Having Fun With A New Loss Function

Jan de Leeuw

First created on June 26, 2022. Last update on July 10, 2022

Abstract

The abstract

1 Introduction

1.1 Notation

Notation xy, x/y and x>1. Vectorized, Hadamard.

diag , MP, Loewner order

Notation D, D_1, D_11. Spivak.

1.2 Terminology

In this paper a *loss function* is a non-negative function σ on $\mathbb{X} \otimes \mathbb{Y}$, with $\mathbb{X} \subseteq \mathbb{R}^n$ and $\mathbb{Y} \subseteq \mathbb{R}^n$, where $\sigma(x, y) = 0$ iff $x = y$. A loss function is *additive* if there are a non-negative real-valued *base loss function*, also called σ, defined on $\mathfrak{x} \otimes \mathfrak{y}$, where $\mathfrak{x} \subseteq \mathbb{R}$ and $\mathfrak{y} \subseteq \mathbb{R}$, and there are positive weights w_i, such that

$$\sigma(x, y) = \sum_{i=1}^{n} w_i \sigma(x_i, y_i). \tag{1}$$

Loss functions are used to estimate *parameters*. This means that in our loss function x stands for the *model* and y for the *data*. The model $x = \xi(\theta)$ is a real-valued function from $\Theta \subseteq \mathbb{R}^p$ to the positive orthant of $\mathbb{R}^n$. We want to find θ such that $\xi(\theta)$ approximates the data y. In most applications the data are fixed constants, and it makes sense to define

$$\sigma(\theta) := \sigma(\xi(\theta), y). \tag{2}$$

Sorry for overloading σ even more, but which one of the three definitions we have in mind will always be clear from the context and/or the arguments.

In minimizing loss we are interested in computing

$$\sigma_\star = \inf_{\theta \in \Theta} \sigma(\theta), \tag{3}$$

and

$$\hat{\theta} := \operatorname*{Argmin}_{\theta \in \Theta} \sigma(\theta) = \{\theta \in \Theta \mid \sigma_\star = \sigma(\theta)\}. \tag{4}$$

Note that in general $\hat{\theta}$ is a subset of Θ, with possibly more than one element, and that $\hat{\theta}$ can be empty if the infimum is not attained at some θ.

1.3 Credo

Now if I say that we will discuss a *new* loss function, I obviously mean "new for me".

> What has been will be again, what has been done will be done again; there is nothing new under the sun. (Ecclesiastes 1: 9)

I do discuss some earlier work in which similar or even identical loss functions have been proposed and/or applied, but I am sure my literature review is far from complete. Thus, instead of referring to my loss function as a "new" loss function, I will just call it the "fun loss function", or FLF.

2 Fun Loss Function

2.1 The FLF

To construct the FLF we start with the *base function* f, defined on the positive real axis by $f(x) := x - \log(x) - 1$. There is a plot in figure 1.

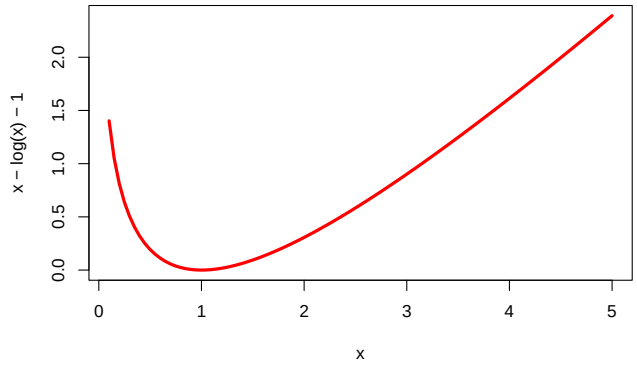


Figure 1: Base Function

We see that $f(x) \geq 0$ on the open interval $(0, +\infty)$, with $f(x) = 0$ iff $x = 1$. Since the logarithm is strictly concave base function f is strictly convex. At $x = 0$ there is a vertical asymptote, i.e. $f(0) = +\infty$ (so to speak).

The partial derivative is $\mathcal{D}f(x) = 1 - x^{-1}$ and thus $\mathcal{D}f$ strictly increases with a horizontal asymptote $\mathcal{D}f(\infty) = 1$ and with a vertical asymptote $\mathcal{D}f(0) = -\infty$. The second partial derivative is $\mathcal{D}^2 f(x) = x^{-2}$, which is positive and strictly decreases to $\mathcal{D}^2 f(+\infty)0$. More generally, for $s \geq 2$ we have $\mathcal{D}^s f(x) =) = (-1)^s (s-1)! x^{-s}$. Thus $\mathcal{D}^s f$ is strictly increasing and stricty concave for s odd and strictly decreasing and strictly convex for s even.

The function f is used to define the base loss function σ.

$$\sigma(x_i, y_i) := f(\frac{x_i}{y_i}) = \frac{x_i}{y_i} - \log \frac{x_i}{y_i} - 1. \tag{5}$$

Clearly $\sigma(x_i, y_i) \geq 0$ with equality iff $x_i = y_i$. Also the function is defined iff $x_i > 0$ and $y_i > 0$. The base loss function is used to define the final FLF.

$$\sigma(x, y) := \sum_{i=1}^{n} w_i \sigma(x_i, y_i) = \sum_{i=1}^{n} w_i \left\{ \frac{x_i}{y_i} - \log \frac{x_i}{y_i} - 1 \right\}. \tag{6}$$

We learn more about the FLF by looking at its gradient and Hessian, assuming throughout that x, y and w are strictly positive.

$$\mathcal{D}_1 \sigma(x, y) = w \left\{ \frac{1}{y} - \frac{1}{x} \right\}, \tag{7}$$

$$\mathcal{D}_2 \sigma(x, y) = \frac{w}{y} \left\{ 1 - \frac{x}{y} \right\}. \tag{8}$$

The second partials are

$$\mathcal{D}_{11} \sigma(x, y) = \text{diag}(\frac{w}{x^2}), \tag{9}$$

$$\mathcal{D}_{12} \sigma(x, y) = -\text{diag}(\frac{w}{y^2}), \tag{10}$$

$$\mathcal{D}_{22} \sigma(x, y) = \text{diag}\left(\frac{w}{y^2} \left\{ 2\frac{x}{y} - 1 \right\} \right). \tag{11}$$

Thus the four submatrices of the Hessian are all diagonal. We see that the FLF is always convex in x for fixed y, and convex in y for fixed x iff $x/y \geq \frac{1}{2}$. It is concave in y for fixed x iff $x/y \leq \frac{1}{2}$. Since $\mathcal{D}_{11} \sigma(x, y) > 0$ the FLF is jointly convex in x and y iff

$$\mathcal{D}_{21} \sigma(x, y) \{\mathcal{D}_{11} \sigma(x, y)\}^{-1} \mathcal{D}_{12} \sigma(x, y) \lesssim \mathcal{D}_{22} \sigma(x, y). \tag{12}$$

This simplifies to

$$\left(\frac{x}{y} - 1 \right)^2 \leq 0, \tag{13}$$

in other words iff $x = y$, which is true iff $\sigma(x, y) = 0$.

Note that if x_i is close to y_i we have the approximation, from Taylor's theorem,

$$\sigma(x,y) = \frac{1}{2}\sum_{i=1}^n w_i \frac{(x_i-y_i)^2}{y_i^2} + o(\|x-y\|), \tag{14}$$

which can be compared with

$$\sum_{i=1}^n w_i (\log x_i - \log y_i)^2 = \sum_{i=1}^n w_i \frac{(x_i-y_i)^2}{y_i^2} + o(\|x-y\|). \tag{15}$$

From definition (2) of FLF as a function of the parameters for fixed data we obtain

$$\sigma(\theta) = \sum_{i=1}^n w_i \left\{ \frac{\xi_i(\theta)}{y_i} - \log \frac{x_i}{y_i} - 1 \right\}, \tag{16}$$

with partial derivatives

$$\mathcal{D}\sigma(\theta) = \sum_{i=1}^n w_i \left\{ \frac{1}{y_i} - \frac{1}{\xi_i(\theta)} \right\} \mathcal{D}\xi_i(\theta), \tag{17}$$

$$\mathcal{D}^2\sigma(\theta) = \sum_{i=1}^n w_i \left\{ \frac{1}{y_i} - \frac{1}{\xi_i(\theta)} \right\} \mathcal{D}^2\xi_i(\theta) + \sum_{i=1}^n w_i \frac{1}{\xi_i^2(\theta)} \mathcal{D}\xi_i(\theta)\mathcal{D}\xi_i(\theta)'. \tag{18}$$

A sufficient condition for convexity of σ on Θ is that the ξ_i are linear. Another sufficient condition is that the ξ_i are convex, and that $\xi_i(\theta) \geq y_i$.

2.2 Antithesis and Synthesis

Note that the FLF has a twin in which we swap x and y, i.e. the data and the model.

$$\sigma_L(x,y) := \sum_{i=1}^n w_i f(\frac{y_i}{x_i}) = \sum_{i=1}^n w_i \left\{ \frac{y_i}{x_i} - \log \frac{y_i}{x_i} - 1 \right\}. \tag{19}$$

We use σ_L, with "L" for left. The idea is that our FLF of definition (6) is σ_R, with "R" for right. There is also a symmetrized version σ_M, with "M" for middle.

$$\sigma_M(x,y) := \frac{1}{2}\{\sigma_R(x,y) + \sigma_L(x,y)\} = \frac{1}{2}\sum_{i=1}^n w_i \frac{(x_i-y_i)^2}{x_i y_i} \tag{20}$$

We will concentrate on the FLF of definition (6), because in most cases it leads to (slightly) simpler equations and algorithms. For the partials of σ_L we find, for instance,

$$\mathcal{D}\sigma_L(\theta) = \sum_{i=1}^n w_i \frac{1}{\xi_i(\theta)} \left\{ 1 - \frac{y_i}{\xi_i(\theta)} \right\} \mathcal{D}\xi_i(\theta), \tag{21}$$

$$\mathcal{D}^2\sigma_L(\theta) = \sum_{i=1}^n w_i \frac{1}{\xi_i(\theta)} \left\{ 1 - \frac{y_i}{\xi_i(\theta)} \right\} \mathcal{D}^2\xi_i(\theta) + \sum_{i=1}^n w_i \frac{1}{\xi_i^2(\theta)} \left\{ 2\frac{y_i}{\xi_i(\theta)} - 1 \right\} \mathcal{D}\xi_i(\theta)\mathcal{D}\xi_i(\theta)'. \tag{22}$$

A sufficient condition for σ_L to be convex on Θ is that the ξ_i are convex and $y_i < \xi_i(\theta) < 2y_i$. Another sufficient condition is that the ξ_i are linear and $\xi_i(\theta) < 2y_i$.

3 Applications in Regression

3.1 Multiplicative Regression

I start with the simplest application I can think of, in which $\xi(\alpha) = \alpha x_i$, where α is a one-dimensional parameter. Thus

$$\sigma(\alpha) = \sum_{i=1}^n w_i \left\{ \frac{\alpha x_i}{y_i} - \log \frac{\alpha x_i}{y_i} - 1 \right\}, \tag{23}$$

and

$$\mathcal{D}\sigma(\alpha) = \sum_{i=1}^n w_i \left\{ \frac{x_i}{y_i} - \frac{1}{\alpha} \right\}. \tag{24}$$

Consequently

$$\hat{\alpha} = \frac{\sum_{i=1}^n w_i}{\sum_{i=1}^n w_i \frac{x_i}{y_i}}, \tag{25}$$

which is the weighted harmonic mean of the y_i/x_i.

3.2 Additive Regression

In a slighty more complicated application we have a single additive parameter, instead of a multiplicative one.

$$\sigma(\beta) = \sum_{i=1}^n w_i \left\{ \frac{x_i - \beta}{y_i} - \log \frac{x_i - \beta}{y_i} - 1 \right\}, \tag{26}$$

$$\mathcal{D}f(\beta) = \sum_{i=1}^n \left\{ -\frac{1}{y_i} + \frac{1}{x_i - \beta} \right\}, \tag{27}$$

$$\mathcal{D}^2 f(\beta) = \sum_{i=1}^n \frac{1}{(x_i - \beta)^2}. \tag{28}$$

The function σ is defined only if $x_i - \beta > 0$ for all i and thus for $\beta < x_{\min}$. In that region σ is convex and we can apply Newton's method. But, as always with Newton, we have to be careful. We start, obviously, with $\beta < x_{\min}$, but the crux is to stay in the region. As σ is very steep near the vertical asymptote, the minimum is likely to be close to $x_{\min}$. See figure 2. If we start with choosing β too small then Newton will possibly take us to a $\beta > x_{\min}$, and we are in trouble. The algorithm resolves this by a combination of Newton and bisection. It makes sure that $\mathcal{D}\sigma(\beta)$ is always positive, and thus it remains between the asymptote and the minimum and converges from the right. If we start with $\beta = -3$ the algorithm converges after one or two iterations. The output shows the function values, the value of β, and the first and second partial derivative.

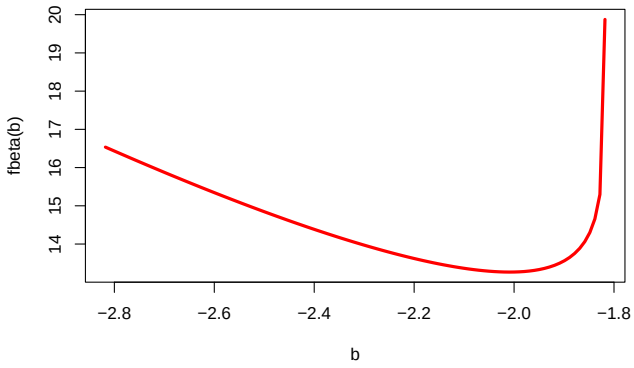


Figure 2: Additive Regression

```
## itel  1 fold 207.29428 fnew 207.29396 bnew -1.82742 gnew 0.06544 hnew 11167.08336
## itel  2 fold 207.29396 fnew 207.29396 bnew -1.82743 gnew 0.00004 hnew 11153.27193
## itel  3 fold 207.29396 fnew 207.29396 bnew -1.82743 gnew 0.00000 hnew 11153.26338
```

Note the second derivative is very large, which means the function is very flat near the minimum.

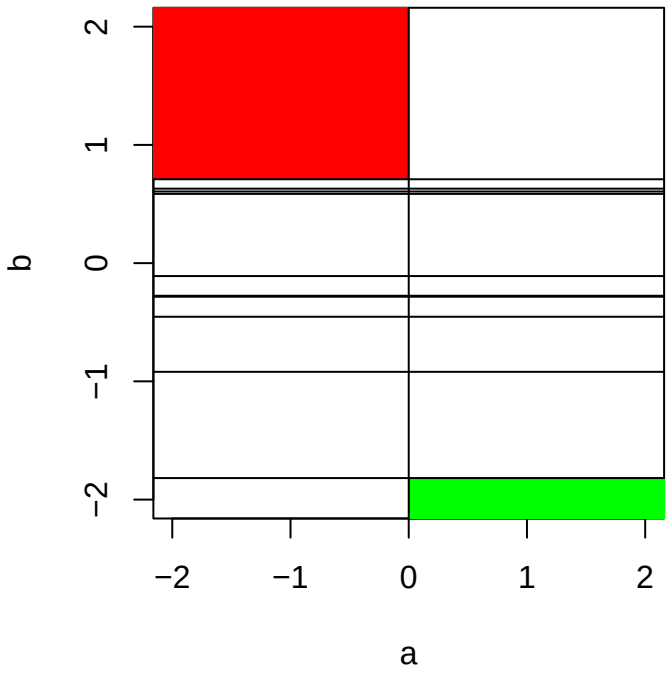
3.3 Multiple Linear Regression

$y_i > 0$ define $z_i = x_i/y_i$

$$\sigma(\beta) = \sum_{i=1}^{n} \left\{ \frac{\beta' x_i}{y_i} - \log \frac{\beta' x_i}{y_i} - 1 \right\}$$

Convex region $\beta' z_i > 0$. Newton.

$a(x_i - b) \geq 0$ for all i iff either $a > 0$ and $b < x_{\min}$ or $a < 0$ and $b > x_{\max}$.



$$\mathcal{D}\sigma(\beta) = \sum_{i=1}^{n} w_i \{1 - \frac{1}{\beta' z_i}\} z_i$$

$$\mathcal{D}^2\sigma(\beta) = \sum_{i=1}^{n} w_i \frac{1}{(\beta' z_i)^2} z_i z_i'$$

Cao, Eggermont, and Terebey (1999)

Itakura-Saito distance, negative cross Burg entropy, negative cross log entropy

$x_i \geq 0$ minimize over $\beta \geq 0$.

$$\sigma(\beta) = \sum_{i=1}^{n} w_i \left\{ \frac{y_i}{x_i'\beta} - \log \frac{y_i}{x_i'\beta} \right\}$$

without the weights. Majorize

$$\mathcal{D}\sigma(\beta) = \sum_{i=1}^{n} w_i \frac{1}{z_i'\beta} \left\{ 1 - \frac{1}{z_i'\beta} \right\} z_i$$

$$\mathcal{D}\sigma(\beta) = \sum_{i=1}^{n} w_i \frac{1}{(z_i'\beta)^2} \left\{ 2\frac{1}{z_i'\beta} - 1 \right\} z_i z_i'$$

Convex for $z_i'\beta \leq 2$

3.4 Monotone Regression

In Monotone Regression (for a weak linear order) we minimize

$$\sum_{i=1}^{n} w_i \{\frac{x_i}{y_i} - \log \frac{x_i}{y_i} - 1\} \tag{29}$$

over all $x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n$.

Lemma 3.1 (Monotone Regression Lemma)**.** *Suppose $n = 2$ and we require $x_1 \leq x_2$. If $y_1 \leq y_2$ then $\hat{x}_1 = y_1$ and $\hat{x}_2 = y_2$. If $y_i \geq y_2$ then $\hat{x}_1 = \hat{x}_2 = \mathcal{H}(y_1, y_2; w_1, w_2)$*

Proof. We must minimize the convex function σ on the cone $0 < x_1 \leq x_2$. There are only two possibilities: either (y_1, y_2) is inside the cone, or it is outside. If it is inside the cone the gradient must vanish, which means $(\hat{x}_1, \hat{x}_2) = (y_1, y_2)$. If (y_1, y_2) is outside the cone, i.e. $0 < y_2 < y_1$, we project on the boundary line $x_1 = x_2$, which produces the weighted harmonic mean of y_1 and y_2. □

Theorem 3.1 (PAVA Theorem)**.** *If $y_i > y_{i+1}$ then $\hat{x}_i = \hat{x}_{i+1}$.*

Proof. If $\hat{x}_i < \hat{x}_{i+1}$ then the Monotone Regression Lemma shows that merging improves the fit. And of course the merged values are still feasible. □

From theorem 3.1 it follows we can compute the monotonic regression by a Pooled Adjacent Violaters Algorithm, or PAVA (De Leeuw, Hornik, and Mair (2009)). It also follows that the reasoning in De Leeuw (1977b) applies for the case in which the y_i have ties.

The PAVA algorithm looks for a violation $y_i > y_{i+1}$. It then uses this violation, and theorem @3.1, to reduce the problem from one of size n to one of size $n-1$. w_i and w_{i+1} are replaced by $w_i + w_{i+1}$ and y_i and y_{i+1} by $\mathcal{H}(y_i, y_{i+1}; w_i, w_{i+1})$. We then continue in the same way with the smaller problem until we are ultimately left with a vector of merged elements that are in the correct order.

We give a simple example where we proceed strictly from left to right. All weights are one, which means that after merging the weights are the sizes of the blocks. The matrices below have the merged y_1 in the first row and the block sizes in the second row. If the weights are arbitrary positive numbers we maintain three rws: block values, block weights, and block sizes. Efficient ways of merging and selecting which violations to use are discussed by Busing (2022).

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]    2    1    3    1    1    5    3
## [2,]    1    1    1    1    1    1    1
```

```
##             [,1] [,2] [,3] [,4] [,5] [,6]
## [1,] 1.333333333    3    1    1    5    3
## [2,] 2.000000000    1    1    1    1    1
```

```
##              [,1] [,2] [,3] [,4] [,5]
## [1,] 1.333333333  1.5    1    5    3
## [2,] 2.000000000  2.0    1    1    1
```

```
##              [,1]        [,2] [,3] [,4]
## [1,] 1.333333333 1.285714286    5    3
## [2,] 2.000000000 3.000000000    1    1
```

```
##              [,1] [,2] [,3]
## [1,] 1.304347826    5    3
## [2,] 5.000000000    1    1
```

```
##              [,1] [,2]
## [1,] 1.304347826 3.75
## [2,] 5.000000000 2.00
```

Now the two remaining elements are in the correct order and we expand the solution to its original length, using the block sizes.

```
## [1] 1.304347826 1.304347826 1.304347826 1.304347826 1.304347826 3.750000000
## [7] 3.750000000
```

Just to make sure we did no go astray, we check the Karush-Kuhn-Tucker conditions for minimizing $\sigma(x, y)$ over x satisfying $Ax \leq 0$, where A is

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]    1   -1    0    0    0    0    0
## [2,]    0    1   -1    0    0    0    0
## [3,]    0    0    1   -1    0    0    0
## [4,]    0    0    0    1   -1    0    0
## [5,]    0    0    0    0    1   -1    0
## [6,]    0    0    0    0    0    1   -1
```

The gradient $\nabla\sigma(\hat{x})$ is

```
## [1] -0.26666666667  0.23333333333 -0.43333333333  0.23333333333  0.23333333333
## [6] -0.06666666667  0.06666666667
```

and the Lagrange multipliers $\hat{\lambda}$ are the solution of $A'\lambda = \nabla\sigma(\hat{x})$, which gives

```
## [1] 0.26666666667 0.03333333333 0.46666666667 0.23333333333 0.00000000000
## [6] 0.06666666667
```

The right-hand sides $\hat{r} = A\hat{x}$ are

```
## [1]  0.000000000  0.000000000  0.000000000  0.000000000 -2.445652174
## [6]  0.000000000
```

The KKT conditions $\hat{\lambda} \geq 0$ and $\hat{r} \leq 0$ are satisfied, and so is strict complementarity which requires that for each i either $\hat{r}_i = 0$ or $\hat{\lambda}_i = 0$ (but not both). Thus we have indeed computed the unique minimum of σ over the cone $x_1 \leq \cdots \leq x_n$.

4 Applications in Log-linear Analysis

$$\sigma(\theta, p) = \sum_{i=1}^{n} w_i \left\{ \frac{\pi_i(\theta)}{p_i} - \log \frac{\pi_i(\theta)}{p_i} - 1 \right\}$$

$$\mathcal{D}\sigma(\theta) = \sum_{i=1}^{n} w_i \left\{ \frac{1}{p_i} - \frac{1}{\pi_i(\theta)} \right\} \mathcal{D}\pi_i(\theta) = G(\theta)'h(\theta)$$

$$\mathcal{D}^2\sigma(\theta) = \sum_{i=1}^{n} w_i \left\{ \frac{1}{p_i} - \frac{1}{\pi_i(\theta)} \right\} \mathcal{D}^2\pi_i(\theta) + \sum_{i=1}^{n} w_i \frac{1}{\pi_i^2(\theta)} \mathcal{D}\pi_i(\theta)\mathcal{D}\pi_i(\theta)$$

4.1 Choice of Weights

FLF gives us some freedom, because we can choose weights that depend on the data. That does not change the basic propertie sof the loss function, and it does not change first and second derivative formulas.

Suppose the model is such that $\sum_{i=1}^{n} \pi_i(\theta) = 1$ for all θ, or, alternatively, we require $\sum_{i=1}^{n} \pi_i(\theta) = 1$ as a side condition. If we choose $w = p$ then

$$\sigma(\theta) = \sum_{i=1}^{n} p_i \log \frac{p_i}{\pi_i(\theta)},$$

which is the Kullback-Leibler distance between p and $\pi(\theta)$. It follows that minimizing FLF with $w = p$ is equivalent to computing maximum likelihood estimates of θ.

There is some freedom in choosing weights. We could also consider, for example, $w = -\log p$, which means a zero weight if $p = 1$ and an infinitely large weight if $p = 0$.

4.2 Estimates

Fisher-consistent $\hat{\theta}(p)$ converges in probability to $\hat{\theta}(\pi(\theta)) = \theta$

$$\mathcal{D}\theta(p) = -\{\mathcal{D}_{22}\sigma(p, \theta(p))\}^{-1}\mathcal{D}_{21}\sigma(p, \theta(p))$$

$$\{\mathcal{D}_{21}\sigma(p,\theta(p))\}_{si} = -w_i/p_i^2$$

$$\mathcal{D}_{22}\sigma(p,\theta(p)) = \sum_{i=1}^{n} w_i \left\{\frac{1}{p_i} - \frac{1}{\pi_i(\theta)}\right\} \mathcal{D}_{ik}\pi(\theta) + \sum_{i=1}^{n} w_i \frac{1}{\pi_i^2(\theta)} \mathcal{D}\pi_i(\theta)\mathcal{D}\pi_i(\theta)'$$

4.3 Independence

$$\sigma(\alpha,\beta) = \sum_{i=1}^{n}\sum_{j=1}^{m} w_i \left\{\frac{\alpha_i\beta_j}{p_{ij}} - \log\frac{\alpha_i\beta_j}{p_{ij}} - 1\right\}$$

$$\alpha_i = \mathcal{H}(\frac{p_{i1}}{\beta_1}, \cdots, \frac{p_{im}}{\beta_m})$$
$$\beta_j = \mathcal{H}(\frac{p_{1j}}{\alpha_1}, \cdots, \frac{p_{nj}}{\alpha_n})$$

4.4 No Second-Order Interaction

$$\mathcal{L} = \sum_i \sum_j \sum_k \left\{\frac{\alpha_{ij}\beta_{ik}\gamma_{jk}}{p_{ijk}} - \log\frac{\alpha_{ij}\beta_{ik}\gamma_{jk}}{p_{ijk}} - 1\right\}$$
$$\frac{\partial\mathcal{L}}{\partial\alpha_{ij}} = \sum_k \frac{\beta_{ik}\gamma_{jk}}{p_{ijk}} - \frac{1}{\alpha_{ij}}$$

instead of $\mathcal{L}(\pi, p)$ also $\mathcal{L}(p, \pi)$

$$\mathcal{L} = \sum_i \sum_j \sum_k \{p_{ijk}\alpha_{ij}\beta_{ik}\gamma_{jk} - \log p_{ijk}\alpha_{ij}\beta_{ik}\gamma_{jk} - 1\}$$

$$\sum_j \sum_k p_{ijk}\beta_{ik}\gamma_{jk} - \frac{1}{\alpha_{ij}} = 0$$

Missing Data

5 Applications in Multivariate Analysis

5.1 Factor Analysis

Let me first report a happy coincidence. In a really excellent, and unfortunately somewhat neglected paper, Swain (1975) discussed several loss functions for (random orthogonal exploratory) factor analysis that give tests and estimates with the same statistical properties as the maximum likelihood method. All these loss functions are functions of the eigenvalues θ_i of $S^{-1}\Sigma$, where S is the sample covarinace matrix and $\Sigma = AA' + \Delta^2$ is the factor analysis

mode. In particular, Swain shows that the maximum likelihood estimation corresoinds with minimizing

$$\sum_{i=1}^{n}\left\{\frac{1}{\theta_i}+\log(\theta_i)-1\right\},$$

which is obviously the same as minimizing

$$\sigma(\lambda)=\sum_{i=1}^{n}\left\{\frac{\lambda_i}{1}-\log(\frac{\lambda_i}{1})-1\right\},$$

where the λ_i are the eigenvalues of $\Sigma^{-1}S$. Thus maximum likelihood estimation in factor analysis is a special case of minimizing our new fun loss function σ.

5.2 Maximum Likelihood Ratio

The original motivation for the new loss function, however, comes from De Leeuw (n.d.), and is more closely related to Mc Donald's maximum likelihood ratio estimate in (fixed orthogonal exploratory) factor analysis (McDonald (1979)).

Suppose $\underline{z}_i$ are independent normallly distributed random variables, with

$$\underline{z}_i\sim\mathcal{N}(y_i-f_i(\theta),\ \delta_i^2(\eta)).$$

Both means and variances depend on parameters, respectively θ and η. The the deviance (two times the negative log-likelihood) is

$$\Delta(\theta,\eta)=\sum_{i=1}^{n}\left\{\log\delta_i^2(\eta)+\frac{(y_i-f_i(\theta))^2}{\delta_i^2(\eta)}\right\}$$

Presumably we want to minimize this over θ and η, but this runs into problems if there is at least one index k, a $\hat{\theta}$, and a $\hat{\eta}$ for which $y_k-f_k(\theta)=0$ and $\lim_{\eta\to\hat{\eta}}\delta_k^2(\eta)=0$. If that is the case then $\inf_\theta\inf_\eta\Delta(\theta,\eta)=-\infty$ and the minimum (and thus the maximum lilkeihood estimate) does not exist (Anderson and Rubin (1956)).

McDonald (1979) proposes to subtract the minimum deviance over the unconstrained δ_i for given residuals $y_i-f_(\theta)$ from the deviance in equation ... Now

$$\min_\delta\Delta(\theta,\delta)=\min_\delta\sum_{i=1}^{n}\left\{\log\delta_i^2+\frac{(y_i-f_i(\theta))^2}{\delta_i^2}\right\}=\sum_{i=1}^{n}\left\{\log(y_i-f_i(\theta))^2+1\right\},$$

and thus

$$\Delta(\theta,\eta)-\min_\delta\Delta(\theta,\delta)=\sum_{i=1}^{n}\left\{\frac{(y_i-f_i(\theta))^2}{\delta_i^2(\eta)}-\log\frac{(y_i-f_i(\theta))^2}{\delta_i^2(\eta)}-1\right\}=\sigma((y_i-f_i(\theta))^2,\ \delta_i^2(\eta))$$

Subtracting the term $\min_\delta\Delta(\theta,\delta)$ works as a barrier penalty function, preventing $y_i-f_i(\theta)=0$. Now I am not sure about the statistical optimality properties of maximum likelihood ratio estimates, but in this case they lead us to our fun loss function and that's all I care about here. This approach is analyzed, with examples, in more detail in De Leeuw (n.d.). Clearly it can be used to fit not just fixed score factor analysis, but a wide variety of normal mean structure/variance structure models.

6 MDS/Unfolding

Another reason for looking at loss function σ is that its zero-avoidance properties may be helpful in multidimensional scaling (MDS), in particular in unfolding, and even more in particular in non-metric unfolding. We have developed an appropriate form of monotone regression in section 3.4, and consequently alternating minimization can be used to turn any metric MDS into a non-metric MDS.

6.1 Theory

In order to get relatively simple formulas for the MDS loss function and its derivatives we use some convenient notation. This will also bring out the similarity with the corresponding formulas for smacof (De Leeuw and Mair (2009), Mair, Groenen, and De Leeuw (2022)).

First, define

$$\mathfrak{A}_{ij} := (e_i - e_j)(e_i - e_j)', \tag{30}$$

where e_i and e_j are n-element unit vectors, which respectively have elements $i$and j equal to one, and all other elements equal to zero. The matrix $\mathfrak{A}_{ij}$, with $i \neq j$, is symmetric and of order n, with elements (i, i) and (j, j) equal to $+1$ and elements (i, j) and (j, i) equal to -1. Thus $\mathfrak{A}_{ij}$ is doubly-centered (rows and columns add up to zero) and of rank one, with tr $A_{ij} = 2$.

Next, using the Kronecker product,

$$A_{ij} := I_p \otimes \mathfrak{A}_{ij}, \tag{31}$$

which can alternatively be written as a direct sum

$$A_{ij} := \underbrace{\mathfrak{A}_{ij} \oplus \cdots \oplus \mathfrak{A}_{ij}}_{p \text{ times}}. \tag{32}$$

In equation (31) matrix I_p is the identity matrix of order p, and thus A_{ij} is block-diagonal of order $n \times p$, with p copies of $\mathfrak{A}_{ij}$ along the diagonal. Equation (32) says the same thing.

The main reason for defining A_{ij} is that we can now write the squared Euclidean distance $d_{ij}^2(X)$ between rows i and j of the $n \times p$ *configuration matrix* X as

$$d_{ij}^2(X) = x' A_{ij} x, \tag{33}$$

where $x = \text{vec}(X)$, i.e. x is a vector with the p columns of X stacked on top of each other.

Our loss function becomes

$$\sigma(x) = \frac{1}{2} \sum_{1 \leq i < j \leq n} \sum w_{ij} \left\{ \frac{x' A_{ij} x}{\delta_{ij}^2} - \log \frac{x' A_{ij} x}{\delta_{ij}^2} - 1 \right\}. \tag{34}$$

Note that we are aiming at $d_{ij}^2(x) \approx \delta_{ij}^2$, and not at $d_{ij}(x) \approx \delta_{ij}$, which would mean a different weighting of errors. We assume that all $\delta_{ij} > 0$ and that the w_{ij} are irreducible, i.e. the MDS problem does not separate into a number of smaller MDS problems (De Leeuw (1977a)).

The gradient is

$$\mathcal{D}\sigma(x) = (S - T(x))x, \tag{35}$$

with matrix

$$S := \sum_{1\leq i<j\leq n}\sum w_{ij}\frac{1}{\delta_{ij}^2}A_{ij}, \tag{36}$$

and matrix-valued function

$$T(x) := \sum_{1\leq i<j\leq n}\sum w_{ij}\frac{1}{d_{ij}^2(x)}A_{ij}. \tag{37}$$

Both S and each $T(x)$ are block-diagonal, doubly-centered, and positive semi-definite. Because of irreducibility they are both of rank $n-1$, with only the constant vectors in their null-spaces (De Leeuw (1977a)). We say that x is *stationary* if $\mathcal{D}\sigma(x) = 0$, i.e. if $x = S^+T(x)x$.

For the Hessian we find

$$\mathcal{D}^2\sigma(x) = S - (T(x) - U(x)) + U(x), \tag{38}$$

with

$$U(x) = \sum_{1\leq i<j\leq n}\sum w_{ij}\frac{1}{d_{ij}^4(x)}A_{ij}xx'A_{ij}. \tag{39}$$

Unlike S and T the matrix-valued function U does not not take block-diagonal values but produces full matrices with non-zero off-diagonal blocks. Like S and T matrix $U(x)$ is positive semi-definite and doubly-centered. Also

$$T(x) - U(x) = \sum_{1\leq i<j\leq n}\sum w_{ij}\frac{1}{d_{ij}^2(x)}\left(A_{ij} - \frac{A_{ij}xx'A_{ij}}{x'A_{ij}x}\right), \tag{40}$$

which shows $T(x) \gtrsim U(x)$ for all x. This implies

$$S - T(x) \lesssim S - T(x) + U(x) \lesssim \mathcal{D}^2\sigma(x) \lesssim S + U(x) \lesssim S + T(x), \tag{41}$$

and thus $\mathcal{D}^2\sigma(x) \gtrsim 0$ if $S^+T(x) \lesssim I_{pn}$. Note that at a stationary point x is an eigenvector of $S^+T(x)$ with eigenvalue equal to one.

Because of equations (14) and (15) we expect that our solutions may be similar to the ones computed by MULTISCAL (Ramsay (1977)). But note that there are several interesting variations to MDS-FLF. We could use weights $w_{ij} = \delta_{ij}^2$, for instance. We could switch from fitting squared distances to fitting distances, or to other powers of distances. The logarithm in FLF makes fitting powers of distances especially simple. And we could use the FLF twin σ_L of definition (19), or the symmetrized version σ_M of definition (20). All this remains to be explored (or not).

6.2 Algorithm

Although we have formula (38) for the Hessian, applying Newton iterations directly is fraught with perils. Various safeguards are surely necessary. Instead we have opted for the iteration

$$x^{(k+1)} = x^{(k)} - (S + T(x^{(k)}))^{-1}(S - T(x^{(k)}))x^{(k)} = 2(S + T(x^{(k)}))^{-1}T(x^{(k)})x^{(k)} \tag{42}$$

The default start sets $x^{(0)}$ equal to the classical scaling solution (Torgerson (1958)).

Because equation (41) tells us that $\mathcal{D}^2\sigma(x) \lesssim S + T(x)$ process (42) can be thought of as a dampened version of Newton's method. In the few examples we have tried the loss function values $\sigma(x^k)$ decrease monotonically, even after thousands of iterations (no proof yet !). We have given up on quadratic convergence and obtained stability instead.

6.3 Examples

6.3.1 De Gruijter

Our first metric MDS example uses the data from De Gruijter (1967), giving dissimilarities between the nine Dutch political parties that were in parliament at the time. The data are averaged over a politically heterogeneous group of students, and consequently all dissimilarities regress to the mean. We applied a monotone transformation to stretch the scale and then rounded to the nearest integer.

The data are

```
##      KVP PvdA VVD ARP CHU CPN PSP BP D66
## KVP    0    2   1   1   1   5   3  4   2
## PvdA   2    0   3   2   2   1   1  4   2
## VVD    1    3   0   2   1   9   5  3   1
## ARP    1    2   2   0   1   7   3  4   2
## CHU    1    2   1   1   0   6   4  3   2
## CPN    5    1   9   7   6   0   1  2   5
## PSP    3    1   5   3   4   1   0  3   2
## BP     4    4   3   4   3   2   3  0   4
## D66    2    2   1   2   2   5   2  4   0
```

Metric MDS to minimize the FLF uses 391 iterations to arrive at loss 4.7928377744. We iterate until loss decreases less than 1e-15, a precision obviously not warranted by the data. The gradient at convergence is

```
##       [,1]          [,2]
##  [1,] -0.0000000229 +0.0000000637
##  [2,] +0.0000000062 -0.0000000088
##  [3,] +0.0000000373 +0.0000000142
```

```
##  [4,] -0.0000000697 -0.0000000470
##  [5,] +0.0000000637 -0.0000000584
##  [6,] -0.0000000038 +0.0000000233
##  [7,] -0.0000000218 +0.0000000025
##  [8,] +0.0000000117 +0.0000000136
##  [9,] -0.0000000007 -0.0000000030
```

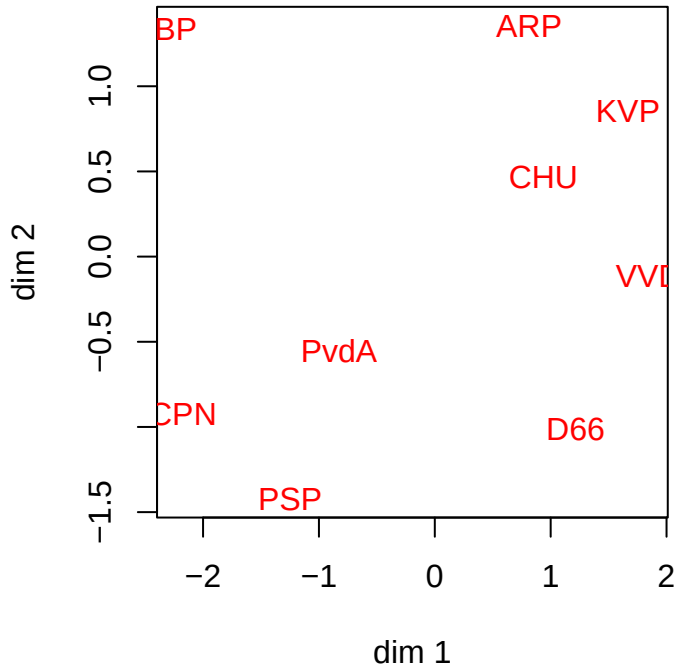


Figure 3: FLF De Gruijter Solution 1

The solution, in figure 3, shows the cluster of leftist parties (CPN, PSP, PvdA), the cluster of religious parties (KVP, CHU, ARP), the pragmatic centrists (D66, brand new in 1967), the traditional European conservative liberals (VVD), and the unavoidable right-wing protest party (BP).

To look at the local minimum situation we repeated the analysis ten times, each tme from a different random start. The number of iterations and the final FLF values are

```
##    1503          2.0039193801
##    1591          1.7563295930
##    1207          1.4093409943
##    1113          1.4093409943
##    1030          1.5881708149
##    2834          1.7627681641
##    1119          2.0039193801
##    1158          1.4093409943
##    1322          1.4093409943
##    1324          1.5881708149
```

The ten runs give five local minima, together with the solution from the original run there are at least six. The solution with the smallest FLF of 1.4093409943 is in figure 4.

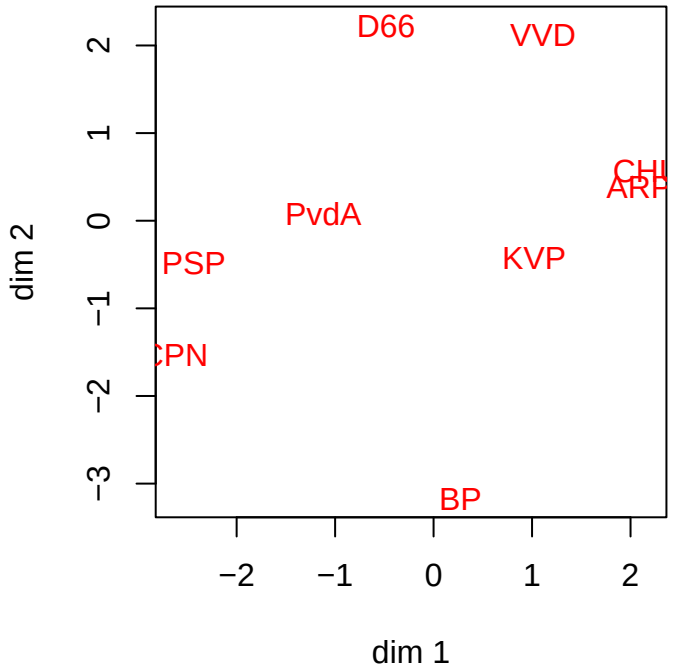


Figure 4: FLF De Gruijter Solution 2

We see the same clusters of parties, and basically the same overall arrangement, with some small differences.

The Shepard Plot (De Leeuw and Mair (2015)) in figure 5 shows an atrocious fit. We must remember, however, that FLF does not aim at $x_i - y_i = 0$ but at $x_i/y_i = 1$, i.e. at $\log x_i - \log y_i = 0$. The log Shepard Plot in figure 6 gives a better picture of the fit.

Another way of showing the fLF fit it to plot the empirical cumulative distribution of the residuals

$$r_{ij}(X) = \frac{d_{ij}^2(X)}{\delta_{ij}^2} - 1. \tag{43}$$

Figure 7 show they are tightly clustered around zero. The outlier is the distance between VVD (conservatives) and CPN (communists).

6.3.2 Ekman

The second example uses the color data from Ekman (1954), a matrix of similarities between 14 colors. We know from many previous MDS analysis of these data that MDS in two dimensions finds a global minimum with very good fit that puts the 14 colors on the color circle, with a gap between the two endpoints (see De Leeuw (2019)). Given previous results we apply the transformation $(1-s_{ij})^3$ top the similartities, and to make it easy for non-metric scxaling we then round to the nearest integer. The data are

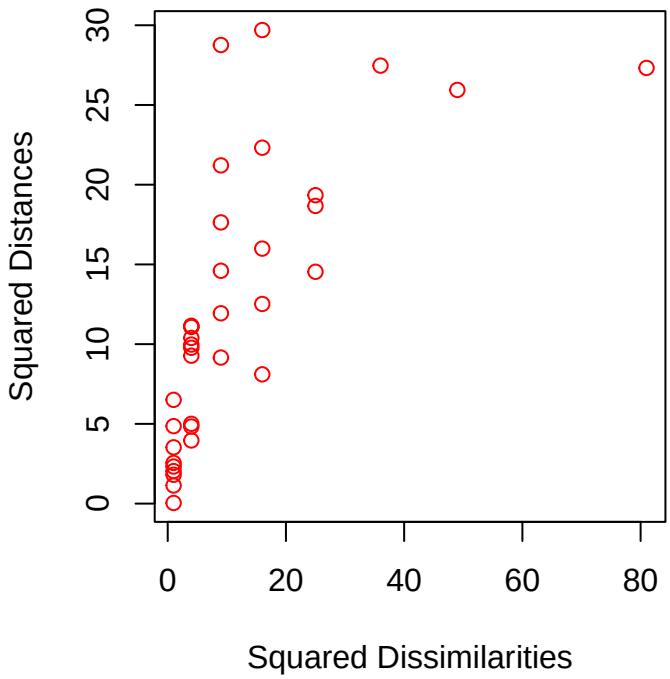


Figure 5: FLF De Gruijter Shepard Plot

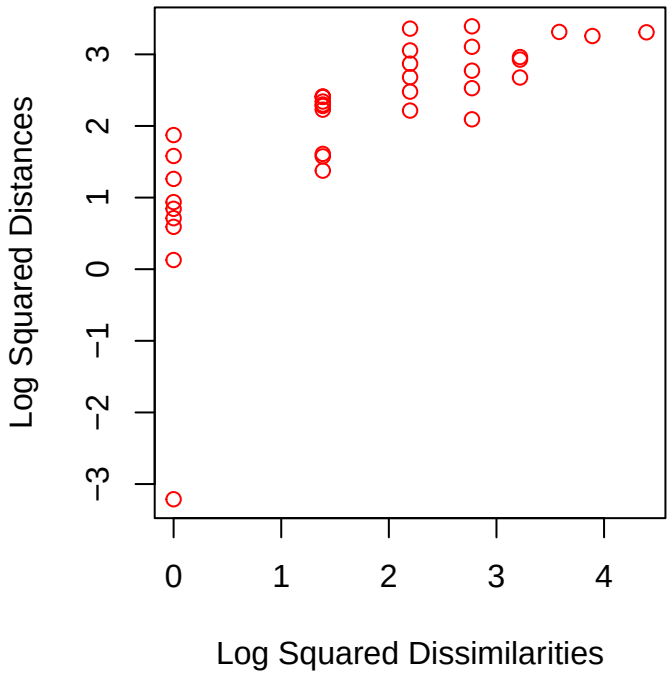


Figure 6: FLF De Gruijter Log Shepard Plot

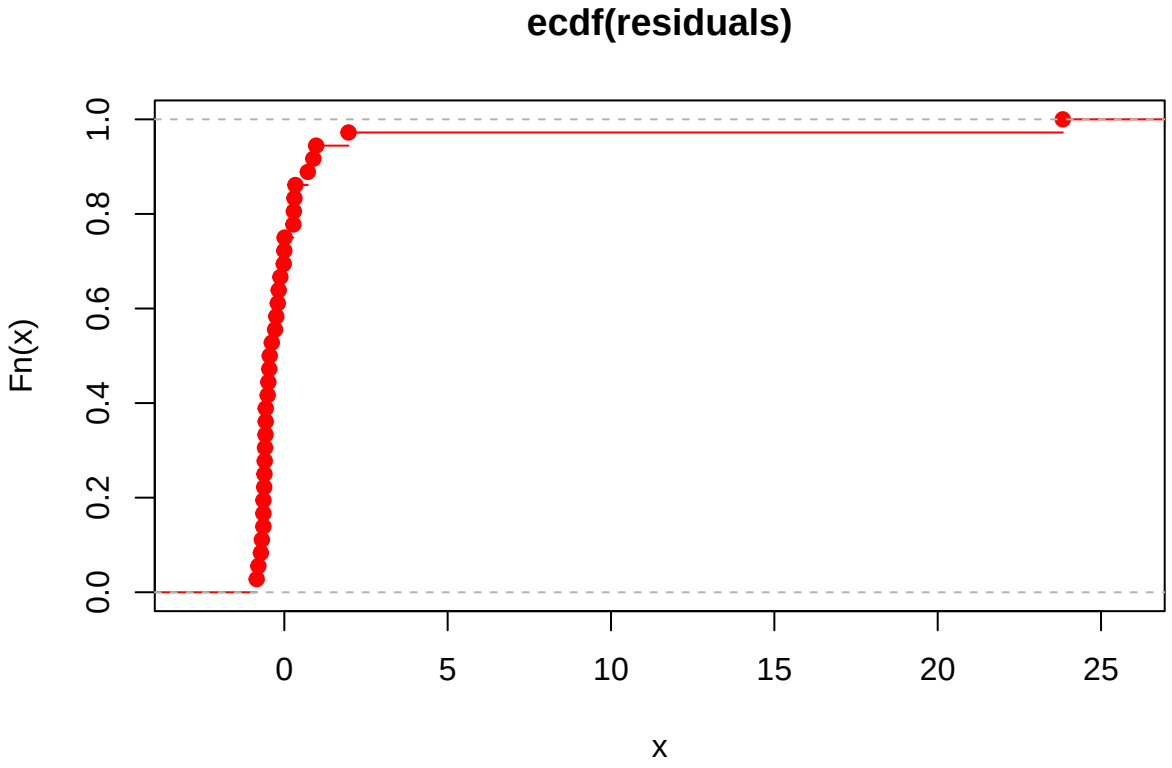


Figure 7: FLF Residuals De Gruijter ECDF

```
##      434 445 465 472 490 504 537 555 584 600 610 628 651 674
## 434    0   1   3   3   7   9   9  10  10   9   9   8   8   7
## 445    1   0   2   3   6   9   9   9  10  10   9   8   8   7
## 465    3   2   0   1   2   7   8   9  10  11  10  11  10  10
## 472    3   3   1   0   2   5   8   9  10  11  11  11  10  10
## 490    7   6   2   2   0   2   4   5   9  10  10  11  10  11
## 504    9   9   7   5   2   0   2   3   7   9  10  10  10  11
## 537    9   9   8   8   4   2   0   1   6   7  10  10  10  11
## 555   10   9   9   9   5   3   1   0   4   6  10  10  10  10
## 584   10  10  10  10   9   7   6   4   0   2   4   5   6   6
## 600    9  10  11  11  10   9   7   6   2   0   1   2   3   5
## 610    9   9  10  11  10  10  10  10   4   1   0   1   2   2
## 628    8   8  11  11  11  10  10  10   5   2   1   0   1   1
## 651    8   8  10  10  10  10  10  10   6   3   2   1   0   1
## 674    7   7  10  10  11  11  11  10   6   5   2   1   1   0
```

After 890 iterations MDS-FLF arrives at a loss of 2.7831466825 (once again we iterate to precision 1e-15). The gradient at the solution is

```
##        [,1]          [,2]
##  [1,] +0.0000000768 +0.0000001044
##  [2,] -0.0000000873 -0.0000001084
```

```
##  [3,] -0.0000000059 -0.0000000168
##  [4,] +0.0000000105 +0.0000000226
##  [5,] +0.0000000026 +0.0000000009
##  [6,] +0.0000000010 +0.0000000007
##  [7,] +0.0000000008 +0.0000000010
##  [8,] +0.0000000013 -0.0000000007
##  [9,] +0.0000000011 -0.0000000002
## [10,] +0.0000000017 -0.0000000001
## [11,] +0.0000000010 -0.0000000010
## [12,] +0.0000000005 -0.0000000013
## [13,] -0.0000000018 +0.0000000014
## [14,] -0.0000000024 -0.0000000025
```

The Hessian is positive semi-definite, with three zero eigenvalues (one to take care of rotational indeterminacy, and two for translational indeterminacy, see De Leeuw (1988)). The solution is

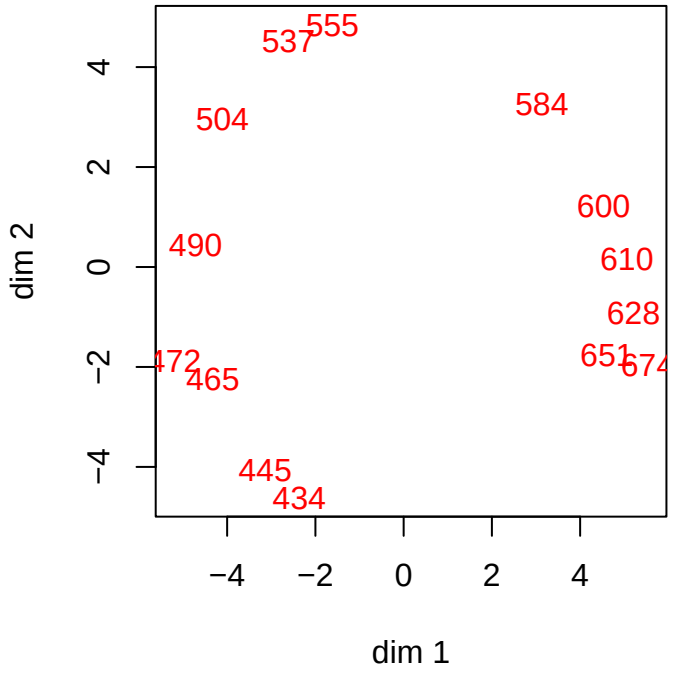


Figure 8: FLF Ekman Configuration

6.3.3 Roskam

References

Anderson, T. W., and H. Rubin. 1956. "Statistical Inference in Factor Analysis." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, edited by J. Neyman, V:111–50. Berkeley; Los Angeles: University of California Press.

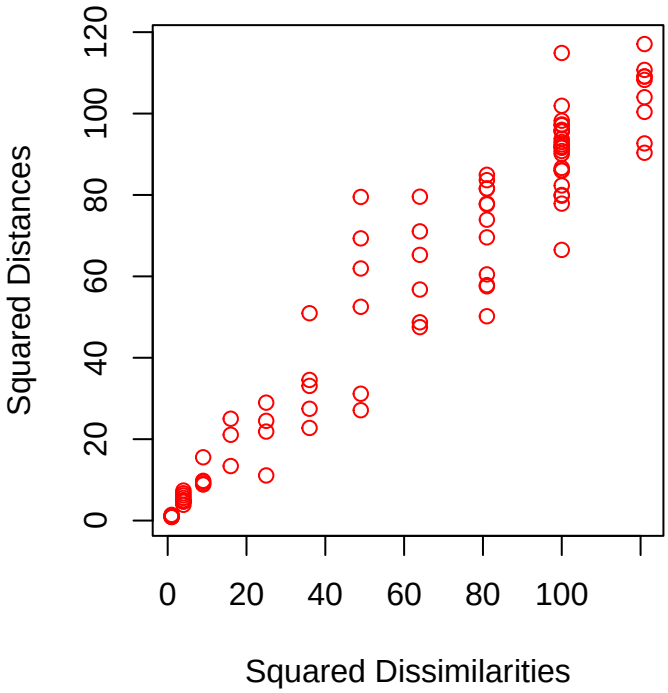


Figure 9: FLF Ekman Shepard Plot

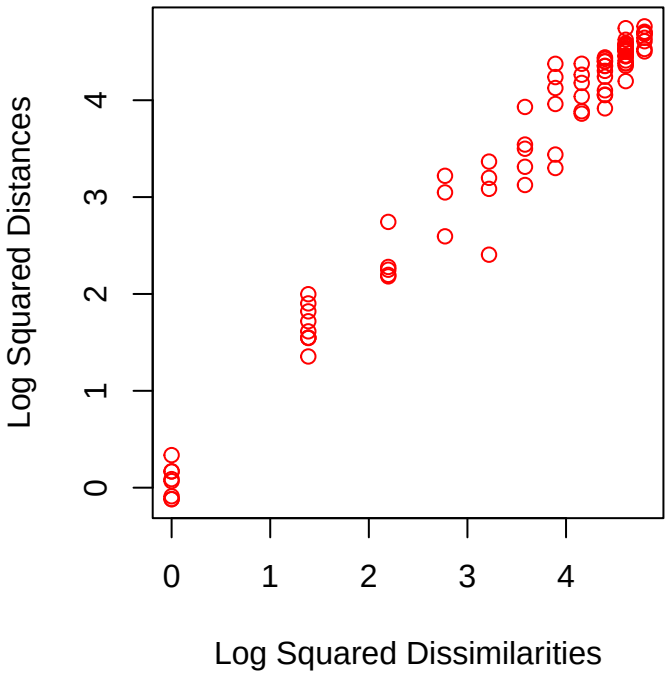


Figure 10: FLF Ekman Log Shepard Plot

Busing, F. M. T. A. 2022. "Monotone Regression: A Simple and Fast O(n) PAVA Implementation." *Journal of Statistical Software* 102 (Code Snippet 1).

Cao, Y., P. P. B. Eggermont, and S. Terebey. 1999. "Cross Log Entropy Maximization and Its Application to Ringing Suppression in Image Reconstruction." *IEEE Transactions on Image Processing* 8 (2): 1–5.

De Gruijter, D. N. M. 1967. "The Cognitive Structure of Dutch Political Parties in 1966." Report E019-67. Psychological Institute, University of Leiden.

De Leeuw, J. 1977a. "Applications of Convex Analysis to Multidimensional Scaling." In *Recent Developments in Statistics*, edited by J. R. Barra, F. Brodeau, G. Romier, and B. Van Cutsem, 133–45. Amsterdam, The Netherlands: North Holland Publishing Company.

———. 1977b. "Correctness of Kruskal's Algorithms for Monotone Regression with Ties." *Psychometrika* 42: 141–44.

———. 1988. "Convergence of the Majorization Method for Multidimensional Scaling." *Journal of Classification* 5: 163–80.

———. 2019. "Global Minima by Penalized Full-dimensional Scaling." 2019.

———. n.d. "Factor Analysis, Correspondence Analysis, ANOVA." https://jansweb.netlify.app/publication/deleeuw-e-22-c.

De Leeuw, J., K. Hornik, and P. Mair. 2009. "Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods." *Journal of Statistical Software* 32 (5): 1–24.

De Leeuw, J., and P. Mair. 2009. "Multidimensional Scaling Using Majorization: SMACOF in R." *Journal of Statistical Software* 31 (3): 1–30.

———. 2015. "Shepard Diagram." In *Wiley StatsRef: Statistics Reference Online*, 1–3. Wiley.

Ekman, G. 1954. "Dimensions of Color Vision." *Journal of Psychology* 38: 467–74.

Mair, P., P. J. F. Groenen, and J. De Leeuw. 2022. "More on Multidimensional Scaling in R: smacof Version 2." *Journal of Statistical Software* 102 (10): 1–47.

McDonald, R. P. 1979. "The Simultaneous Estimation of Factor Loadings and Scores." *British Journal of Mathematical and Statistical Psychology* 32 (212–228).

Ramsay, J. O. 1977. "Maximum Likelihood Estimation in Multidimensional Scaling." *Psychometrika* 42: 241–66.

Swain, A. J. 1975. "A Class of Factor Analysis Estimation Procedures with Common Asymptotic Sampling Properties." *Psychometrika* 40 (3): 315–36.

Torgerson, W. S. 1958. *Theory and Methods of Scaling*. New York: Wiley.