ for $p = 1 - \mathbb{P}(X = 0)$, Y is equal in distribution to $X|X \neq 0$, and $\overset{\text{d}}{=}$ denotes equality in distribution.

Proposition 5. *Let X be a random variable that has four finite moments with $\mathbb{E}[X] = 0$. Rewrite $X \overset{d}{=} BY$ where $B \sim \text{Bernoulli}(p)$ for $p = \mathbb{P}(X \neq 0)$ and where Y is equal in distribution to $X|X \neq 0$. Let $\widetilde{\mu}_4$ be the kurtosis of Y. Then X is leptokurtic if and only if*

$$\mathbb{P}(X \neq 0) < \frac{1}{3}\widetilde{\mu}_4 \, .$$

The proof of Proposition 5 is given in Section A.4 in the Appendix. The proposition can be interpreted as saying that X is leptokurtic if its underlying non-zero part Y is leptokurtic, and that otherwise X needs to have at least a certain degree of sparsity in order for it to be leptokurtic.

The following corollary is our extension of the result by Rohe and Zeng (2020). The proof of the corollary immediately follows from Proposition 5 and the fact that the kurtosis of any distribution is bounded below by one (Pearson, 1916).

Corollary 6. *Any random variable X that has four finite moments with $\mathbb{E}[X] = 0$ and that has $\frac{2}{3} < \mathbb{P}(X = 0) < 1$ is leptokurtic.*

Analogous results can be derived for the identifiability assumptions for Entromin$_2$ based on similar reasoning. However, the resulting restriction on $\mathbb{P}(X = 0)$ is not as interpretable due to being an interval in terms of the fourth and sixth standardized moments of the non-zero part of X. The following result for the non-leptokurtic case is only included here for completeness but otherwise we do not spend too much effort trying to interpret its conditions.

Proposition 7. *Let X be a random variable that has six finite moments and is not leptokurtic. Rewrite $X \overset{d}{=} BY$ where $B \sim \text{Bernoulli}(p)$ for $p = 1 - \mathbb{P}(X = 0)$ and where Y is equal in distribution to $X|X \neq 0$. Let $\widetilde{\mu}_n$ be the n^{th} standardized moment of Y. If the conditions*

$$\widetilde{\mu}_6 < \frac{5}{3}\widetilde{\mu}_4^2 \tag{18}$$

and

$$\frac{\widetilde{\mu}_4}{3} \leq p < \frac{\widetilde{\mu}_4}{4}\left(1 + \sqrt{1 - \frac{8\widetilde{\mu}_6}{15\widetilde{\mu}_4^2}}\right) \tag{19}$$

both hold, then X satisfies the inequality

$$-\frac{\mathbb{E}\left[(X - \mathbb{E}[X])^6\right]}{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]^3} + 15\frac{\mathbb{E}\left[(X - \mathbb{E}[X])^4\right]}{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]^2} - 30 > 0\,.$$

The proof of Proposition 7 is given in Section A.5 in the Appendix. Equation 18 is only necessary to ensure that the interval in Equation 19 is valid. The lower bound in Equation 19 is a consequence of Proposition 5 in conjunction with the assumption that X is not leptokurtic. The upper bound on the probability of X being non-zero is the main quantity of interest, but it is difficult to interpret due to being a non-linear quantity of the standardized moments of Y.

3.5 Recovering sparsity through Entromin

For making a more meaningful connection between minimizing an Entromin criterion and recovering sparsity in the factors, we find that the Inagaki entropy criterion described in Section 2.5.3 has useful statistical properties. In particular, the expected Inagaki entropy for a vector $X \in \mathbb{R}^k$ of k independent standard normal random variables has a closed form (Lemma 11 in Section A.6.1 in the Appendix). A sparse version of X can also be considered where independent random variables $B_j \sim \text{Bernoulli}(p)$, $p \in [0,1]$, $j \in \{1,\ldots,k\}$, are introduced to obtain the vector

$$Y \overset{\text{d}}{=} (X_1 B_1, \ldots, X_k B_k) \; .$$

The expected Inagaki entropy of Y also has a closed form (Lemma 12 in Section A.6.3 in the Appendix). The following corollary then says that the expected Inagaki entropy of Y is always less than that of X for $p < 1$.

Corollary 8. *Let the Inagaki entropy criterion be defined as in Equation 8. Let $X \in \mathbb{R}^k$ be a vector of k i.i.d. standard normal random variables and let $B \in \{0,1\}^k$ be a vector of k i.i.d. $\text{Bernoulli}(p)$ random variables where $p \in [0,1)$. Define $Y_j = X_j B_j$ for $j \in \{1,\ldots,k\}$ and $Y = (Y_1,\ldots,Y_k)$. Then*

$$\mathbb{E}_Y \left[h_{IG}(Y, I_k) \right] < \mathbb{E}_X \left[h_{IG}(X, R) \right]$$

for any rotation matrix $R \in \mathcal{O}(k)$.

The proof of Corollary 8 follows from a quick inspection of the expected entropies given in Lemma 11 and Lemma 12. The significance of the result is that the smaller expected Inagaki entropy is entirely a consequence of the sparsity assumption. Therefore, it may be reasonable to assume that a sparse factor representation is more likely to exist and be recovered by minimizing the expected Inagaki entropy criterion when the factor distributions are inherently sparse. The intuition is made more obvious in Figure 2, which shows the expected entropy of Y as a function of p for $k \in \{2, 10, 50\}$. Notice that each entropy curve is monotone increasing with p. Hence, sparser distributions are associated with smaller expected Inagaki entropies.

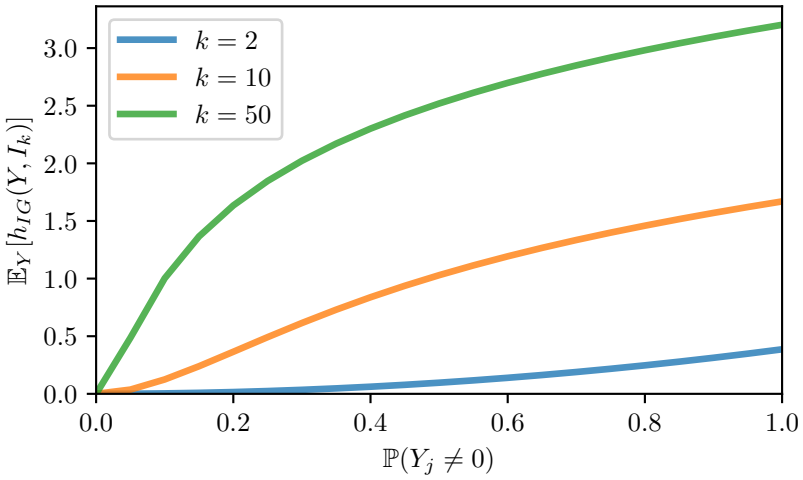

Figure 2: The expected Inagaki entropy for a sparse, i.i.d. standard normal vector Y for several numbers of factors k and varying levels of sparsity controlled by $\mathbb{P}(Y_j \neq 0)$, $j \in \{1, \ldots, k\}$. The expected entropy is equal to that of a non-sparse vector when $\mathbb{P}(Y_j \neq 0) = 1$.

The following theorem generalizes the idea in Corollary 8 to any factor distribution that is sufficiently sparse.

Theorem 9. *Let the Inagaki entropy criterion be defined as in Equation 8. Let $X \in \mathbb{R}^k$ be a vector of k i.i.d. standard normal random variables. Suppose that the rows of the latent factor matrix $Z \in \mathbb{R}^{n \times k}$ are i.i.d. and that the columns are independent. Let $Z^o \in \mathbb{R}^k$ be the first row of Z. If for all $j \in \{1, \ldots, k\}$, Z_j^o satisfies*

$$0 < \mathbb{P}(Z_j^o \neq 0) \leq 1 - \left(1 - \frac{e\left(\psi\left(\frac{k}{2} + 1\right) - \psi\left(\frac{1}{2} + 1\right)\right)}{k}\right)^{\frac{1}{k-1}}$$

where ψ is the digamma function, then

$$\mathbb{E}_{Z^o}\left[h_{IG}(Z^o, I_k)\right] \leq \mathbb{E}_X\left[h_{IG}(X, R)\right]$$

for any rotation matrix $R \in \mathcal{O}(k)$.

The proof of Theorem 9 is given in Section A.6 in the Appendix. The theorem states that the expected Inagaki entropy for any sufficiently sparse latent factors is less than that of non-sparse normally distributed factors (which can be standardized by scaling and centering). From the intuition developed through Corollary 8 and Figure 2, it is reasonable to expect that maximizing the Inagaki entropy objective defined in Equation 9 (minimizing the Inagaki entropy criterion) is associated with finding a sparse configuration of the factors.

The theorem may be partially viewed as an identifiability result by appealing to Maxwell's theorem. Minimizing the Inagaki entropy objective may be seen as trying to move away from a Gaussian-like representation of the factors, and hence there is potential to identify the latent factors if the distributions are truly not Gaussian. However, the theorem is not a true identifiability result as it makes no claims about the possible non-trivial rotations that produce an even smaller expected entropy for the given factors.

Figure 3 shows the upper bound of $\mathbb{P}(Z_j^o \neq 0)$ given in Theorem 9 as a function of the number of factors k. As the number of factors grow, the sparsity condition on the factor distributions also becomes more strict where the probability of a non-zero value must be very small for moderate k. We note that the given upper bound is overly conservative due to time constraints on this thesis, and that a bound with weaker sparsity restrictions should be possible.

It is worth mentioning that Theorem 9 also resembles a statement about the *maximum entropy distribution* under the given assumptions. We provide a brief discussion of this perspective in Appendix C.

3.6 Higher-order approximations of Entromin

Although only a first and second-order approximation of Entromin have been studied so far in (Rohe and Zeng, 2020) and in this thesis, the results that have been

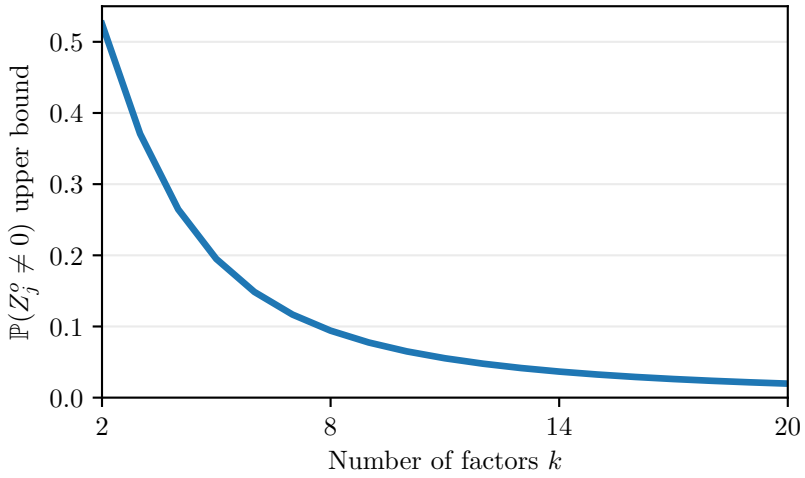

Figure 3: The upper bound restriction on $\mathbb{P}(Z_j^o \neq 0)$ that must be satisfied for Theorem 9 to hold as the number of latent factors k increases.

obtained have notable implications regarding what can be expected from higher-order approximations of Entromin. From the identifiability results of Varimax and Entromin$_2$, it is expected that the conditions under which the n^{th}-order approximation of Entromin is identifiable will only grow increasingly complex with n in terms of both the number of involved moments and their interpretation. It would also be reasonable to assume that the conditions for higher-order approximations will not be complete relaxations of the conditions for lower-order approximations similar to the relationship between the assumptions of Entromin$_2$ and Varimax.

Furthermore, we expect that the conditions for higher-order approximations will have natural expressions in terms of the cumulants of the latent factors. This can be seen by inspecting the relevant summations that give rise to the cumulant connection. As described in Section 3.3, the expansion of the CGF in Equation 16 before collecting the coefficients of terms with the same exponent has the form

$$K_X(t) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} \left(\sum_{\ell=1}^{\infty} \mathbb{E}_X \left[X^\ell \right] \frac{t^\ell}{\ell!} \right)^n$$

where the inner ℓ-sum is obtained from expanding the inside of the expectation and the outer n-sum is obtained from expanding the outer logarithm. The cumulants are then defined as the collected coefficients of the same powers of t. From a similar perspective, the expectation of the Entromin criterion in Equation 6 with respect to a row random variable $Z \in \mathbb{R}^k$ can be rewritten as

$$\begin{aligned}\mathbb{E}_Z\left[h(Z,R)\right] &= -\sum_{j=1}^{k}\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n}\mathbb{E}_Z\left[[ZR]_j^2\left([ZR]_j^2-1\right)^n\right] \\ &= -\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n}\sum_{j=1}^{k}\mathbb{E}_X\left[X_j^2\left(X_j^2-1\right)^n\right]\end{aligned}$$

where the n-sum is obtained from expanding the logarithm and the second equality takes $X_j = [ZR]_j$. Comparing this form of the expected Entromin criterion to the expansion of the CGF makes it easy to see their similarities from which the cumulant connection arises. The expanded CGF obtains moments of X from the inner power series, while the expanded expectation of the Entromin criterion obtains (even) moments of X_j, $j \in \{1,\ldots,k\}$, by raising the inner polynomial to a power before taking its expectation. The comparison also highlights the differences where in particular, the finite-order polynomial in the inner sum of the expanded expected Entromin criterion implies that higher-order cumulants will be missing certain moments. However, as observed with Entromin$_2$, the cumulant connection can likely still be made, and it may be reasonable to assume that the more complicated conditions are due to these missing moments.

Currently, it is unclear what the above conjectures imply for Entromin. If the generalizations hold for all finite-order approximations, then it is tempting to assume the same holds for when the order is taken to the limit. However, the results would then suggest that the identifiability conditions for Entromin are unfathomably complex to satisfy. This does not appear to be the case in practice where Entromin is observed to return better results compared to its first and second-order approximations. Possible explanations include that the generalizations do not hold in the limit, or that the identifiability assumptions do not generally hold for data in practice. We leave further investigation of these ideas for future work.

Chapter 4

Algorithms for orthogonal factor rotations

While the results in Chapter 3 provide theoretical guarantees for the factor rotation methods, it is also important to consider the practical aspects and how the procedure in Section 2.3 will be implemented. In this chapter, we provide an overview of the two iterative algorithms historically used to implement orthogonal rotations. We also show that Varimax and Entromin$_2$ have convexity properties that provide them with a practical advantage over Entromin. We end the chapter with a brief comment on the computational complexity of factor rotation methods.

4.1 Pairwise optimization algorithm

For early forms of Varimax and Entromin, the objective optimization step (Step 4 of the procedure in Section 2.3) was originally implemented using what we refer to as the *pairwise optimization algorithm* (Kaiser, 1958; Nemec and Peron, 1979). Each iteration of the algorithm involves iterating through every pair of columns in the input factor matrix and performing multiple smaller two-dimensional rotations that each do not decrease the value of the objective. The algorithm terminates when every two-dimensional optimization problem in a single iteration leaves the columns unchanged. For columns $X_j \in \mathbb{R}^n$ and $X_\ell \in \mathbb{R}^n$, let $[X_j, X_\ell] \in \mathbb{R}^{n\times 2}$ denote the matrix formed by horizontally stacking the columns. The basic framework

of the pairwise optimization algorithm is shown in Algorithm 1.

Algorithm 1: (Naive) pairwise optimization

Input : $n \times k$ matrix X
Output: rotated $n \times k$ matrix X

1 **while** *not yet converged* **do**
2 **foreach** *pair of columns* (X_j, X_ℓ), $j, \ell \in \{1, \ldots, k\}$, $j < \ell$ **do**
3 find optimal rotation matrix $R_{[X_j, X_\ell]}$;
4 $X_j, X_\ell \leftarrow [X_j, X_\ell] R_{[X_j, X_\ell]}$;
5 **end**
6 **end**
7 return X;

The success of the algorithm hinges on the fact that every pairwise rotation does not decrease the value of the objective, and that any high-dimensional rotation can be represented by a sequence of two-dimensional rotations (Richard et al., 2011; Golub and Ortega, 2014). The algorithm is considered greedy as it may find a suboptimal rotation for non-convex objectives. It is recommended to run the algorithm multiple times and starting with random initial rotations of the input matrix. However, suboptimal rotations are not a problem for convex objectives like Varimax, and the rotations found even for non-convex objectives like Entromin appear to be quite reasonable in practice. We discuss this point further in Section 4.3.

The two components of the pairwise optimization algorithm that determine its performance are the convergence criterion in Line 1 of Algorithm 1 and the method of finding the optimal rotation in Line 3. Possible choices for the convergence criteria include the relative increase of the objective and the value of the largest optimal rotation angle across all pairs of columns. Small values suggest convergence for both of these criteria, and so a convergence threshold below which the algorithm terminates can be selected.

For finding the optimal rotation in Line 3, Varimax is advantageous in that a closed form solution for the rotation angle exists (Kaiser, 1958). For Entromin and its approximations, a line search for the optimal rotation angle is necessary. While

early references pre-dating the rise of numerical optimization techniques suggest discretizing and searching a bounded interval for the optimal angle (Nemec and Peron, 1979), modern techniques such as gradient descent and Newton-Raphson are likely to be more efficient (assuming that the derivatives exist). Section D.1 in the Appendix derives the gradients that can be used in these numerical methods.

One change to the pairwise optimization algorithm that we found to generally improve performance is to modify the sequence of column pairs to rotate. The sequence specified in Line 2 of Algorithm 1 requires solving $\frac{k(k-1)}{2}$ optimization problems in every iteration. However, empirically we found that these optimizations are often wasteful as many of these optimizations do not increase the objective by a significant margin. We instead suggest pre-computing the gradient for each pair of columns and selecting the column pair (j, ℓ) with the largest absolute gradient to optimize. After optimizing the pair, the pre-computed gradients of every pair that involves either j or ℓ are updated, which requires at most $2k$ gradient calculations. The next iteration then chooses the next pair of columns to optimize based on the updated gradients. The modified algorithm is shown in Algorithm 2.

While the pairwise optimization algorithm works and is simple to understand, we find that the line search for the optimal angle required by Entromin and Entromin$_2$ is a major computational bottleneck in obtaining good performance as the objectives appear to have high curvature. Also, modern orthogonal rotation methods are typically implemented using the algorithm discussed in the following section. For these reasons, we do not consider the pairwise optimization algorithm in the empirical analyses in Chapter 5.

4.2 Gradient projection algorithm

Most modern orthogonal rotation methods are implemented using the *gradient projection algorithm* (Jennrich, 2001). The gradient projection algorithm is a form of projected gradient descent where the rotation matrix is first updated by taking a step in the direction of the gradient of the objective, and then projected back onto the orthogonal manifold. The gradient projection algorithm is outlined in Algorithm 3.

Algorithm 2: Pairwise optimization

Input : $n \times k$ matrix X, objective function f_X
Output: rotated $n \times k$ matrix X

```
1  initialize k × k matrix of gradients G;
2  foreach pair of columns (X_j, X_ℓ), j, ℓ ∈ {1, ..., k}, j < ℓ do
3  |  G_jℓ ← ∇f_[X_j,X_ℓ](I_k);
4  end
5  while not yet converged do
6  |  j, ℓ ← arg max_{j,ℓ} |G_jℓ|;
7  |  find optimal rotation matrix R_[X_j,X_ℓ];
8  |  X_j, X_ℓ ← [X_j, X_ℓ]R_[X_j,X_ℓ];
9  |  G_jℓ ← ∇f_[X_j,X_ℓ](I_k);
10 |  for m ∈ {1, ..., k} do
11 |  |  if m > j then
12 |  |  |  G_jm ← ∇f_[X_j,X_m](I_k);
13 |  |  end
14 |  |  if m < ℓ then
15 |  |  |  G_mℓ ← ∇f_[X_m,X_ℓ](I_k);
16 |  |  end
17 |  end
18 |  return X;
19 end
```

The key steps of Algorithm 3 include the update step in Line 5 and the projection step in Line 6. Projection of the updated matrix $R + \alpha G$ onto the orthogonal group (a Stiefel manifold) is performed using the products of the SVD (Manton, 2002). As the projected matrix may actually lead to a decrease in the objective if the step size α is too large, the update and projection steps are repeated with a continually halved step size until the resulting projected matrix increases the objective.

Although the SVD step can be a computational bottleneck for matrices with large numbers of columns, the gradient projection algorithm is favoured over the pairwise optimization algorithm as it generally requires much fewer iterations to converge in practice and is therefore faster. Like the pairwise optimization algorithm, the performance is dependent on the convergence criterion in Line 2. Possible op-

Algorithm 3: Gradient projection

Input : $n \times k$ matrix X, step size α, objective function f_X
Output: rotated $n \times k$ matrix XR

initialize rotation matrix R;
while *not yet converged* **do**
  compute gradient G;
  loop
    $U, D, V \leftarrow \mathrm{SVD}(R + \alpha G)$;
    $R' \leftarrow UV^T$;
    if $f_X(R) \leq f_X(R')$ **then**
      $R \leftarrow R'$;
      break;
    end
    $\alpha \leftarrow \alpha/2$;
  end
end
return XR;

tions for the criterion include the relative increase in the objective and the relative change in the size of the gradient G. However, our experimental results suggest that the step size is also a critical factor for the performance. Too small of a step size results in slow convergence, while too large of a step size results in many iterations of the inner loop. For objectives that have particular convexity properties like in the case of Varimax and Entromin$_2$, there is theory that allows for a modification to Algorithm 3 that resolves the issue of picking a step size. We discuss the advantage of convexity in the following section and conduct a thorough empirical analysis of the factors relevant to convergence in Section 5.1.1 and Section 5.1.2.

4.3 Convexity

Optimizing convex objective functions are ideal as the convexity guarantees that any local solution is also a global solution. As an added bonus for gradient projection algorithms, it can be shown that a convex objective resolves the issue of having to pick a step size, which is necessary for achieving a reasonable convergence speed and can be difficult and problem dependent. Jennrich (2001) showed

that if the objective function is convex and differentiable over the set

$$\mathcal{B}(n,k) = \left\{ X \in \mathbb{R}^{n \times k} : \|X\|_F \leq \sqrt{k} \right\}$$

where $\|X\|_F$ is the Frobenius norm of X, then Algorithm 3 can be modified to not depend on any step size α. The modified algorithm for a convex objective is shown in Algorithm 4.

Algorithm 4: Gradient projection for convex objective

Input : $n \times k$ matrix X
Output: rotated $n \times k$ matrix XR

initialize rotation matrix R;
while *not yet converged* **do**
    compute gradient G;
    $U, D, V \leftarrow \mathrm{SVD}(G)$;
    $R \leftarrow UV^T$;
end
return XR;

Jennrich (2001) also showed that the Quartimax objective (equivalent to the Varimax objective by Proposition 1) is convex over $\mathcal{B}(k,k)$. Note that $\mathcal{B}(k,k)$ is a superset of the set of orthogonal rotation matrices and that Jennrich's proof for the above result does not actually depend on the specification of $\mathcal{B}$. Hence, we show that the objective for Entromin$_2$ is convex over the smaller set

$$\mathcal{B}_{\mathcal{O}}(k) = \left\{ X \in \mathbb{R}^{k \times k} : \|X\|_F \leq \sqrt{k},\ |X_{ij}| \in [0,1],\ i,j \in \{1,\ldots,k\} \right\}$$

and therefore is compatible with Algorithm 4.

Proposition 10. *Let the Entromin$_2$ criterion be defined as in Equation 13. If the matrix U has orthonormal columns, then the second-order objective $f_U^{h_2}$ defined as in Equation 11 is convex over the set $\mathcal{B}_{\mathcal{O}}(k)$.*

The proof of Proposition 10 is given in Section A.7 in the Appendix.

Note that while the Entromin$_2$ objective is convex, the objective for Entromin itself is not convex. Despite resembling the sum of negative entropy functions of the form $x \mapsto x \log x$ which is well-known to be convex (Rao, 1984), the squared values in the Entromin objective results in the loss of convexity. This can be verified empirically where implementing Entromin using Algorithm 4 does not appear to lead to convergence. However, we conjecture that any solution in a locally convex neighbourhood is a global solution and that the multiple solutions are restricted to column permutations and sign-flips. This may explain why the greedy pairwise optimization algorithm tends to work well for Entromin even when its objective is non-convex, where the pairwise optimization algorithm stays in the neighbourhood of one local solution until convergence. Due to time constraints on this thesis, we leave investigating this potential property as future work.

4.4 Computational complexity

We briefly comment on the computational complexity of Algorithm 3 and Algorithm 4. Based on the gradients derived in Section D.2 in the Appendix, assume that computing the gradient G is $\mathscr{O}(n^2k)$ where $\mathscr{O}$ denotes big-O notation. Assume that computing the SVD is $\mathscr{O}(k^3)$ (Vasudevan and Ramakrishna, 2019), as is projecting the updated matrix back onto the orthogonal manifold. Then the complexity of each iteration is approximately $\mathscr{O}(n^2k + k^3)$ for Algorithm 4 and $\mathscr{O}(n^2k + tk^3)$ for Algorithm 3 where t is the number of iterations to find an appropriate step size. Therefore, it is expected that each iteration of Algorithm 3 may take at least as long as each iteration of Algorithm 4 to complete, if not longer.

In practice, we generally have $k \ll n$ and $t < 10$. It is then expected that Algorithm 3 and Algorithm 4 scale similarly with the size of the input matrix. Therefore, when comparing factor rotation methods, it may be more informative to compare the number of iterations that a method takes to converge rather than the measured wall time. We report the number of iterations until convergence for most of the empirical analyses we conduct in Chapter 5.

Chapter 5

Empirical analyses of Entromin

In this chapter, we empirically study the factor rotation methods and orthogonal rotation algorithms discussed in previous chapters on simulated and real datasets. We start by investigating specific properties of interest on simple simulated datasets before evaluating the overall performance on a dataset of New York Times articles. The R code used in the analyses discussed in this chapter can be found on the GitHub repository for this thesis.[2]

Unless otherwise specified, Varimax and Entromin$_2$ is implemented using Algorithm 4 and Entromin is implemented using Algorithm 3 with an initial step size of one. A threshold of 10^{-5} is used for soft-thresholding zeros.

5.1 Simulated datasets

We start our analysis of the gradient projection algorithm and of Varimax, Entromin and Entromin$_2$ on simple and intuitive simulated datasets. Section 5.1.1 examines the effect of the convergence criterion on the performance of the gradient projection algorithm, Section 5.1.2 examines the effect of the initial step size in Algorithm 3 on Entromin, and Section 5.1.3 compares the sparsity of the output for the three factor rotation methods.

[2] https://github.com/chiukenny/MScThesis-entromin

5.1.1 Convergence criterion analysis

We evaluate the effect of various parameters on the convergence of the gradient projection algorithms (Algorithm 3 and Algorithm 4). For simplicity, we consider a simulated Gaussian dataset A with $n = 50{,}000$ rows and $d = 20$ columns where

$$\begin{aligned} A_{ij} &= B_{ij} X_{ij} \ , \\ B_{ij} &\sim \text{Bernoulli}(0.5) \ , \\ X_{ij} &\sim \text{Normal}(0, 1) \end{aligned}$$

for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, d\}$. For the rotation procedure described in Section 2.3, we take $k = d$ principal components and do not apply the normalization and centering steps. The sparsity induced by B_{ij} ensures that there are non-trivial rotations of the principal components that are more optimal than others.

We first examine the effect of the choice of the convergence criterion. Let $\varepsilon > 0$ be a tolerance parameter. We consider the following two criteria in particular:

(C_1) The relative increase in the sum of singular values. Let $D^{(t)}$ denote the singular value matrix D obtained in iteration t of either algorithm. Then the algorithm is considered to have converged if

$$\frac{\sum_{j=1}^{k} D_{jj}^{(t)}}{\sum_{j=1}^{k} D_{jj}^{(t-1)}} < 1 + \varepsilon \ .$$

The sum of singular values of a matrix G is equal to $\|G\|_F^2$ and so this convergence criterion is a measure of the change in the gradient. The algorithm is then considered to have converged if the overall change in the gradient is below a certain threshold in between iterations. Note that this is the convergence criterion used by the Varimax implementation in base R.

(C_2) The relative change in the objective. Let X denote the input matrix to rotate, and let $R^{(t)}$ denote the rotation matrix obtained in iteration t of either

algorithm. Then the algorithm is considered to have converged if

$$\max\left\{\frac{f_X(R^{(t)})}{f_X(R^{(t-1)})}, \frac{f_X(R^{(t-1)})}{f_X(R^{(t)})}\right\} < 1 + \varepsilon \ .$$

The maximum is used to allow for objective functions f_X that can take on negative values.

To allow for a fair comparison, we set $\varepsilon = 10^{-5}$ for all methods. Also, note that C_2 is sensitive to the scale of the gradient relative to the objective value. In particular, the Entromin$_2$ objective has a very large objective value relative to the gradient because of the constant term in the criterion that scales with the number of columns (when U has orthonormal columns). While this is generally not an issue in practice as the size of ε can be adjusted accordingly, it can make comparisons between methods difficult where we set ε to the same value across methods. Without being considerate of the problem, the C_2 criterion will prematurely terminate the algorithm as can be seen in Table 4 in the Appendix. For this reason, we drop off all constants from the objective when evaluating the C_2 criterion.

Table 2 shows the results for the three factor rotation methods over 100 randomly generated datasets. For each rotation method, the mean value of the objectives, the mean ($\pm$ one standard deviation) number of recovered soft zeros, and the mean ($\pm$ one standard deviation) number of iterations until convergence appear to be similar regardless of the convergence criterion used. Therefore, we use the C_1 convergence criterion in all of our following analyses to align with base R's Varimax implementation. Comparing across methods, Entromin achieves a notably higher number of soft zeros in the rotated principal components although also requiring more iterations for convergence.

5.1.2 Step size analysis

We next investigate the effect of the initial step size in Algorithm 3 for Entromin. In each iteration, the rotation matrix is updated by taking a step in the direction of the gradient. However, too large of a step may result in a rotation matrix that decreases the objective, and so the step size is repeatedly halved until the objective

		$v' \times 10^3$	$h_2' \times 10^3$	h	Soft zeros	Iterations
Var.						
	C_1	2.400	-4.800	187.987	$41{,}238 \pm 2103$	18 ± 3
	C_2	2.400	-4.800	187.987	$41{,}207 \pm 2097$	17 ± 3
Entr.$_2$						
	C_1	2.400	-4.800	187.987	$41{,}223 \pm 2107$	22 ± 3
	C_2	2.400	-4.800	187.987	$41{,}189 \pm 2114$	21 ± 3
Entr.						
	C_1	2.398	-4.796	187.966	$98{,}173 \pm 4926$	73 ± 13
	C_2	2.398	-4.796	187.966	$97{,}562 \pm 4815$	72 ± 13

Table 2: Results from the convergence criteria analysis over 100 simulated sparse Gaussian datasets $A \in \mathbb{R}^{50{,}000 \times 20}$. C_1 and C_2 are the sum of singular values and the relative objective increase criteria, respectively. The mean values of the Varimax, Entromin$_2$ and Entromin criteria with constant terms dropped are denoted by v', h_2' and h, respectively. (Standard deviations are not shown due to being orders of magnitude smaller.) The rounded mean $\pm$ one standard deviation number of soft zeros (threshold 10^{-5}) in the 20 rotated principal components and number of iterations until convergence are also shown.

finally increases. We run Entromin on one simulated dataset from Section 5.1.1 but with two different initial step sizes. Figure 4 shows how the value of the Entromin criterion and the step size changes across iterations for both runs. The first run (dashed line) starts with an initial step size of $\alpha_0 = 1$ but immediately decreases to $\alpha_1 = 2^{-4} = 0.0625$ on the first iteration, after which it remains constant until the algorithm converges after 92 iterations. The second run (solid line) starts with an initial step size of $\alpha_0 = 0.1$ and remains constant until the 12^{th} iteration where the step size halves to $\alpha_{12} = 0.05$. The second run converges after 53 iterations. The final value of the objective is roughly the same for both runs.

Even though the first run started with a larger step size, the large step size could not be maintained and resulted in many more iterations compared to the second run. The key conclusions here are that the initial step size may have a significant impact

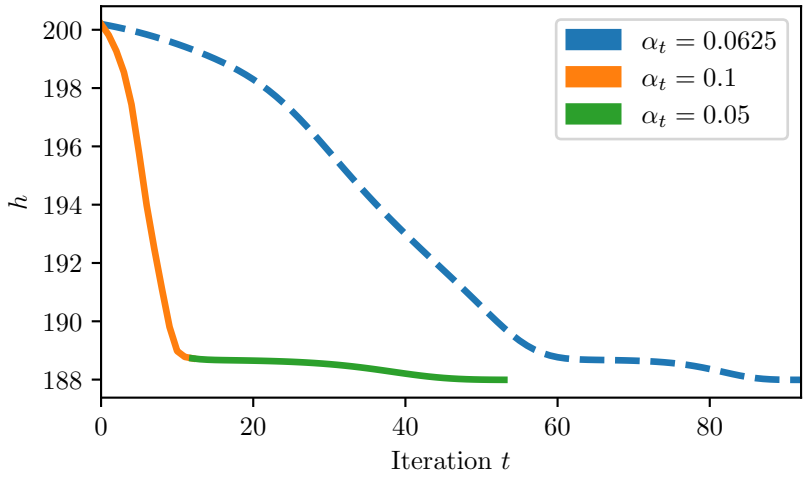

Figure 4: The value of the Entromin criterion h and the step size (colour) in the gradient projection algorithm as they change across iterations for two Entromin runs on a simulated sparse Gaussian dataset $A \in \mathbb{R}^{50,000 \times 20}$. Run 1 (dashed line) starts with an initial step size of $\alpha_0 = 1$ and Run 2 (solid line) starts with an initial step size of $\alpha_0 = 0.1$. (Best viewed in colour.)

on the runtime of the algorithm, and that choosing the optimal initial step size may not be easy without experimenting. The results observed in this particular example may suggest that obtaining a new step size by halving the current step size is overly conservative and that decreasing the current step size by some fixed amount would fair better. However, choosing the amount to decrease by is then another problem on its own. A line search for the optimal step size may be possible and is one direction of research that we leave for future work. If the performance of the algorithm is an important point of consideration, it is recommended to first try various values for the initial step size and take the best performing one.

5.1.3 Sparsity analysis

We evaluate the sparsity quality of the factor rotation methods on a small, simulated Gaussian dataset A with $n = 1000$ rows and $d = 3$ columns that is generated through the following procedure:

1. generate $n \times d$ standard normal random variables X_{ij} for $i \in \{1, \ldots, n\}$ and $j \in \{1, 2, 3\}$;

2. generate $n \times (d-1)$ Bernoulli(p) random variables B_{ij} where $p \in [0, 1]$ for $i \in \{1, \ldots, n\}$ and $j \in \{1, 2\}$; finally,

3. take $A_{ij} = B_{ij}X_{ij}$ for $j \in \{1, 2\}$ and $A_{ij} = X_{ij}$ for $j = 3$.

The dataset A generated through this particular procedure resembles a spherical Gaussian cluster with perpendicular radial streaks on a two-dimensional subspace. When the dataset is fed into the rotation procedure described in Section 2.3 (without the normalization and centering steps), the leading $k = 2$ principal components resemble a two-dimensional Gaussian cluster with a streak that generally aligns with the x-axis when plotted (SVD is able to find relatively sparse components in this example due to the simple construction). We apply a random rotation to the principal components and evaluate the methods depending on how well they can visually recover the sparse structure. Depending on the sparsity parameter p, the signal from the points in the streak may be weak compared to the noise from the Gaussian cluster and may make recovery of the sparse structure difficult.

Figure 5 shows the two principal components as well as the rotated components for Varimax, Entromin$_2$ and Entromin for $p \in \{0.3, 0.4, 0.5\}$. As discussed in Section 5.1.1, all three rotation methods use the sum of singular value convergence criterion (C_1) with $\varepsilon = 10^{-5}$ for a fair comparison. Convergence details for the three methods are provided in Section E.2 in the Appendix. For all p, it can be seen that the rotation found by Varimax and Entromin$_2$ are very similar. It can also be seen that Entromin tends to align the streak more closely with the axis compared to the other two methods. For $p = 0.3$, none of the three methods are able to align the streak with the axis. For $p = 0.4$, Entromin can be considered a success or very close while the other two methods still have room for improvement. For $p = 0.5$, all three methods are able to align the streak with the axis.

We also quantitatively evaluate the sparsity by counting the number of soft zeros in the rotated principal components. Figure 6 shows the mean and standard de-

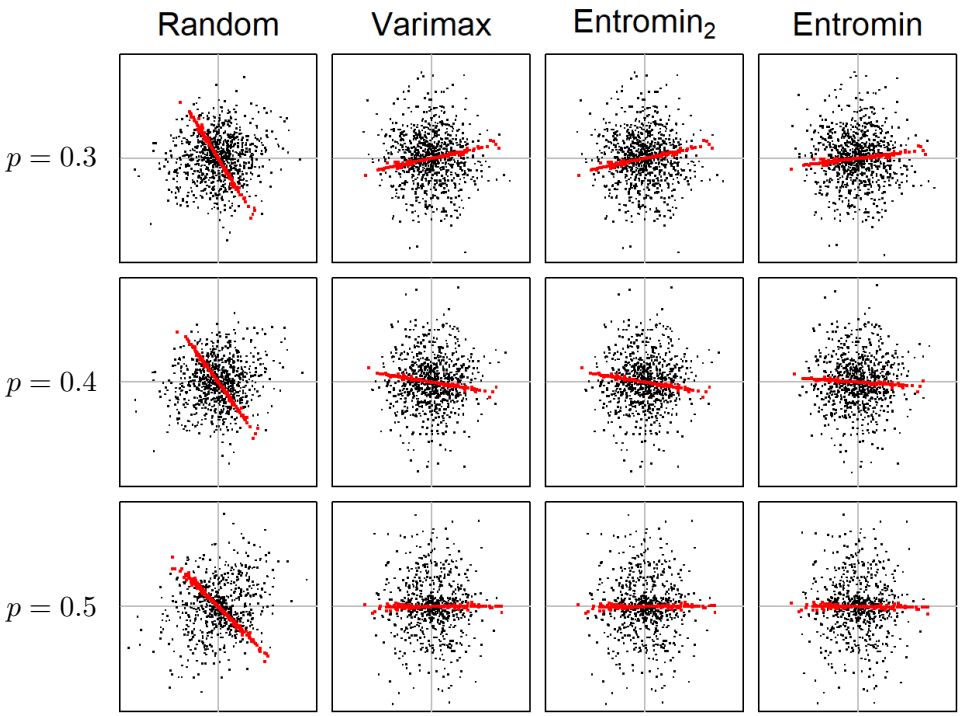

Figure 5: Plots of two randomly rotated and method-rotated principal components for the simulated sparse Gaussian dataset $A \in \mathbb{R}^{1000\times 3}$ with varying levels of sparsity. Sparsity is controlled by the parameter $p = \mathbb{P}(A_{ij} = 0)$ for $i \in \{1, \ldots, 1000\}$ and $j \in \{1, 2\}$. The observed radial streaks are roughly highlighted in red. (Best viewed in colour.)

viation of the number of soft zeros in the rotated principal components over 100 datasets for each of $p \in \{0, 0.1, \ldots, 0.9\}$. For datasets constructed through the procedure above, it is clear that Entromin generally recovers greater sparsity in the rotated principal components compared to Varimax and Entromin$_2$. The difference between Varimax and Entromin$_2$ is not as obvious.

5.2 New York Times articles dataset

We examine the performance of Entromin and Entromin$_2$ on the same New York Times articles dataset (Dua and Graff, 2017) that Rohe and Zeng (2020) examined Varimax under. The dataset consists of the bag-of-words counts for 102,660 words

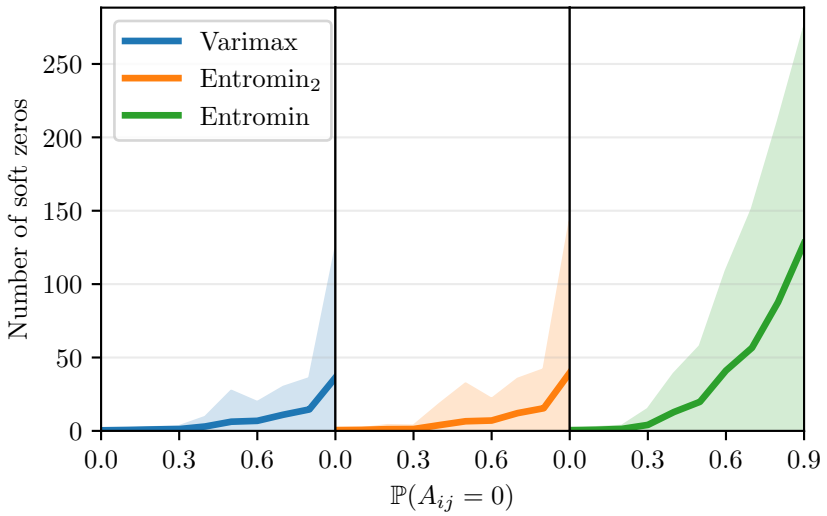

Figure 6: Mean $\pm$ one standard deviation number of soft zeros (threshold 10^{-5}) in the rotated principal components over 100 simulated sparse Gaussian datasets $A \in \mathbb{R}^{1000\times 3}$ for each $\mathbb{P}(A_{ij} = 0) \in \{0, 0.1, \ldots, 0.9\}$ where $i \in \{1, \ldots, 1000\}$, $j \in \{1, 2\}$. (Best viewed in colour.)

across 300,000 news articles. The data are represented in the form of a 300,000 $\times$ 102,660 matrix A where A_{ij} is the number of times word j occurs in article i. The data matrix is very sparse with approximately 99.8% of its elements being 0.

We apply the same data preprocessing procedure that Rohe and Zeng (2020) applied for Varimax. Because the row and column sums of A are highly heterogeneous, both the rows and the columns of the matrix are normalized. The optional centering and recentering steps are also performed. An analysis of the leading 50 principal components shows that twelve of the components are dominated by a relatively few number of outlier articles (based on a somewhat arbitrary threshold of $\|\widehat{U}_{\cdot j}\|_4 > 0.15$) and so they are discarded (Zhang and Rohe, 2018). A scree plot of the remaining 38 components shows an eigengap at $k = 8$. Therefore, only the first eight remaining principal components are retained and rotated. The mentioned plots used to analyze the principal components can be found in Section E.3 in the Appendix.

Figure 7 shows the rotated principal components for Varimax (lower left triangular) and Entromin$_2$ (upper right triangular). Figure 8 shows the same but with Entromin on the upper right triangular. The sign of each component is set so that the third sample moment (the skew) is positive. Note that the Varimax-rotated principal components are arranged in order of decreasing variance, whereas the components for Entromin and Entromin$_2$ are manually ordered to best match Varimax. The plots on the diagonals show the Varimax-rotated component plotted against the matched component of the other method. Although each component consists of $300{,}000$ elements, the plots only show a sample of $5{,}000$ for clarity. The same $5{,}000$ elements are shown in every plot.

The diagonal plots in Figure 7 show that the Varimax-rotated principal components and the Entromin$_2$-rotated components are in near one-to-one correspondence. This suggests that the rotation found by Entromin$_2$ is nearly identical to the rotation found by Varimax (up to column permutations and sign-flips). Compared to the unrotated principal components in Figure 1, the rotated components in Figure 7 align with the axes more closely and thus are expected to be more interpretable.

Compared to Figure 7, the diagonal plots in Figure 8 show a more notable difference between the Varimax-rotated principal components and the Entromin-rotated components. The rotated components are also clearly more aligned with the axes compared to the unrotated principal components in Figure 1, but it is difficult to visually discern the quality of the Entromin-estimated factors compared to that of the Varimax-estimated factors.

Table 3 shows the values of the Varimax, Entromin$_2$ and Entromin criteria for the rotated principal components. The number of recovered soft zeros, the number of iterations until convergence and the total runtime are also shown. It can be seen that the criteria values for all three rotation methods are relatively similar for the New York Times articles dataset. As observed in Figure 7 and Figure 8, the results for Varimax and Entromin$_2$ are nearly identical while the results for Entromin differ slightly from the other two methods. In terms of sparsity in the rotated prin-

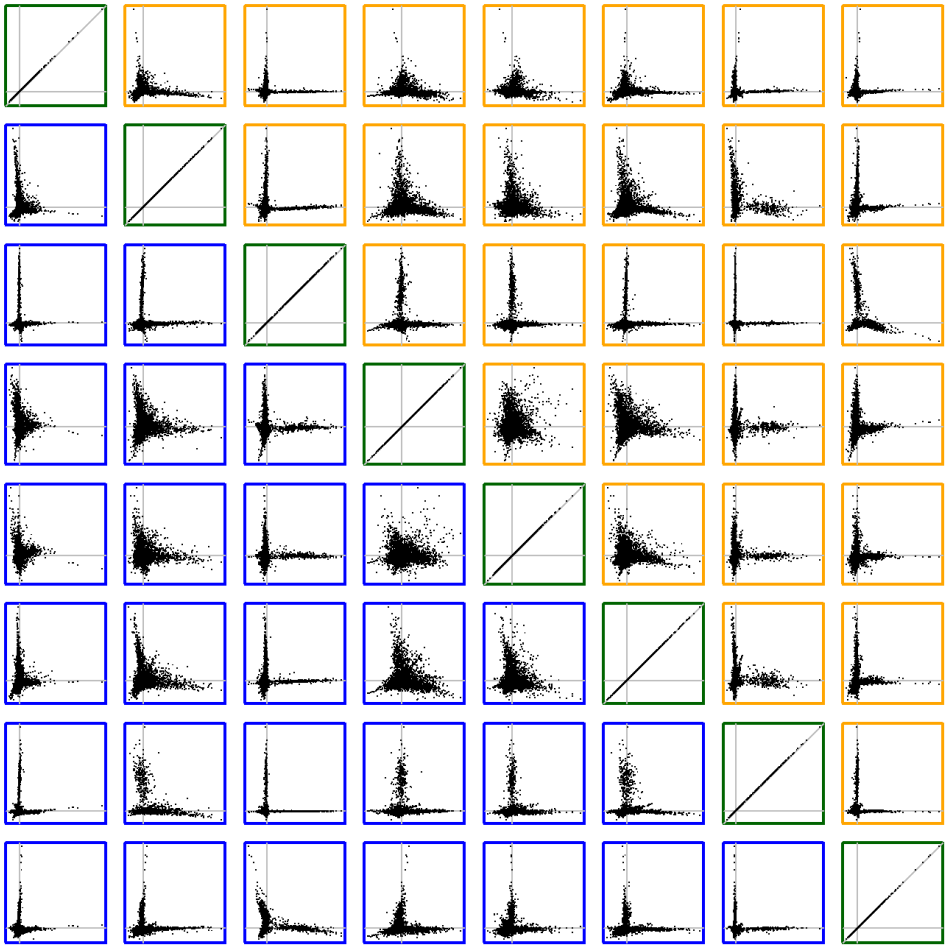

Figure 7: Pairwise plots of the principal components rotated by Varimax (blue) and Entromin$_2$ (orange) for the New York Times articles dataset. Each plot on the diagonal (green) shows one of the Varimax-rotated components plotted against its manually matched Entromin$_2$-rotated component. Of the 300,000 elements in each component, only 5000 are randomly sampled and shown. (Best viewed in colour.)

cipal components, Entromin recovers the greatest number of soft zeros, followed by Entromin$_2$ and Varimax. In terms of the number of iterations and runtime, the two methods that use Algorithm 4 are significantly faster than Entromin as both methods converged within a quarter of the number of iterations and of the time that

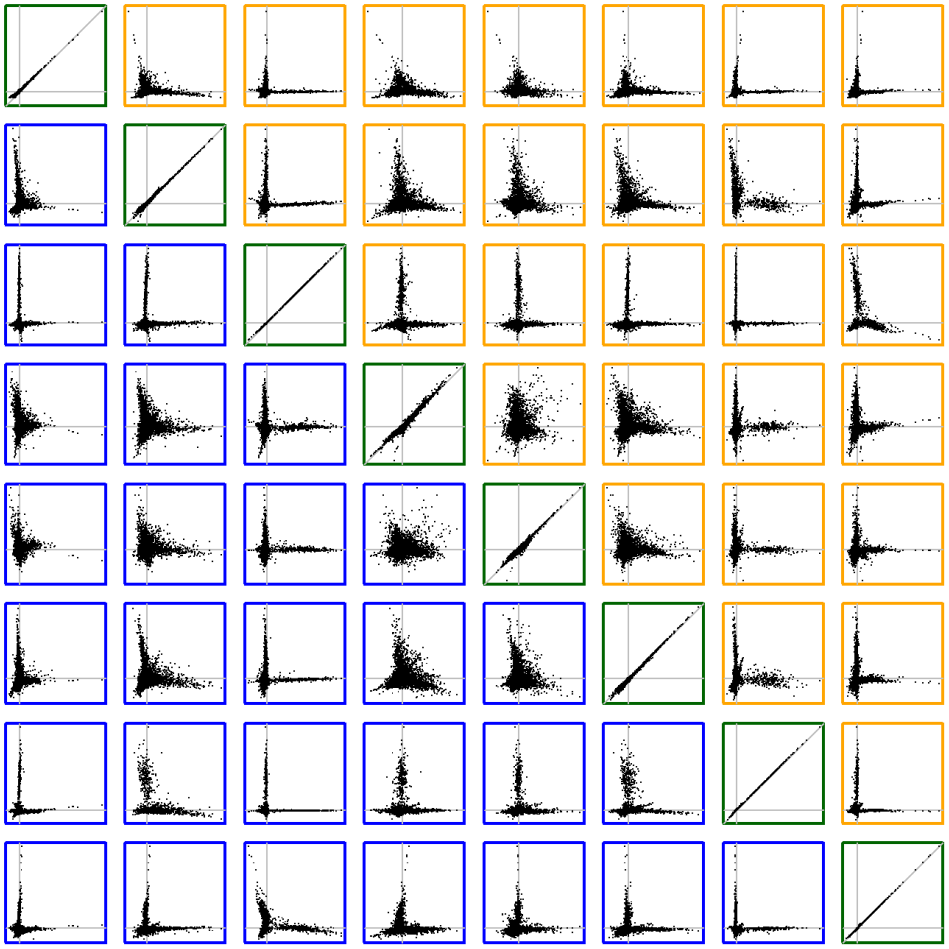

Figure 8: Pairwise plots of the principal components rotated by Varimax (blue) and Entromin (orange) for the New York Times articles dataset. Each plot on the diagonal (green) shows one of the Varimax-rotated components plotted against its manually matched Entromin-rotated component. Of the 300,000 elements in each component, only 5000 are randomly sampled and shown. (Best viewed in colour.)

it took for Entromin (Algorithm 3) to converge.

Rotation	$v \times 10^9$	h_2	h	Soft zeros	Iterations	Runtime
Varimax	3.087	11.998	82.879	19,674	8	3.48
Entromin$_2$	3.087	11.998	82.878	19,725	9	4.58
Entromin	3.025	11.998	82.754	19,810	44	24.13

Table 3: Results from the analysis of the New York Times articles dataset. The Varimax, Entromin$_2$ and Entromin criteria are denoted by v, h_2 and h, respectively. The number of soft zeros (threshold 10^{-5}) in the eight rotated principal components is shown. The number of iterations and the total runtime (in seconds) until convergence are also shown.

Chapter 6

Conclusion

Various Entromin criteria have been introduced and associated with the historic claim that they find relatively sparser factors compared to Varimax. The claim appears to hold some truth as evident by the persistent usage of Entromin today and as observed in our empirical analyses. However, despite now being a few decades old, the success of Entromin as an orthogonal rotation method remains an enigma.

As a contribution to the renewed interest in factor rotations in recent literature, we have dedicated this thesis to taking steps towards developing a rigorous statistical understanding of Entromin and its properties. The main contributions of our work include several theoretical results. We have shown that Varimax is a first-order approximation of Entromin, and that generalizing this connection leads to an entire family of Entromin approximations. We have studied the second-order member of the approximating family, Entromin_2, and identified the exact conditions under which it is expected to recover the true latent factors. We have also made the connection to cumulants of the factor distributions, which allow for a more natural expression of the identifiability conditions discussed in (Rohe and Zeng, 2020) and in this thesis. Our final main theoretical result formalized the connection between minimizing a particular Entromin criterion (the Inagaki entropy criterion) and recovering sparsity in the factors, and its implications lead to a stronger intuition for the Entromin sparsity claim that remains to be explained.

Furthermore, we have provided an overview of the two algorithms commonly used to implement orthogonal rotation methods. The older pairwise optimization algorithm is simple to understand and implement but has somewhat fallen out of favour due to its need for a line search. The relatively newer gradient projection algorithm generally requires fewer iterations to converge and is much faster—particularly when the objective is convex and free of a step size parameter as in the case of Varimax and Entromin$_2$. The results of our empirical analyses of the factor rotation methods suggest that Varimax and Entromin$_2$ are generally similar in both performance and quality of the estimated factors, while Entromin generally recovers sparser factors at the expense of being slower. All three factor rotation methods are able to produce reasonable factors for the New York Times articles dataset.

We emphasize that while more work needs to be done in order to develop a complete understanding of Entromin and its statistical properties, our results have made progress towards this goal and provide a foundation for a deeper study of Entromin. We conclude the main body of this thesis by highlighting some relevant directions of theoretical research that are currently ongoing or that we believe may lead to something insightful.

Family of Entromin approximations

The main purpose of studying the approximating family is to develop the properties of the finite-order members and eventually generalize to Entromin by taking the order to the limit. In this thesis, we have only examined the second-order member. We have discussed what we expect from the general members of the approximating family of Entromin in Section 3.6, but rigorous theory is currently lacking. The next step would be to derive a general representation for the expectation of the members of the family. While a general formula in terms of moments of the factors can be derived (Section B.2 in the Appendix), such a formula is not easy to work with for developing statistical theory as each moment may appear in multiple terms and collecting the coefficients of each moment is not straightforward. We believe that a general formula in terms of cumulants would be easier to analyze if such a formulation is possible.

Unsatisfiable moment conditions

Theorem 4 states that Entromin$_2$ may identify the latent factors even when their distributions are not leptokurtic as long as Equation 14 is satisfied. However, as noted in Section 3.2, we have so far been unable to find nor construct a distribution that satisfies the condition under Assumption 2. It remains to be seen whether there exists a distribution that satisfies the condition or if the condition is theoretically impossible to satisfy.

Tighter sparsity bound

The upper bound for the probability of non-zero given in Theorem 9 is very loose due to time constraints on this thesis. We believe that a tighter general bound (i.e., a larger upper bound) is possible. Our experiments suggest that a universal bound independent of the number of factors k may even be possible (an upper bound of $\frac{1}{3}$ appears to be sufficient in our tests).

Finite-sample results

In this thesis, the statistical properties of Entromin are examined under the semi-parametric factor model which involves the population data matrix $\mathbb{E}[A|Z, Y]$, i.e., the data matrix A as the number of samples goes to infinity. While it is important to understand the limiting behaviour of Entromin and other factor rotation methods, datasets in practice will always be finite and so finite-sample results are also desirable.

Relaxing identifiability assumptions

The theoretical results presented in this thesis all assume that the latent factors are independent. This assumption is statistically convenient for simplifying the analyses and proofs. However, it is unlikely that the assumption holds on real datasets and so the results may not generalize to the practical setting. The proofs of the results rely heavily on the independence assumption and so a different approach for understanding the statistical properties of factor rotation methods would likely be necessary.

Soft sparsity

This thesis is mainly concerned with the case of hard sparsity (exact zeros). However, soft sparsity (small values close to zero) is more common in practice as rotations that result in multiple new zeros may not be possible for a given dataset. Our results may not directly translate to the case of soft sparsity and would likely require different proof techniques.

Global solutions of Entromin

As the objective for Entromin is non-convex, it would be beneficial to have theoretical guarantees for the solution returned by Entromin. We conjecture in Section 4.3 that Entromin solutions in locally convex neighbourhoods are global Entromin solutions. Our hypothesis is based on the fact that it holds for the function $x \mapsto x^2 \log x^2$, which is symmetric and has a unique (but identical) solution on both orthants of $\mathbb{R}$. It may be the case that this property does not hold for the Entromin objective as a function of the rotation R. Further analysis will be needed to determine whether the conjecture or a related result hold.

Bibliography

Bernaards, C. A. and Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, 65(5):676–696. → page 8

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022. → page 16

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36(1):111–150. → page 8

Chen, F. and Rohe, K. (2020). A new basis for sparse PCA. arXiv:2007.00596. → page 16

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314. Higher Order Statistics. → page 16

Dowson, D. and Wragg, A. (1973). Maximum-entropy distributions having prescribed first and second moments (corresp.). *IEEE Transactions on Information Theory*, 19(5):689–693. → page 83

Dua, D. and Graff, C. (2017). UCI machine learning repository. http://archive.ics.uci.edu/ml. → page 46

Dytso, A., Bustin, R., Poor, H. V., and Shamai, S. (2018). Analytical properties of generalized Gaussian distributions. *Journal of Statistical Distributions and Applications*, 5(1):6. → page 23

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons Inc., New York. → page 13

Golub, G. H. and Ortega, J. M. (2014). *Scientific Computing: An Introduction with Parallel Computing*. Academic Press. → page 34

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137. → page 16

Hu, Z., Pan, G., Wang, Y., and Wu, Z. (2016). Sparse principal component analysis via rotation and truncation. *IEEE Transactions on Neural Networks and Learning Systems*, 27(4):875–890. → page 16

Inagaki, A. (1993). Entromin : A new factor solution and a new orthogonal factor rotation method based upon entropy. *Abstracts of Japan Society of Physical Education, Health and Sport Sciences Conference*, page 580. → pages 2, 13, 15

Inagaki, A. (1994). Entromin and entromax : Criteria for orthogonal factor rotation and direct factor solution based on entropy. *Abstracts of Japan Society of Physical Education, Health and Sport Sciences Conference*, page 455. → page 13

Jennrich, R. (2001). A simple general procedure for orthogonal rotation. *Psychometrika*, 66(2):289–306. → pages 35, 37, 38, 87

Jennrich, R. (2004). Rotation to simple loadings using component loss functions: The orthogonal case. *Psychometrika*, 69:257–273. → page 13

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200. → pages 2, 10, 19, 33, 34

Manton, J. H. (2002). Optimization algorithms exploiting unitary constraints. *IEEE transactions on signal processing*, 50(3):635–650. → page 36

Maxwell, J. C. (1860). V. illustrations of the dynamical theory of gases.—part i. on the motions and collisions of perfectly elastic spheres. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 19(124):19–32. → page 13

McCammon, R. B. (1970). *Minimum entropy criterion for analytic rotation*. Number 43 in Computer Contributions. Kansas Geological Survey. → pages 2, 13, 14

Nemec, W. and Peron, J. (1979). Entromin: a minimum entropy criterion and Fortran program for analytic rotation in factor analysis. *Acta Universitatis Wratislaviensis, Prace Geol.-Mineral.* → pages 33, 35

Nemenman, I., Shafee, F., and Bialek, W. (2001). Entropy and inference, revisited. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 471–478, Cambridge, MA, USA. MIT Press. → pages 74, 83

Neuhaus, J. O. and Wrigley, C. (1954). The quartimax method. *British Journal of Statistical Psychology*, 7:81–91. → page 19

Pastor, G., Mora-Jimenez, I., Jantti, R., and Caamano, A. (2020). Constructing measures of sparsity. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1. → page 14

Pearson, K. (1916). Mathematical contributions to the theory of evolution. xix. second supplement to a memoir on skew variation. *Philosophical transactions of the Royal Society of London. Series A, Containing papers of a mathematical or physical character*, 216(538-548):429–457. → page 27

Ramírez, I. and Sapiro, G. (2011). Sparse coding and dictionary learning based on the MDL principle. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2160–2163. → page 16

Rao, C. R. (1984). Convexity properties of entropy functions and analysis of diversity. *Lecture Notes-Monograph Series*, 5:68–77. → page 39

Richard, A., Fuchs, L., Largeteau-Skapin, G., and Andres, E. (2011). Decomposition of nD-rotations: Classification, properties and algorithm. *Graph. Models*, 73(6):346–353. → page 34

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471. → page 16

Rohe, K. and Zeng, M. (2020). Vintage factor analysis with varimax performs statistical inference. arXiv:2004.05387. → pages v, 2, 6, 7, 10, 11, 12, 16, 17, 26, 27, 30, 46, 47, 52, 91

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423. → page 13

Smith, J. O. (accessed July 25, 2021). *Spectral Audio Signal Processing*. http://ccrma.stanford.edu/~jos/sasp/. Online book, 2011 edition. → pages 13, 22

Squires, S., Bennett, A. P., and Niranjan, M. (2019). Minimum description length as an objective function for non-negative matrix factorization. arXiv:1902.01632. → page 16

Thurstone, L. L. . (1935). *The vectors of mind; multiple-factor analysis for the isolation of primary traits*. The University of Chicago Press, Illinois. → pages 2, 5, 6

Thurstone, L. L. (1947). *Multiple factor analysis.* Multiple factor analysis. Chicago, University of Chicago Press. → page 6

Thurstone, L. L. (1954). An analytical method for simple structure. *Psychometrika*, 19(3):173–182. → page 6

Vasudevan, V. and Ramakrishna, M. (2019). A hierarchical singular value decomposition algorithm for low rank matrices. arXiv:1710.02812. → page 39

Wall, M. E., Rechtsteiner, A., and Rocha, L. M. (2003). *Singular Value Decomposition and Principal Component Analysis*, pages 91–109. Springer US, Boston, MA. → page 7

Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. R.I.P. *The American Statistician*, 68(3):191–195. → pages 12, 22

Willink, R. (2008). A unique property of the normal distribution associated with perturbing a general random variable. *The American Statistician*, 62(2):144–146. → page 25

Zhang, Y. and Rohe, K. (2018). Understanding regularized spectral clustering via graph conductance. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 10654–10663, Red Hook, NY, USA. Curran Associates Inc. → page 47

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286. → page 16

Appendix A

Proofs

This chapter of the Appendix contains the mathematical proofs for the theoretical results presented in this thesis.

A.1 Proof of Proposition 1

Proof. Because U has orthonormal columns and $R \in \mathcal{O}(k)$, the matrix UR has orthonormal columns. The Varimax objective in Equation 4 can then be written as

$$\begin{aligned}
\underset{R\in\mathcal{O}(k)}{\arg\max}\ f_U^v(R) &= \underset{R\in\mathcal{O}(k)}{\arg\max}\ \sum_{j=1}^{k}\left(\frac{1}{n}\sum_{i=1}^{n}[UR]_{ij}^4 - \left(\frac{1}{n}\sum_{i=1}^{n}[UR]_{ij}^2\right)^2\right) \\
&= \underset{R\in\mathcal{O}(k)}{\arg\max}\ \sum_{j=1}^{k}\left(\frac{1}{n}\sum_{i=1}^{n}[UR]_{ij}^4 - \frac{1}{n^2}\right) \\
&= \underset{R\in\mathcal{O}(k)}{\arg\max}\ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^4 - \frac{k}{n^2} \\
&= \underset{R\in\mathcal{O}(k)}{\arg\max}\ \sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^4 \\
&= \underset{R\in\mathcal{O}(k)}{\arg\max}\ f_U^q(R)
\end{aligned}$$

which is exactly the Quartimax objective. Thus, maximizing the Varimax criterion is equivalent to maximizing the Quartimax criterion. □

A.2 Proof of Proposition 2

Proof. Because U has orthonormal columns and $R \in \mathcal{O}(k)$, the matrix UR has orthonormal columns. The first-order approximation of the Entromin criterion then has the form

$$\begin{aligned} h_1(U,R) &= -\sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^2\left([UR]_{ij}^2-1\right) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^4+\sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^2 \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^4+k\ . \end{aligned}$$

Therefore, its corresponding objective is

$$\begin{aligned} \underset{R\in\mathcal{O}(k)}{\arg\max}\ f_U^{h_1}(R) &= \underset{R\in\mathcal{O}(k)}{\arg\max}\ -h_1(U,R) \\ &= \underset{R\in\mathcal{O}(k)}{\arg\max}\ \sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^4-k \\ &= \underset{R\in\mathcal{O}(k)}{\arg\max}\ \sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^4 \\ &= \underset{R\in\mathcal{O}(k)}{\arg\max}\ f_U^{q}(R) \end{aligned}$$

which is exactly the Quartimax objective. Thus, minimizing the criterion of the first-order member of the Entromin approximating family is equivalent to maximizing the Quartimax criterion. □

A.3 Proof of Theorem 4

We use the following notation for summations of cross terms in the proof:

$$\sum_{\ell \neq m} Z_\ell Z_m = \frac{1}{2} \sum_{\ell=1}^{k} \sum_{\substack{m=1 \\ m \neq \ell}}^{k} Z_\ell Z_m \ ,$$

$$\sum_{\ell \neq m} Z_\ell^2 Z_m = \sum_{\ell=1}^{k} \sum_{\substack{m=1 \\ m \neq \ell}}^{k} Z_\ell^2 Z_m \ ,$$

$$\sum_{\ell \neq m \neq q} Z_\ell Z_m Z_q = \frac{1}{6} \sum_{\ell=1}^{k} \sum_{\substack{m=1 \\ m \neq \ell}}^{k} \sum_{\substack{q=1 \\ q \neq \ell \\ q \neq m}}^{k} Z_\ell Z_m Z_q \ .$$

Also, define the set of $k \times k$ doubly-stochastic matrices

$$\mathcal{Q}(k) = \left\{ O^{(2)} : O \in \mathcal{O}(k) \right\} \ .$$

Proof. Let $f^{h_2}_{Z^o \widetilde{R}^T}$ be the corresponding Entromin$_2$ objective to maximize. The expectation of the objective with respect to the distribution of the row Z_1 is given by

$$\begin{aligned} \mathbb{E}\left[f^{h_2}_{Z^o \widetilde{R}^T}(R) \right] &= -\mathbb{E}\left[h_2(Z^o \widetilde{R}^T, R) \right] \\ &= -\frac{1}{2} \sum_{j=1}^{k} \mathbb{E}\left[[Z^o \widetilde{R}^T R]_j^6 \right] + 2 \sum_{j=1}^{k} \mathbb{E}\left[[Z^o \widetilde{R}^T R]_j^4 \right] \\ &\quad - \frac{3}{2} \sum_{j=1}^{k} \mathbb{E}\left[[Z^o \widetilde{R}^T R]_j^2 \right] \ . \end{aligned}$$

Define $O = \widetilde{R}^T R$. Because $\widetilde{R} \in \mathcal{O}(k)$ and $R \in \mathcal{O}(k)$, $O \in \mathcal{O}(k)$. Then

$$\begin{aligned}
\mathbb{E}\left[f^{h_2}_{Z^o\widetilde{R}^T}(R)\right] &= -\frac{1}{2}\sum_{j=1}^{k}\mathbb{E}\left[[Z^oO]_j^6\right] + 2\sum_{j=1}^{k}\mathbb{E}\left[[Z^oO]_j^4\right] - \frac{3}{2}\sum_{j=1}^{k}\mathbb{E}\left[[Z^oO]_j^2\right] \\
&= -\frac{1}{2}\sum_{j=1}^{k}\mathbb{E}\left[[Z^oO]_j^6\right] + 2\sum_{j=1}^{k}\mathbb{E}\left[[Z^oO]_j^4\right] \\
&\quad - \frac{3}{2}\sum_{j=1}^{k}\mathbb{E}\left[\sum_{\ell=1}^{k}(Z^o_\ell)^2O^2_{\ell j} + 2\sum_{\ell\neq m}Z^o_\ell Z^o_m O_{\ell j}O_{mj}\right] \\
&= -\frac{1}{2}\sum_{j=1}^{k}\mathbb{E}\left[[Z^oO]_j^6\right] + 2\sum_{j=1}^{k}\mathbb{E}\left[[Z^oO]_j^4\right] \\
&\quad - \frac{3}{2}\sum_{\ell=1}^{k}\mathbb{E}\left[(Z^o_\ell)^2\right]\sum_{j=1}^{k}O^2_{\ell j} \\
&= -\frac{1}{2}\sum_{j=1}^{k}\mathbb{E}\left[[Z^oO]_j^6\right] + 2\sum_{j=1}^{k}\mathbb{E}\left[[Z^oO]_j^4\right] - \frac{3}{2}k\,.
\end{aligned}$$

The sixth-order term is

$$\begin{aligned}
\mathbb{E}\left[[Z^oO]_j^6\right] &= \mathbb{E}\left[\left(\sum_{\ell=1}^{k}Z^o_\ell O_{\ell j}\right)^6\right] \\
&= \sum_{\ell=1}^{k}\mathbb{E}\left[(Z^o_\ell)^6\right]O^6_{\ell j} + 15\sum_{\ell\neq m}\mathbb{E}\left[(Z^o_\ell)^4\right]\mathbb{E}\left[(Z^o_m)^2\right]O^4_{\ell j}O^2_{mj} \\
&\quad + 90\sum_{\ell\neq m\neq q}\mathbb{E}\left[(Z^o_\ell)^2\right]\mathbb{E}\left[(Z^o_m)^2\right]\mathbb{E}\left[(Z^o_q)^2\right]O^2_{\ell j}O^2_{mj}O^2_{qj} \\
&\quad + 20\sum_{\ell\neq m}\mathbb{E}\left[(Z^o_\ell)^3\right]\mathbb{E}\left[(Z^o_m)^3\right]O^3_{\ell j}O^3_{mj} \\
&= \sum_{\ell=1}^{k}\mathbb{E}\left[(Z^o_\ell)^6\right]O^6_{\ell j} + 15\sum_{\ell\neq m}\mathbb{E}\left[(Z^o_\ell)^4\right]O^4_{\ell j}O^2_{mj} \\
&\quad + 90\sum_{\ell\neq m\neq q}O^2_{\ell j}O^2_{mj}O^2_{qj} \qquad (1)
\end{aligned}$$

where (1) follows from Conditions (iii) and (iv) of Assumption 2.

The fourth-order term is

$$\begin{aligned}
\mathbb{E}\left[[Z^o O]_j^4\right] &= \mathbb{E}\left[\left(\sum_{\ell=1}^{k} Z_\ell^o O_{\ell j}\right)^4\right] \\
&= \sum_{\ell=1}^{k} \mathbb{E}\left[(Z_\ell^o)^4\right] O_{\ell j}^4 + 6 \sum_{\ell \neq m} \mathbb{E}\left[(Z_\ell^o)^2\right] \mathbb{E}\left[(Z_m^o)^2\right] O_{\ell j}^2 O_{mj}^2 \\
&= \sum_{\ell=1}^{k} \mathbb{E}\left[(Z_\ell^o)^4\right] O_{\ell j}^4 + 6 \sum_{\ell \neq m} O_{\ell j}^2 O_{mj}^2 \qquad (2)
\end{aligned}$$

where (2) follows from Condition (iii) of Assumption 2. Then

$$\begin{aligned}
\mathbb{E}\left[f^{h_2}_{Z^o \widetilde{R}^T}(R)\right] = 2 \sum_{j=1}^{k} \Bigg(& \sum_{\ell=1}^{k} \mathbb{E}\left[(Z_\ell^o)^4\right] O_{\ell j}^4 + 6 \sum_{\ell \neq m} O_{\ell j}^2 O_{mj}^2 \Bigg) \\
& - \frac{1}{2} \sum_{j=1}^{k} \Bigg(\sum_{\ell=1}^{k} \mathbb{E}\left[(Z_\ell^o)^6\right] O_{\ell j}^6 + 15 \sum_{\ell \neq m} \mathbb{E}\left[(Z_\ell^o)^4\right] O_{\ell j}^4 O_{mj}^2 \\
& + 90 \sum_{\ell \neq m \neq q} O_{\ell j}^2 O_{mj}^2 O_{qj}^2 \Bigg) - \frac{3}{2} k \, .
\end{aligned}$$

Notice that the expectation is only dependent on elements of O to the even powers. Rewrite the expectation in terms of a doubly stochastic matrix $Q = O^{(2)}$. The

expectation as a function of Q is

$$
\begin{aligned}
F(Q) := & \, 2\sum_{j=1}^{k}\left(\sum_{\ell=1}^{k}\mathbb{E}\left[(Z_\ell^o)^4\right]Q_{\ell j}^2 + 6\sum_{\ell\neq m}Q_{\ell j}Q_{mj}\right) \\
& - \frac{1}{2}\sum_{j=1}^{k}\left(\sum_{\ell=1}^{k}\mathbb{E}\left[(Z_\ell^o)^6\right]Q_{\ell j}^3 + 15\sum_{\ell\neq m}\mathbb{E}\left[(Z_\ell^o)^4\right]Q_{\ell j}^2Q_{mj}\right. \\
& \left. +90\sum_{\ell\neq m\neq q}Q_{\ell j}Q_{mj}Q_{qj}\right) - \frac{3}{2}k \\
= & \, 2\sum_{j=1}^{k}\left(\sum_{\ell=1}^{k}\mathbb{E}\left[(Z_\ell^o)^4\right]Q_{\ell j}^2 + 3\left(\left(\sum_{\ell=1}^{k}Q_{\ell j}\right)^2 - \sum_{\ell=1}^{k}Q_{\ell j}^2\right)\right) \\
& - \frac{1}{2}\sum_{j=1}^{k}\left(\sum_{\ell=1}^{k}\mathbb{E}\left[(Z_\ell^o)^6\right]Q_{\ell j}^3 + 15\sum_{\ell\neq m}\mathbb{E}\left[(Z_\ell^o)^4\right]Q_{\ell j}^2Q_{mj}\right. \\
& \left. +15\left(\left(\sum_{\ell=1}^{k}Q_{\ell j}\right)^3 - \sum_{\ell=1}^{k}Q_{\ell j}^3 - 3\sum_{\ell\neq m}Q_{\ell j}^2Q_{mj}\right)\right) - \frac{3}{2}k \\
= & \, 2\sum_{\ell=1}^{k}\left(\mathbb{E}\left[(Z_\ell^o)^4\right]-3\right)\sum_{j=1}^{k}Q_{\ell j}^2 + \frac{1}{2}\left(\sum_{\ell=1}^{k}\left(15-\mathbb{E}\left[(Z_\ell^o)^6\right]\right)\sum_{j=1}^{k}Q_{\ell j}^3\right. \\
& \left. +15\sum_{\ell=1}^{k}\left(3-\mathbb{E}\left[(Z_\ell^o)^4\right]\right)\sum_{j=1}^{k}\sum_{m=1,m\neq\ell}^{k}Q_{\ell j}^2Q_{mj}\right) - 3k \\
= & \, 2\sum_{\ell=1}^{k}\left(\mathbb{E}\left[(Z_\ell^o)^4\right]-3\right)\sum_{j=1}^{k}\left(Q_{\ell j}^2 - \frac{15}{4}\sum_{\substack{m=1\\m\neq\ell}}^{k}Q_{\ell j}^2Q_{mj}\right) \\
& + \frac{1}{2}\sum_{\ell=1}^{k}\left(15-\mathbb{E}\left[(Z_\ell^o)^6\right]\right)\sum_{j=1}^{k}Q_{\ell j}^3 - 3k\,.
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
F(Q) &= 2\sum_{\ell=1}^{k}\left(\mathbb{E}\left[(Z_\ell^o)^4\right]-3\right)\sum_{j=1}^{k}Q_{\ell j}^2\left(1-\frac{15}{4}\sum_{\substack{m=1\\m\neq\ell}}^{k}Q_{mj}\right)\\
&\quad+\frac{1}{2}\sum_{\ell=1}^{k}\left(15-\mathbb{E}\left[(Z_\ell^o)^6\right]\right)\sum_{j=1}^{k}Q_{\ell j}^3-3k\\
&= 2\sum_{\ell=1}^{k}\left(\mathbb{E}\left[(Z_\ell^o)^4\right]-3\right)\sum_{j=1}^{k}Q_{\ell j}^2\left(1-\frac{15}{4}\left(1-Q_{\ell j}\right)\right)\\
&\quad+\frac{1}{2}\sum_{\ell=1}^{k}\left(15-\mathbb{E}\left[(Z_\ell^o)^6\right]\right)\sum_{j=1}^{k}Q_{\ell j}^3-3k\\
&= \frac{1}{2}\sum_{\ell=1}^{k}\left(\mathbb{E}\left[(Z_\ell^o)^4\right]-3\right)\sum_{j=1}^{k}\left(15Q_{\ell j}^3-11Q_{\ell j}^2\right)\\
&\quad+\frac{1}{2}\sum_{\ell=1}^{k}\left(15-\mathbb{E}\left[(Z_\ell^o)^6\right]\right)\sum_{j=1}^{k}Q_{\ell j}^3-3k\\
&= \frac{1}{2}\sum_{\ell=1}^{k}\left(15\mathbb{E}\left[(Z_\ell^o)^4\right]-\mathbb{E}\left[(Z_\ell^o)^6\right]-30\right)\sum_{j=1}^{k}Q_{\ell j}^3\\
&\quad+\frac{11}{2}\sum_{\ell=1}^{k}\left(3-\mathbb{E}\left[(Z_\ell^o)^4\right]\right)\sum_{j=1}^{k}Q_{\ell j}^2-3k\,.
\end{aligned}
$$

Now consider the following two cases:

I. Suppose that $\mathbb{E}\left[(Z_\ell^o)^4\right] \leq 3$. Then

$$\begin{aligned} F(Q) &\leq \frac{1}{2}\sum_{\ell=1}^{k}\left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right)\sum_{j=1}^{k} Q_{\ell j} \\ &\quad + \frac{11}{2}\sum_{\ell=1}^{k}\left(3 - \mathbb{E}\left[(Z_\ell^o)^4\right]\right)\sum_{j=1}^{k} Q_{\ell j} - 3k \qquad (3) \\ &= \frac{1}{2}\sum_{\ell=1}^{k}\left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right) + \frac{11}{2}\sum_{\ell=1}^{k}\left(3 - \mathbb{E}\left[(Z_\ell^o)^4\right]\right) - 3k \end{aligned}$$

where (3) follows from Condition (v) of Assumption 2. Notice that for any $O \in \mathcal{P}(k)$,

$$\begin{aligned} F(O^{(2)}) &= \frac{1}{2}\sum_{\ell=1}^{k}\left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right)\sum_{j=1}^{k} O_{\ell j}^6 \\ &\quad + \frac{11}{2}\sum_{\ell=1}^{k}\left(3 - \mathbb{E}\left[(Z_\ell^o)^4\right]\right)\sum_{j=1}^{k} O_{\ell j}^4 - 3k \\ &= \frac{1}{2}\sum_{\ell=1}^{k}\left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right) + \frac{11}{2}\sum_{\ell=1}^{k}\left(3 - \mathbb{E}\left[(Z_\ell^o)^4\right]\right) - 3k \end{aligned}$$

and hence

$$\begin{aligned} \max_{Q\in\mathcal{Q}(k)} F(Q) &= \frac{1}{2}\sum_{\ell=1}^{k}\left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right) \\ &\quad + \frac{11}{2}\sum_{\ell=1}^{k}\left(3 - \mathbb{E}\left[(Z_\ell^o)^4\right]\right) - 3k\,. \end{aligned}$$

For any $O \notin \mathcal{P}(k)$,

$$\begin{aligned}
F(O^{(2)}) &= \frac{1}{2}\sum_{\ell=1}^{k}\left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right)\sum_{j=1}^{k} O_{\ell j}^6 \\
&\quad + \frac{11}{2}\sum_{\ell=1}^{k}\left(3 - \mathbb{E}\left[(Z_\ell^o)^4\right]\right)\sum_{j=1}^{k} O_{\ell j}^4 - 3k \\
&< \frac{1}{2}\sum_{\ell=1}^{k}\left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right)\sum_{j=1}^{k} O_{\ell j} \\
&\quad + \frac{11}{2}\sum_{\ell=1}^{k}\left(3 - \mathbb{E}\left[(Z_\ell^o)^4\right]\right)\sum_{j=1}^{k} O_{\ell j} - 3k \\
&= \max_{Q\in\mathcal{Q}(k)} F(Q)\ .
\end{aligned}$$

Therefore, the set of solutions that maximize F is $\left\{O^{(2)} : O \in \mathcal{P}(k)\right\}$.

II. Suppose that $\mathbb{E}\left[(Z_\ell^o)^4\right] > 3$. Then

$$\begin{aligned}
F(Q) &= \frac{1}{2}\sum_{\ell=1}^{k}\left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right)\sum_{j=1}^{k} Q_{\ell j}^3 \\
&\quad - \frac{11}{2}\sum_{\ell=1}^{k}\left(\mathbb{E}\left[(Z_\ell^o)^4\right] - 3\right)\sum_{j=1}^{k} Q_{\ell j}^2 - 3k \\
&\leq \frac{1}{2}\sum_{\ell=1}^{k}\left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right)\sum_{j=1}^{k} Q_{\ell j}^3 \\
&\quad - \frac{11}{2}\sum_{\ell=1}^{k}\left(\mathbb{E}\left[(Z_\ell^o)^4\right] - 3\right)\sum_{j=1}^{k} Q_{\ell j}^3 - 3k \\
&= \frac{1}{2}\sum_{\ell=1}^{k}\left(4\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] + 3\right)\sum_{j=1}^{k} Q_{\ell j}^3 - 3k \\
&\leq \frac{1}{2}\sum_{\ell=1}^{k}\left(4\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] + 3\right)\sum_{j=1}^{k} Q_{\ell j} - 3k \qquad (4) \\
&= \frac{1}{2}\sum_{\ell=1}^{k}\left(4\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] + 3\right) - 3k
\end{aligned}$$

where (4) follows from Condition (v) of Assumption 2. For any $O \in \mathcal{P}(k)$,

$$\begin{aligned}
F(O^{(2)}) &= \frac{1}{2}\sum_{\ell=1}^{k}\left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right)\sum_{j=1}^{k} O_{\ell j}^6 \\
&\quad - \frac{11}{2}\sum_{\ell=1}^{k}\left(\mathbb{E}\left[(Z_\ell^o)^4\right] - 3\right)\sum_{j=1}^{k} O_{\ell j}^4 - 3k \\
&= \frac{1}{2}\sum_{\ell=1}^{k}\left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right) \\
&\quad - \frac{11}{2}\sum_{\ell=1}^{k}\left(\mathbb{E}\left[(Z_\ell^o)^4\right] - 3\right) - 3k \\
&= \frac{1}{2}\sum_{\ell=1}^{k}\left(4\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] + 3\right) - 3k
\end{aligned}$$

and hence

$$\max_{Q \in \mathcal{Q}(k)} F(Q) = \frac{1}{2} \sum_{\ell=1}^{k} \left(4\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] + 3\right) - 3k \; .$$

For any $O \notin \mathcal{P}(k)$,

$$\begin{aligned}
F(O^{(2)}) &= \frac{1}{2} \sum_{\ell=1}^{k} \left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right) \sum_{j=1}^{k} O_{\ell j}^6 \\
&\quad - \frac{11}{2} \sum_{\ell=1}^{k} \left(\mathbb{E}\left[(Z_\ell^o)^4\right] - 3\right) \sum_{j=1}^{k} O_{\ell j}^4 - 3k \\
&< \frac{1}{2} \sum_{\ell=1}^{k} \left(15\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] - 30\right) \sum_{j=1}^{k} O_{\ell j}^6 \\
&\quad - \frac{11}{2} \sum_{\ell=1}^{k} \left(\mathbb{E}\left[(Z_\ell^o)^4\right] - 3\right) \sum_{j=1}^{k} O_{\ell j}^6 - 3k \\
&= \frac{1}{2} \sum_{\ell=1}^{k} \left(4\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] + 3\right) \sum_{j=1}^{k} O_{\ell j}^6 - 3k \\
&< \frac{1}{2} \sum_{\ell=1}^{k} \left(4\mathbb{E}\left[(Z_\ell^o)^4\right] - \mathbb{E}\left[(Z_\ell^o)^6\right] + 3\right) \sum_{j=1}^{k} O_{\ell j} - 3k \\
&= \max_{Q \in \mathcal{Q}(k)} F(Q)
\end{aligned}$$

and so the set of solutions that maximize F is $\left\{O^{(2)} : O \in \mathcal{P}(k)\right\}$.

In both cases, the expectation of the Entromin$_2$ objective as a function of a doubly stochastic matrix $O^{(2)}$ is maximized when $O \in \mathcal{P}(k)$. The matrix O is defined as $O = \widetilde{R}^T R$ and thus $O \in \mathcal{P}(k)$ if and only if $R = \widetilde{R}P$ for some $P \in \mathcal{P}(k)$. It then follows that

$$\begin{aligned}
\operatorname*{arg\,min}_{R \in \mathcal{O}(k)} \; \mathbb{E}_{Z_1}\left[h_2(Z^o \widetilde{R}^T, R)\right] &= \operatorname*{arg\,max}_{R \in \mathcal{O}(k)} \; \mathbb{E}_{Z_1}\left[f^{h_2}_{Z^o \widetilde{R}^T}(R)\right] \\
&= \left\{\widetilde{R}P : P \in \mathcal{P}(k)\right\} \; .
\end{aligned}$$

□

A.4 Proof of Proposition 5

Proof. Define a random variable B that takes values

$$B = \begin{cases} 1 & \text{if } X \neq 0 \\ 0 & \text{if } X = 0 \end{cases} .$$

Then $B \sim \text{Bernoulli}(p)$ where $p = \mathbb{P}(X \neq 0)$. Define a random variable Y where

1. when $B = 1$ then $Y = X$, otherwise
2. when $B = 0$ then Y is equal in distribution to X on the set $X \neq 0$.

It then follows that

$$X \overset{\text{d}}{=} BY$$

where B and Y are independent by construction. Also, because X has four finite moments, Y also has four finite moments. Note that because $\mathbb{E}[X] = 0$ by assumption and that

$$\mathbb{E}[X] = p\mathbb{E}[Y] ,$$

we must have $\mathbb{E}[Y] = 0$. Also,

$$\begin{aligned} \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] &= \mathbb{E}\left[X^2\right] = p\mathbb{E}\left[Y^2\right] , \\ \mathbb{E}\left[(X - \mathbb{E}[X])^4\right] &= \mathbb{E}\left[X^4\right] = p\mathbb{E}\left[Y^4\right] . \end{aligned}$$

We then have

$$\begin{aligned} & \mathbb{E}[(X - \mathbb{E}[X])^4] - 3\mathbb{E}[(X - \mathbb{E}[X])^2]^2 > 0 \\ \Leftrightarrow \quad & p\mathbb{E}\left[Y^4\right] - 3p^2\mathbb{E}\left[Y^2\right]^2 > 0 \\ \Leftrightarrow \quad & \mathbb{E}\left[Y^4\right] - 3p\mathbb{E}\left[Y^2\right]^2 > 0 \\ \Leftrightarrow \quad & \frac{\mathbb{E}\left[Y^4\right]}{3\mathbb{E}\left[Y^2\right]^2} > p . \end{aligned}$$

The sequence of inequalities say that X is leptokurtic if and only if the probability $\mathbb{P}(X \neq 0)$ is less than $\frac{1}{3}$ of the kurtosis of Y, which is what was to be shown. $\square$

A.5 Proof of Proposition 7

Proof. The independent random variables $B \sim \text{Bernoulli}(p)$ and Y can be defined through the same setup as in the proof for Proposition 5. By assumption,

$$\begin{aligned}
\mathbb{E}[X] = \mathbb{E}[Y] &= 0 \ , \\
\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] &= p\mathbb{E}\left[Y^2\right] \ , \\
\mathbb{E}\left[(X - \mathbb{E}[X])^4\right] &= p\mathbb{E}\left[Y^4\right] \ , \\
\mathbb{E}\left[(X - \mathbb{E}[X])^6\right] &= \mathbb{E}\left[X^6\right] = p\mathbb{E}\left[Y^6\right] \ .
\end{aligned}$$

Then

$$\begin{aligned}
& -\frac{\mathbb{E}\left[(X - \mathbb{E}[X])^6\right]}{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]^3} + 15\frac{\mathbb{E}\left[(X - \mathbb{E}[X])^4\right]}{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]^2} - 30 > 0 \\
\Leftrightarrow \quad & -\frac{p\mathbb{E}\left[Y^6\right]}{p^3\mathbb{E}\left[Y^2\right]^3} + 15\frac{p\mathbb{E}\left[Y^4\right]}{p^2\mathbb{E}\left[Y^2\right]^2} - 30 > 0 \\
\Leftrightarrow \quad & \frac{\widetilde{\mu}_6}{30} - \frac{\widetilde{\mu}_4}{2}p + p^2 < 0 \ .
\end{aligned}$$

The final inequality holds when

$$\frac{\widetilde{\mu}_4}{4}\left(1 - \sqrt{1 - \frac{8\widetilde{\mu}_6}{15\widetilde{\mu}_4^2}}\right) < p < \frac{\widetilde{\mu}_4}{4}\left(1 + \sqrt{1 - \frac{8\widetilde{\mu}_6}{15\widetilde{\mu}_4^2}}\right) \ .$$

Notice that for any $\widetilde{\mu}_4$ and $\widetilde{\mu}_6$ we have

$$\frac{\widetilde{\mu}_4}{4}\left(1 - \sqrt{1 - \frac{8\widetilde{\mu}_6}{15\widetilde{\mu}_4^2}}\right) < \frac{\widetilde{\mu}_4}{3} \ .$$

Hence, the condition on p becomes

$$\frac{\widetilde{\mu}_4}{3} \leq p < \frac{\widetilde{\mu}_4}{4}\left(1 + \sqrt{1 - \frac{8\widetilde{\mu}_6}{15\widetilde{\mu}_4^2}}\right)$$

which is a valid interval and is satisfied by assumption. Therefore, by the sequence of inequalities above, it holds that

$$-\frac{\mathbb{E}\left[(X-\mathbb{E}[X])^6\right]}{\mathbb{E}\left[(X-\mathbb{E}[X])^2\right]^3}+15\frac{\mathbb{E}\left[(X-\mathbb{E}[X])^4\right]}{\mathbb{E}\left[(X-\mathbb{E}[X])^2\right]^2}-30>0\ .$$

□

A.6 Proof of Lemma 11, Theorem 9 and Lemma 12

The proofs for Theorem 9 and Lemma 12 make use of Lemma 11 which we prove first.

A.6.1 Lemma 11 and proof

Lemma 11. *Let the Inagaki entropy criterion be defined as in Equation 8. Let $X \in \mathbb{R}^k$ be a vector of k independent standard normal random variables. Then*

$$\mathbb{E}_X\left[h_{IG}(X,R)\right]=\psi\left(\frac{k}{2}+1\right)-\psi\left(\frac{1}{2}+1\right)$$

for any rotation matrix $R \in \mathcal{O}(k)$.

Proof. Because X_j for $j \in \{1,\ldots,k\}$ are independent standard normal random variables, X_j^2 are independent chi-square random variables with one degree of freedom or equivalently,

$$X_j^2 \sim \text{Gamma}\left(\frac{1}{2},2\right)$$

and also

$$\sum_{j=1}^{k} X_j^2 \sim \text{Gamma}\left(\frac{k}{2},2\right)\ .$$

Define

$$P_j=\frac{X_j^2}{\sum_{\ell=1}^{k} X_\ell^2}\ .$$

As X_j^2 are independent Gamma random variables, we have

$$P = (P_1, \ldots, P_k) \sim \text{Dirichlet}(\alpha)$$

where $\alpha_j = \frac{1}{2}$ for all $j \in \{1, \ldots, k\}$.

By Maxwell's theorem, normal distributions are rotationally invariant and hence XR is equal in distribution to X. It then follows that the Inagaki entropy criterion for X and any $R \in \mathcal{O}(k)$ is

$$\begin{aligned}
\mathbb{E}_X \left[h_{IG}(X, R) \right] &= \mathbb{E}_X \left[h_{IG}(X, I_k) \right] \\
&= \mathbb{E}_X \left[-\sum_{j=1}^{k} \frac{X_j^2}{\sum_{\ell=1}^{k} X_\ell^2} \log \left(\frac{X_j^2}{\sum_{\ell=1}^{k} X_\ell^2} \right) \right] \\
&= \mathbb{E}_P \left[-\sum_{j=1}^{k} P_j \log P_j \right] .
\end{aligned}$$

Nemenman et al. (2001) showed that the expected Shannon entropy for a sample D of k probabilities from a $\text{Dirichlet}(\beta, \ldots, \beta)$ distribution is

$$\mathbb{E}_D \left[-\sum_{j=1}^{k} D_j \log D_j \right] = \psi(k\beta + 1) - \psi(\beta + 1)$$

and thus

$$\mathbb{E}_X \left[h_{IG}(X, R) \right] = \psi \left(\frac{k}{2} + 1 \right) - \psi \left(\frac{1}{2} + 1 \right) .$$

□

A.6.2 Proof of Theorem 9

Proof. For $j \in \{1, \ldots, k\}$, define independent random variables B_j that take values

$$B_j = \begin{cases} 1 & \text{if } Z_j^o \neq 0 \\ 0 & \text{if } Z_j^o = 0 \end{cases} .$$

Then by assumption, $B_j \sim \text{Bernoulli}(p)$ for some p in the interval

$$0 < p \leq 1 - \left(1 - \frac{e\left(\psi\left(\frac{k}{2}+1\right) - \psi\left(\frac{1}{2}+1\right)\right)}{k}\right)^{\frac{1}{k-1}} . \tag{20}$$

Also, define independent random variables Y_j where

1. when $B_j = 1$ then $Y_j = Z_j^o$, otherwise
2. when $B_j = 0$ then Y_j is equal in distribution to Z_j^o on the set $Z_j^o \neq 0$.

It then follows that

$$Z_j^o \overset{\text{d}}{=} B_j Y_j .$$

Define $\frac{0}{0} = 0$.[3] Then

$$\begin{aligned}
\mathbb{E}_{Z^o}\left[h_{IG}(Z^o, I_k)\right] &= \mathbb{E}_{Z^o}\left[-\sum_{j=1}^{k} \frac{(Z_j^o)^2}{\sum_{\ell=1}^{k}(Z_\ell^o)^2} \log\left(\frac{(Z_j^o)^2}{\sum_{\ell=1}^{k}(Z_\ell^o)^2}\right)\right] \\
&= \mathbb{E}_{B,Y}\left[-\sum_{j=1}^{k} \frac{B_j Y_j^2}{\sum_{\ell=1}^{k} B_\ell Y_\ell^2} \log\left(\frac{B_j Y_j^2}{\sum_{\ell=1}^{k} B_\ell Y_\ell^2}\right)\right] .
\end{aligned}$$

Notice that

$$\mathbb{E}_B\left[\mathbb{E}_Y\left[-\sum_{j=1}^{k} \frac{B_j Y_j^2}{\sum_{\ell=1}^{k} B_\ell Y_\ell^2} \log\left(\frac{B_j Y_j^2}{\sum_{\ell=1}^{k} B_\ell Y_\ell^2}\right) \middle| \sum_{j=1}^{k} B_j \leq 1\right]\right] = 0 .$$

For notational convenience, denote

$$Q_j = \frac{B_j Y_j^2}{\sum_{\ell=1}^{k} B_\ell Y_\ell^2} .$$

[3] If $B_j = 0$ for all $j \in \{1, \ldots, k\}$, then the row only consists of zeros. There is no reason to perform a non-trivial rotation in this case and so it makes sense that the value of the criterion is the minimum value zero.

Then using the fact that

$$\sum_{j=1}^{k} B_j \sim \text{Binomial}(k, p) \ ,$$

it then follows that

$$\begin{aligned}
&\mathbb{E}_{Z^o}\left[h_{IG}(Z^o, I_k)\right] \\
&= \mathbb{E}_{B,Y}\left[-\sum_{j=1}^{k} Q_j \log Q_j\right] \\
&= \left(1-(1-p)^k - kp(1-p)^{k-1}\right)\mathbb{E}_B\left[\mathbb{E}_Y\left[-\sum_{j=1}^{k} Q_j \log Q_j \middle| \sum_{j=1}^{k} B_j \geq 2\right]\right] \\
&= \left(1-(1+(k-1)p)(1-p)^{k-1}\right)\mathbb{E}_B\left[\mathbb{E}_Y\left[-\sum_{j=1}^{k} Q_j \log Q_j \middle| \sum_{j=1}^{k} B_j \geq 2\right]\right] \\
&\leq \left(1-(1-p)^{k-1}\right)\mathbb{E}_B\left[\mathbb{E}_Y\left[-\sum_{j=1}^{k} Q_j \log Q_j \middle| \sum_{j=1}^{k} B_j \geq 2\right]\right] \ .
\end{aligned}$$

The function $H(x) = -x \log x$ is concave with the maximum at $H(e^{-1}) = e^{-1}$ and so

$$\mathbb{E}_{Z^o}\left[h_{IG}(Z^o, I_k)\right] \leq \left(1-(1-p)^{k-1}\right) k e^{-1} \ .$$

Substituting the upper bound from Equation 20 for p and applying Lemma 11 then gives

$$\begin{aligned}
\mathbb{E}_{Z^o}\left[h_{IG}(Z^o, I_k)\right] &\leq \psi\left(\frac{k}{2}+1\right) - \psi\left(\frac{1}{2}+1\right) \\
&= \mathbb{E}_X\left[h_{IG}(X, R)\right] \ .
\end{aligned}$$

□

A.6.3 Lemma 12 and proof

Lemma 12. *Let the Inagaki entropy criterion be defined as in Equation 8. Let $X \in \mathbb{R}^k$ be a vector of k independent standard normal random variables and let $B \in \{0,1\}^k$ be a vector of k independent* Bernoulli(p) *random variables for some $p \in [0,1]$. Define $Y_j = X_j B_j$ for $j \in \{1,\ldots,k\}$ and $Y = (Y_1,\ldots,Y_k)$. Then*

$$\mathbb{E}_Y\left[h_{IG}(Y, I_k)\right] = \sum_{j=2}^{k} \binom{k}{j} p^j (1-p)^{k-j} \left(\psi\left(\frac{j}{2}+1\right) - \psi\left(\frac{1}{2}+1\right)\right) .$$

Proof. Note that $\sum_{\ell=1}^{k} B_\ell \sim \text{Binomial}(k,p)$. Then

$$\begin{aligned}
\mathbb{E}_Y\left[h_{IG}(Y, I_k)\right] &= \mathbb{E}_B\left[\mathbb{E}_X\left[h_{IG}(XB, I_k)\,|B\right]\right] \\
&= \sum_{j=0}^{k} \mathbb{P}\left(\sum_{\ell=1}^{k} B_\ell = j\right) \mathbb{E}_X\left[h_{IG}(XB, I_k) \,\middle|\, \sum_{\ell=1}^{k} B_\ell = j\right] \\
&= \sum_{j=2}^{k} \binom{k}{j} p^j (1-p)^{k-j} \left(\psi\left(\frac{j}{2}+1\right) - \psi\left(\frac{1}{2}+1\right)\right)
\end{aligned}$$

where the final equality follows from Lemma 11 and because $h_{IG}(0, I_k) = 0$. □

A.7 Proof of Proposition 10

Proof. When U is orthonormal, the objective function can be written as

$$f_U^{h_2}(R) = \sum_{i=1}^{n}\sum_{j=1}^{k} \left(-\frac{1}{2}[UR]_{ij}^6 + 2[UR]_{ij}^4\right) - \frac{3}{2}k .$$

Consider the function

$$g(x) = -\frac{1}{2}x^6 + 2x^4$$

which has first and second derivatives

$$\begin{aligned}
\frac{dg}{dx} &= -3x^5 + 8x^3 , \\
\frac{d^2g}{dx^2} &= -15x^4 + 24x^2 .
\end{aligned}$$

For $x \in [-1, 1]$,

$$\begin{aligned}
\frac{d^2 g}{dx^2} &= -15x^4 + 24x^2 \\
&\geq -15x^2 + 24x^2 \\
&= 9x^2 \\
&\geq 0
\end{aligned}$$

and so g is convex on the interval $[-1, 1]$. Because U is orthonormal and $R \in \mathcal{B}_{\mathcal{O}}(k)$, $[UR]_{ij} \in [-1, 1]$ for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, k\}$. Hence, $g([UR]_{ij})$ is convex as a function of R. The objective $f_U^{h_2}$ is the sum of convex functions of R and thus $f_U^{h_2}$ is convex. $\square$

Appendix B

Derivations of finite-order approximations of Entromin

This chapter of the Appendix derives the criterion for the second-order approximation of Entromin and the expected criteria of finite-order approximations.

B.1 Second-order approximation of Entromin

The Entromin$_2$ criterion in Equation 13 is obtained from Equation 10 through the steps

$$
\begin{aligned}
h_2(U,R) &= \sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^2\left(\sum_{q=1}^{2}\frac{(-1)^q}{q}\left([UR]_{ij}^2-1\right)^q\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^2\left(\frac{1}{2}\left([UR]_{ij}^2-1\right)^2-\left([UR]_{ij}^2-1\right)\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}[UR]_{ij}^2\left(\frac{1}{2}\left([UR]_{ij}^4-2[UR]_{ij}^2+1\right)-\left([UR]_{ij}^2-1\right)\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\left(\frac{1}{2}[UR]_{ij}^6-2[UR]_{ij}^4+\frac{3}{2}[UR]_{ij}^2\right)\ .
\end{aligned}
$$

B.2 Expected finite-order approximation of Entromin

We derive the expectation of the criterion for the N^{th}-order approximation of Entromin.

Let Z be a row random variable with independent elements and let $R \in \mathcal{O}(k)$ be any rotation. Define

$$g_q(Z, R) = \frac{(-1)^q}{q} \sum_{j=1}^{k} [ZR]_j^2 \left([ZR]_j^2 - 1 \right)^q$$

and so from Equation 10, the N^{th}-order criterion is given by

$$h_N(Z, R) = \sum_{q=1}^{N} g_q(Z, R) \ .$$

By the binomial theorem, the expectation of $g_q(Z, R)$ with respect to Z is

$$\begin{aligned}
\mathbb{E}_Z \left[g_q(Z, R) \right] &= \frac{(-1)^q}{q} \sum_{j=1}^{k} \mathbb{E} \left[[ZR]_j^2 \left([ZR]_j^2 - 1 \right)^q \right] \\
&= \frac{(-1)^q}{q} \sum_{j=1}^{k} \mathbb{E} \left[[ZR]_j^2 \sum_{i=0}^{q} \binom{q}{i} [ZR]_j^{q-i} (-1)^i \right] \\
&= \sum_{j=1}^{k} \sum_{i=0}^{q} \frac{(-1)^{q+i}}{q} \binom{q}{i} \mathbb{E}_Z \left[[ZR]_j^{2+q-i} \right] \\
&= \sum_{j=1}^{k} \sum_{i=0}^{q} \frac{(-1)^{q+i}}{q} \binom{q}{i} \mathbb{E}_Z \left[\left(\sum_{\ell=1}^{k} Z_\ell R_{\ell j} \right)^{2+q-i} \right] \ .
\end{aligned}$$

For $a_1, \ldots, a_k \in \mathbb{N}$, let

$$\binom{2+q-i}{a_1, \ldots, a_k} = \frac{(2+q-i)!}{a_1! \ldots a_k!}$$

denote the multinomial coefficient. Applying the multinomial theorem to the ex-

pectation gives

$$\mathbb{E}_Z\left[\left(\sum_{\ell=1}^{k} Z_\ell R_{\ell j}\right)^{2+q-i}\right] = \mathbb{E}_Z\left[\sum_{a_1+\ldots+a_k=2+q-i}\binom{2+q-i}{a_1,\ldots,a_k}\prod_{\ell=1}^{k}\left(Z_\ell R_{\ell j}\right)^{a_\ell}\right]$$
$$= \sum_{a_1+\ldots+a_k=2+q-i}\binom{2+q-i}{a_1,\ldots,a_k}\prod_{\ell=1}^{k}\mathbb{E}_{Z_\ell}\left[Z_\ell^{a_\ell}\right]R_{\ell j}^{a_\ell}$$

where the last equality follows by independence of the elements of Z. It then follows that

$$\mathbb{E}_Z\left[h_N(Z,R)\right]$$
$$= \sum_{q=1}^{N}\sum_{j=1}^{k}\sum_{i=0}^{q}\sum_{a_1+\ldots+a_k=2+q-i}\frac{(-1)^{q+i}}{q}\binom{q}{i}\binom{2+q-i}{a_1,\ldots,a_k}\prod_{\ell=1}^{k}\mathbb{E}_{Z_\ell}\left[Z_\ell^{a_\ell}\right]R_{\ell j}^{a_\ell}\ .$$

Appendix C

Theorem 9 and maximum entropy distributions

Theorem 9 resembles a statement about the maximum entropy distribution under the given assumptions. We discuss the connection briefly in this chapter of the Appendix.

Definition 10. Let the Shannon entropy for a discrete random variable be defined as in Equation 7, and define the *differential entropy* for a continuous random variable X with density f as

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx \ .$$

The *maximum entropy distribution* given a class of discrete (continuous) probability distributions is the distribution that has a Shannon (differential) entropy at least as great as that of all the other distributions in the class. ⌟

Interpreting the theorem as a maximum entropy result should be done with caution. The theorem is a result about the expected Inagaki entropy criterion over samples drawn from the underlying distribution rather than about the underlying distribution itself, and the Inagaki entropy criterion is not equivalent to the Shannon entropy that is used to measure the entropy of discrete probability distributions. It

is valid to interpret the theorem as saying that the discrete distribution generated from the sample of a Gaussian factor (by squaring and normalizing the samples) is expected to be the maximum entropy distribution over all discrete distributions generated from the sample of any other factor distribution that has inherent sparsity.

On the other hand, the expected Inagaki entropy criterion for normal random variables can also be recognized as the expected Shannon entropy for a sample drawn from a symmetric Dirichlet distribution with parameter $\alpha = \frac{1}{2}$ (Lemma 11). While the normal distribution is the maximum entropy distribution with support $\mathbb{R}$ under a second moment constraint (Dowson and Wragg, 1973), the expected Shannon entropy of the sample from the Dirichlet distribution is less than the entropy of the discrete uniform distribution (Nemenman et al., 2001), which is the maximum entropy distribution with discrete support. The expected Shannon entropy can be made the entropy of the discrete uniform distribution by taking the Dirichlet parameter $\alpha \to \infty$. From this perspective where the generative procedure of the discrete distribution is ignored, the Dirichlet distribution corresponding to the Inagaki entropy upper bound in Theorem 9 is not a maximum entropy distribution.

Appendix D

Algorithm details

This chapter of the Appendix contains additional details about the algorithms discussed in Chapter 4.

D.1 Gradients for pairwise optimization

We derive the gradient G used in the pairwise optimization algorithm (Algorithm 1 and Algorithm 2) in this section.

Let X_j and X_ℓ denote the j^{th} and ℓ^{th} columns of the input matrix $X \in \mathbb{R}^{n\times k}$, and let $[X_j, X_\ell] \in \mathbb{R}^{n\times 2}$ denote the matrix obtained by horizontally stacking the two columns. For $\theta \in [0, 2\pi]$, the rotation matrix $R_{[X_j,X_\ell]}$ has the form

$$R_{[X_j,X_\ell]} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} .$$

The two-dimensional rotation of $[X_j, X_\ell]$ is then

$$[X_j, X_\ell]R_{[X_j,X_\ell]} = \begin{bmatrix} X_{1j}\cos(\theta) + X_{1\ell}\sin(\theta) & -X_{1j}\sin(\theta) + X_{1\ell}\cos(\theta) \\ \vdots & \vdots \\ X_{nj}\cos(\theta) + X_{n\ell}\sin(\theta) & -X_{nj}\sin(\theta) + X_{n\ell}\cos(\theta) \end{bmatrix} .$$

For notational convenience, denote the rotated elements as

$$X'_{ij} = X_{ij}\cos(\theta) + X_{i\ell}\sin(\theta) \ ,$$
$$X'_{i\ell} = -X_{ij}\sin(\theta) + X_{i\ell}\cos(\theta) \ .$$

The derivatives of the rotated elements with respect to θ are then

$$\frac{dX'_{ij}}{d\theta} = -X_{ij}\sin(\theta) + X_{i\ell}\cos(\theta) = X'_{i\ell} \ ,$$
$$\frac{dX'_{i\ell}}{d\theta} = -X_{ij}\cos(\theta) - X_{i\ell}\sin(\theta) = -X'_{ij} \ .$$

D.1.1 Varimax

Assume that the columns of the input matrix X are orthonormal and so the Varimax criterion is equivalent to the Quartimax criterion by Proposition 1. The Varimax objective as a function of θ is then

$$f^v_{[X_j,X_\ell]}(\theta) = \sum_{i=1}^{n}\left((X'_{ij})^4 + (X'_{i\ell})^4\right) + \frac{2}{n^2}$$

and its gradient with respect to θ is then given by

$$G_{j\ell} = \frac{df^v_{[X_j,X_\ell]}}{d\theta} = 4\sum_{i=1}^{n} X'_{ij}X'_{i\ell}\left((X'_{ij})^2 + (X'_{i\ell})^2\right) \ .$$

For second-order numerical optimization techniques, the second derivative of the objective with respect to θ is

$$\frac{d^2 f^v_{[X_j,X_\ell]}}{d\theta^2} = 12\sum_{i=1}^{n}(X'_{ij}X'_{i\ell})^2 - 4f^v_{[X_j,X_\ell]}(\theta) + \frac{8}{n^2} \ .$$

D.1.2 Entromin

The Entromin objective as a function of θ is

$$f^h_{[X_j,X_\ell]}(\theta) = \sum_{i=1}^{n} \left((X'_{ij})^2 \log(X'_{ij})^2 + (X'_{i\ell})^2 \log(X'_{i\ell})^2\right)$$

and its gradient with respect to θ is given by

$$G_{j\ell} = \frac{df^h_{[X_j,X_\ell]}}{d\theta} = 2\sum_{i=1}^{n} X'_{ij} X'_{i\ell} \log\left(\frac{X'_{ij}}{X'_{i\ell}}\right)^2 .$$

For second-order numerical optimization techniques, the second derivative of the objective with respect to θ is

$$\frac{d^2 f^h_{[X_j,X_\ell]}}{d\theta^2} = 6\sum_{i=1}^{n} \left((X'_{i\ell})^2 \log(X'_{ij})^2 + (X'_{ij})^2 \log(X'_{i\ell})^2\right) + 2f^h_{[X_j,X_\ell]}(\theta) .$$

Notice that the second derivative does not exist if exactly one of X'_{ij} and $X'_{i\ell}$ is zero for any i.

D.1.3 Entromin$_2$

Assume that the columns of the input matrix X are orthonormal. The Entromin$_2$ objective as a function of θ is

$$f^{h_2}_{[X_j,X_\ell]}(\theta) = -\sum_{i=1}^{n} \left(\frac{1}{2}\left((X'_{ij})^6 + (X'_{i\ell})^6\right) - 2\left((X'_{ij})^4 + (X'_{i\ell})^4\right)\right) - 3$$

and its gradient with respect to θ is given by

$$G_{j\ell} = \frac{df^{h_2}_{[X_j,X_\ell]}}{d\theta} = \sum_{i=1}^{n} \left(X'_{ij}\left(3(X'_{i\ell})^5 - 8(X'_{i\ell})^3\right) - X'_{i\ell}\left(3(X'_{ij})^5 - 8(X'_{ij})^3\right)\right) .$$

For second-order numerical optimization techniques, the second derivative of the objective with respect to θ is

$$\frac{d^2 f^{h_2}_{[X_j, X_\ell]}}{d\theta^2} = \sum_{i=1}^{n} \Bigg(3\left((X'_{ij})^6 + (X'_{i\ell})^6\right) - 8\left((X'_{ij})^4 + (X'_{i\ell})^4\right) + (X_{ij}X_{i\ell})^2 \left(48 - 15\left((X'_{ij})^2 - (X'_{i\ell})^2\right)\right) \Bigg) .$$

D.2 Gradients for gradient projection

We derive the gradient G used in the gradient projection algorithm (Algorithm 3 and Algorithm 4) in this section.

For any factor rotation criterion g of the form $g(U, R)$, define $g(X) = g(U, R)$ where $X = UR$. Jennrich (2001) showed that for an objective of the form

$$f_U(R) = g(U, R) ,$$

the gradient of f is given by

$$\nabla f_U(R) = U^T \frac{dg(UR)}{dX}$$

where $\frac{dg}{dX}$ is the $k \times k$ matrix of derivatives

$$\left(\frac{dg}{dX}\right)_{j\ell} = \frac{dg}{dX_{j\ell}}$$

for $j, \ell \in \{1, \ldots, k\}$.

D.2.1 Varimax

Assume that U has orthonormal columns. By Proposition 1, the Varimax objective is equivalent to the Quartimax objective. Then as a function of X, the derivative of

the Quartimax criterion with respect to a particular matrix element is

$$\frac{dq}{dX_{j\ell}} = 4X_{j\ell}^3 \ .$$

Dropping scaling factors, the gradient of the Varimax objective is then given by

$$G = \nabla f_U^v(R) = U^T[UR]^{(3)} \ .$$

D.2.2 Entromin

As a function of X, the derivative of the Entromin criterion with respect to a particular matrix element is

$$\frac{dh}{dX_{j\ell}} = -2\left(X_{j\ell}\log X_{j\ell}^2 + X_{j\ell}\right) \ .$$

Let $\odot$ denote the element-wise product and let $\log X$ denote the element-wise logarithm of a matrix X. Dropping scaling factors, the gradient of the Entromin objective is then given by

$$G = \nabla f_U^h(R) = U^T\left([UR] \odot \log[UR]^{(2)} + UR\right) \ .$$

D.2.3 Entromin$_2$

Assume that U has orthonormal columns and so the order-two term in the Entromin$_2$ criterion is constant. Then as a function of X, the derivative of the Entromin$_2$ criterion with respect to a particular matrix element is

$$\frac{dh_2}{dX_{j\ell}} = 3X_{j\ell}^5 - 8X_{j\ell}^3 \ .$$

Hence, the gradient of the Entromin$_2$ objective is given by

$$G = \nabla f_U^{h_2}(R) = U^T\left(8[UR]^{(3)} - 3[UR]^{(5)}\right) \ .$$

Appendix E

Empirical analysis details

This chapter of the Appendix contains additional details from the empirical analysis discussed in Chapter 5.

E.1 Additional results from convergence criterion analysis

Table 4 shows the numerical results from the convergence analysis in Section 5.1.1 for Varimax, Entromin$_2$ and Entromin when constant terms in the objectives are not dropped. Entromin$_2$ is sensitive to the scale of the objective and the gradient and terminates prematurely as a result.

		$v \times 10^8$	h_2	h	Soft zeros	Iterations
Varimax						
	C_1	3.986	29.995	188.028	39,297	17
	C_2	3.986	29.995	188.028	39,297	17
Entromin$_2$						
	C_1	3.986	29.995	188.028	39,282	22
	C_2	2.011	29.997	199.927	1982	1
Entromin						
	C_1	3.984	29.995	187.991	103,361	92
	C_2	3.984	29.995	187.992	101,470	91

Table 4: Results from the convergence criterion analysis on a simulated sparse Gaussian dataset $A \in \mathbb{R}^{50,000\times 20}$. C_1 and C_2 are the sum of singular values and the relative objective increase convergence criteria, respectively. The Varimax, Entromin$_2$ and Entromin criteria are denoted by v, h_2 and h, respectively. The number of soft zeros (threshold 10^{-5}) in the 20 rotated principal components and the number of iterations until convergence are also shown.

E.2 Convergence results from sparsity analysis

Table 5 shows the numerical convergence results from the sparsity analysis in Section 5.1.3 for Varimax, Entromin$_2$ and Entromin and for varying levels of sparsity.

		$v \times 10^6$	h_2	h	Soft zeros	Iterations
$p = 0.3$						
	Varimax	5.419	2.985	12.038	6	13
	Entromin_2	5.419	2.985	12.038	7	17
	Entromin	5.391	2.985	12.035	7	22
$p = 0.4$						
	Varimax	5.911	2.984	11.962	10	11
	Entromin_2	5.911	2.984	11.962	9	14
	Entromin	5.874	2.984	11.957	13	15
$p = 0.5$						
	Varimax	6.791	2.982	11.689	75	8
	Entromin_2	6.791	2.982	11.689	71	9
	Entromin	6.790	2.982	11.688	109	6

Table 5: Results from the sparsity analysis on the simulated sparse Gaussian dataset $A \in \mathbb{R}^{1000\times 3}$. The Varimax, Entromin_2 and Entromin criteria are denoted by v, h_2 and h, respectively. The number of soft zeros (threshold 10^{-4}) in the two rotated principal components and the number of iterations until convergence are also shown.

E.3 Additional plots from NYT articles analysis

This section contains the additional plots described in Section 5.2 that are used to determine the principal components to retain and rotate in the analysis of the New York Times articles dataset.

Figure 9 shows the L_4-norm of the original 50 principal components. Rohe and Zeng (2020) considered principal components with norms that exceeded a threshold of 0.15 as being localized. The localized components are discarded.

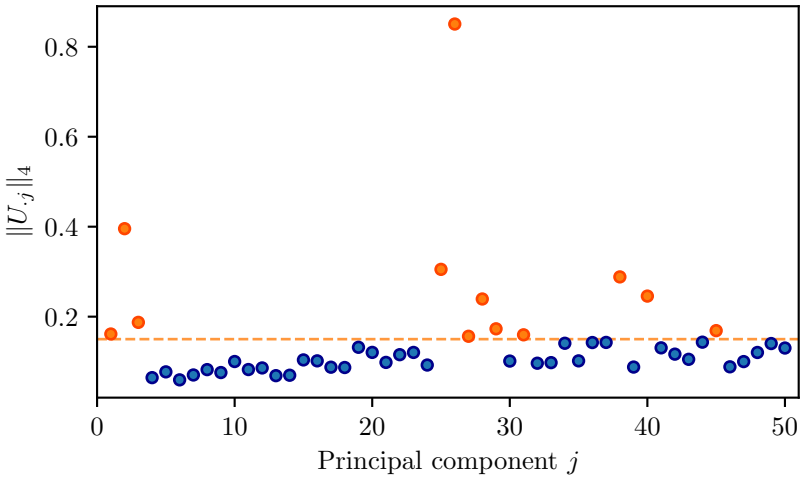

Figure 9: The L_4-norms of the first 50 principal components from the New York Times articles dataset. Principal components with norms greater than 0.15 are considered to be localized.

Figure 10 shows the leading singular values of the remaining non-localized 38 principal components. There is a notable eigengap after the eighth singular value and so only the first eight principal components are retained and rotated.

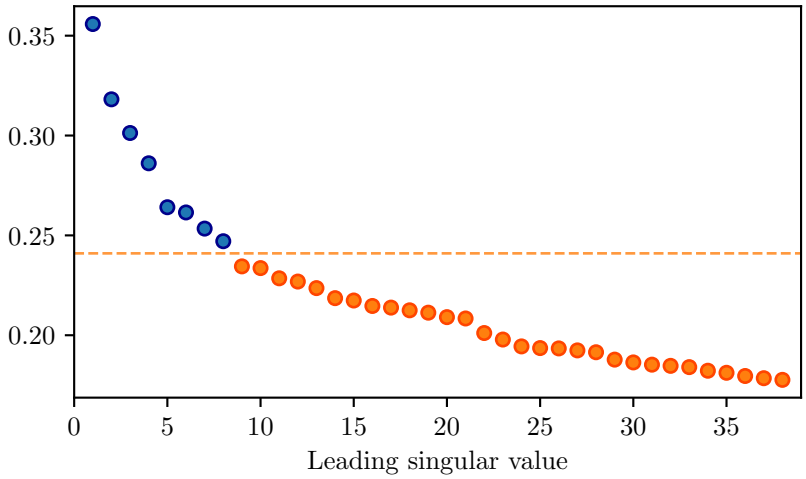

Figure 10: The leading singular values of the remaining 38 principal components from the New York Times articles dataset after 12 localized components were discarded. There is a notable eigengap after the eighth singular value.