# ICA for subspaces

P. Spurek J. Tabor

## 1 ToDo

- modele ktore robia PCA?

- modele ktore nie robia PCA? czy wyszukiwanie podprzestrzeni jest istotnym skladnikiem modeu?

WIKIPEDIA: *Defining component independence.*

*ICA finds the independent components (also called factors, latent variables or sources) by maximizing the statistical independence of the estimated components. We may choose one of many ways to define a proxy for independence, and this choice governs the form of the ICA algorithm. The two broadest definitions of independence for ICA are*

- *Minimization of mutual information*
- *Maximization of non-Gaussianity*

*The Minimization-of-Mutual information (MMI) family of ICA algorithms uses measures like Kullback-Leibler Divergence and maximum entropy. The non-Gaussianity family of ICA algorithms, motivated by the central limit theorem, uses kurtosis and negentropy.*

*Typical algorithms for ICA use centering (subtract the mean to create a zero mean signal), whitening (usually with the eigenvalue decomposition), and dimensionality reduction as preprocessing steps in order to simplify and reduce the complexity of the problem for the actual iterative algorithm. Whitening and dimension reduction can be achieved with principal component analysis or singular value decomposition. Whitening ensures that all dimensions are treated equally a priori before the algorithm is run. Well-known algorithms for ICA include infomax, FastICA, JADE, and kernel-independent component analysis, among others. In general, ICA cannot identify the actual number of source signals, a uniquely correct ordering of the source signals, nor the proper scaling (including sign) of the source signals.*

*ICA is important to blind signal separation and has many practical applications. It is closely related to (or even a special case of) the search for a factorial code of the data, i.e., a new vector-valued representation of each data vector such that it gets uniquely encoded by the resulting code vector (loss-free coding), but the code components are statistically independent.*

## 2 Basic tools

### 2.1 Orthogonal projection onto affine subspaces

Suppose that we have an affine subspace generated over $m \in \mathbb{R}^D, \mathrm{V} \in \mathbb{R}^{D \times d}$, where $\mathrm{V} = [v_1, \ldots, v_k]$ (or more precisely its consecutive columns) is the base of linear part of $P$ with $d$ elements, that is

$$M = m + \mathrm{span}(\mathrm{V}) = m + \{\mathrm{V}r : r \in \mathbb{R}^d\} = \{m + r_1 v_1 + \ldots + r_d v_d : r_i \in \mathbb{R}\}.$$

We are interested in the coordinates of the point $x \in \mathbb{R}^D$ after the orthogonal projection onto $P$ with respect to the base. This can restated as the search for coordinates $r = (r_1, \ldots, r_d)^T \in \mathbb{R}^d$ such that

$$r = \operatorname*{argmin}_{s \in \mathbb{R}^d} \|x - (m + \mathrm{V}s)\|^2 = \operatorname*{argmin}_{s \in \mathbb{R}^d} \|x - (m + s_1 v_1 + \ldots + s_d v_d)\|^2.$$

The formula can be obtained by the least squares solution to the problem $m + \mathrm{V}r = x$:

$$r_1 v_1 + \ldots + r_k v_k = x - m,$$

which is given by:

$$r = (\mathrm{V}^T \mathrm{V})^{-1} \mathrm{V}^T (x - m) \in \mathbb{R}^d.$$

## 2.2 Integration on subspaces

For the integration over $C^1$ submanifolds of $\mathbb{R}^D$ refer the reader to [?, ?]. If we are given an a $C^1$ submanifold $M$ of dimension $d$ of $\mathbb{R}^D$, then we have a default restriction of Lebesgue measure to $M$, which we denote by $\lambda_d$ (formally, it is the normalization of $d$-dimensional Haar measure).

In the case we are interested in, when $M$ is an affine subspace, to integrate a function over $M$ we can take a point $m \in M$ and base V of the linear part of $M$, and then

$$\int_M f(x) d\lambda_d(x) = \det(\mathrm{V}^T \mathrm{V})^{1/2} \int_{\mathbb{R}^d} f(m + \mathrm{V}r) d\lambda_d(r).$$

With respect to measure $\lambda_d$ in $\mathbb{R}^D$ we can consider the singular densities (that is those defined only on $M$, or equivalently zero except for $M$). In the most important case of Gausian densities, if $m \in \mathbb{R}^D$ and $\Sigma$ is a symmetric nonnegative matrix with rank $d$, then by $N(m, \Sigma)$ we denote the function with support in $M = \{m + \Sigma^{1/2} r : r \in \mathbb{R}^D\}$ and the density given by

$$\mathcal{N}(m, \Sigma)(x) = \frac{1}{\sqrt{\det^*(2\pi\Sigma)}} e^{-\frac{1}{2}(x-m)^T \Sigma^\dagger (x-m)} \text{ for } x \in M,$$

where $\Sigma^\dagger$ is the generalized Moore-Penrose inverse and $\det^*$ is the pseudo-determinant[1].

## 2.3 Push-forward of measures

Since we know how to integrate functions on affine subspaces, let us discuss the natural method of defining (by push-forward) measures on such subspaces, for more information see `https://en.wikipedia.org/wiki/Pushforward_measure`, `http://www.mat.univie.ac.at/~gerald/ftp/book-fa/index.html` (page 256) and [?]. We assume as before, that $M$ is an affine subspace of $\mathbb{R}^D$ of dimension $d$, and that we fix $m$ (cordinate center) and V (base of linear part of $M$). Assume that we are given an affine function

$$a : \mathbb{R}^d \ni r \to m + \mathrm{V}r \in M \subset \mathbb{R}^D.$$

Then $m$ and V introduce a coordinate system on $V$, with center at $m$.

Suppose that we are given a measure $\mu$ on $\mathbb{R}^d$ with density $f$. Then we can push-forward (transport) the measure $\mu$ onto $M$ through the map $a$ to obtain the measure by the formula

$$(a_* \mu)(B) = \mu(a^{-1} B) \text{ for } B \subset \mathbb{R}^D.$$

[1] That is the product of all nonzero eigenvalues.

By applying the knowledge of integration over submanifolds, we obtain that the measure $a_*\mu$ with support in $M$ has the singular density with respect to $\lambda_d$ given by

$$f_{m,\mathrm{V}}(x) = \begin{cases} \frac{1}{\sqrt{\det(\mathrm{V}^T\mathrm{V})}} f(a_{m,\mathrm{V}}^{-1}x) \text{ if } x \in M, \\ 0 \text{ otherwise.} \end{cases} \tag{1}$$

Roughly speaking the above means that if we have a data-set $W \subset \mathbb{R}^d$ which comes from the density $f$ on $\mathbb{R}^d$, then $a_{m,V}(W)$ is clearly supported in $M$ and comes from the singular density $f_{m,\mathrm{V}}$ given by (1).

In the particual case when $W$ comes from the normal density $\mathcal{N}(m_d, \Sigma_d)$ in $\mathbb{R}^d$, then $a_{m,\mathrm{V}}(W)$ has the singular normal density $\mathcal{N}(m + \mathrm{V}m_d, \mathrm{V}^T\Sigma_d\mathrm{V})$ in $\mathbb{R}^D$.

### 2.4 Measure of nongaussianity

We consider the similar idea to the Kullback-Leibler.

### 2.5 Construction of densities

We can define the family of singular densities on affine subspaces of dimension $d$, by taking the transport.

In this subsection we describe the basic construction of product measures and densities. Given functions $f_1, f_2$ on $\mathbb{R}^{d_1}, \mathbb{R}^{d_2}$ by

$$(f_1 \otimes f_2)(x_1, x_2) = f_1(x_1) \cdot f_2(x_2) \text{ for } (x_1, x_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$$

we denote the tensor product of $f_1$ and $f_2$. Observe that if $f_1, f_2$ are densities, then so is $f_1 \otimes f_2$.

If $\mathcal{F}$ is a family of densities on $\mathbb{R}$, then by $\mathcal{F}^{\otimes d}$ ($d$-th tensor power of $\mathcal{F}$) we denote the family of densities on $\mathbb{R}^d$ given by

$$\mathcal{F}^{\otimes k} = \{f_1 \otimes \ldots \otimes f_d : f_i \in \mathcal{F}\}.$$

### 2.6 Our case

We assume that we have a family $\mathcal{F}$ larger then Gaussians on $\mathbb{R}$.

We have

$$\mathrm{KL}(X, \mathrm{aff}(\mathcal{S}^{\otimes d}), \mathcal{G}^d) = \inf_{m,\mathrm{V}} \mathrm{KL}((v^T\mathrm{V})^{-1}\mathrm{V}^T(X - m), \mathcal{S}^{\otimes d}, \mathcal{G}).$$

Notation: $x[m, \mathrm{V}]$. By the $i$-th coordinate we denote $x[m, \mathrm{V}]_i$.

Thus

$$\mathrm{KL}(X, \mathrm{aff}(\mathcal{S}^{\otimes d}), \mathcal{G}^d) = \inf_{m,\mathrm{V}} \left( \sum_{i=1}^{d} \mathrm{mle}(X[m, \mathrm{V}]_i, \mathcal{S}) - \mathrm{mle}(X[m, \mathrm{V}], \mathcal{G}) \right),$$

where the minus has the direct formula which can be computed.

## 3 Main idea

We want to find an index which would have the following characteristics:

1. the more non-gaussian data the better,
2. for gaussian data the value zero,

3. invariant under affine transformations.

4. ???? $k(f * N) < k(f)$ which implies the minimization?

**Theorem 3.1.** *? Theorem: in the perfect split we obain original split ?*

*Proof.* We have two random variables which are independent, the second Gaussian. Observe that if the change of coordinates, then sum of independent variables.

We search for minimal entropy (maximal likelihood). Since [Original Entropy Power Inequality]

$$e^{2H(X+Y)} \geq e^{2H(X)} + e^{2H(Y)},$$

and the equality holds only for the gaussians, □

We propose the possible solution for the ICA. We assume that we are given an affine-invariant family $\mathcal{F}$ of densities on $\mathbb{R}^D$, which contains normal densities $\mathcal{G}$ (Gaussians). To measure the distance from normality, we define an analogue of Kullback-Leibler divergence [sprawdzic znak, jak entropia to odwrotnie?]:

$$\mathrm{KL}(X, \mathcal{F}, \mathcal{G}) = \mathrm{mle}(X, \mathcal{F}) - \mathrm{mle}(X, \mathcal{G}).$$

[czy bierzemy znormalizowane - czy sumaryczne?]

Observe that for a fixed data the second element depends only on the covariance of the data. On the other hand, the first component typically has to be optimized by some gradient methods. Since the formula for the previous part is known

$$\mathrm{mle}(X, \mathcal{G}) = \mathrm{card}X(-\frac{1}{2}\ln|\Sigma_X| - \frac{D}{2}\ln(2\pi e)).$$

Now consider the situation where we are given a task of finding dimension on possibly smaller space of dimension $d \leq D$. In this case assume that we are given a family $\mathcal{F}^d$ on $\mathbb{R}^d$, where $d \leq D$ (we do not assume that $\mathcal{F}^d$ is affine invariant, as we obtain it directly from the construction by the fact that we can adapt the base). To fix an affine space $V$ of dimension $d$ in $\mathbb{R}^D$ we choose its center $m$ and $d$ linearly independent elements $\mathrm{V} = v_1, \ldots, v_d \in \mathbb{R}^D$.

Now the coordinates[2] in the base V of orthogonal projection of $x \in \mathbb{R}^D$ onto $V$ is given by

$$\lambda^x_{m,\mathrm{V}} = (\mathrm{V}^T\mathrm{V})^{-1}\mathrm{V}^T(x - m) \in \mathbb{R}^d \text{ and } x_{m,v} = m + \mathrm{V}\lambda^x_{m,\mathrm{V}}. \tag{2}$$

By $\Lambda_{m,\mathrm{V}} = (\lambda^x_{m,v})$ we denote the coordinates of the whole data set. Now we can project the data to this space, and in those coordinates we can measure the previously defined Kullback-Leibler generalized divergence:

$$(m, \mathrm{V}) \to \mathrm{KL}(\Lambda_{m,\mathrm{V}}, \mathcal{F}^d, \mathcal{G}). \tag{3}$$

The minimization of the above function leads to the solution of the ICA problem on the respective subspace.

We will consider it for the family $\mathcal{F}$ of split Gaussians, however, one can apply any family used in the ICA process.

It occurs that under weak assumption we can even rank the base vectors of V. To do so suppose that $\mathcal{S}^{\otimes d}$ is given as tensor product $\mathcal{S}^{\otimes d} = \mathcal{S} \otimes \cdots \otimes \mathcal{S}$, where $\mathcal{S}$ denotes a family of densities on $\mathbb{R}$ (this is the case of split Gaussians). In other words we assume that every element of $F \in \mathcal{S}^d$ can be decomposed in the form

$$F(x_1, \ldots, x_d) = f_1(x_1) \cdot \ldots \cdot f_d(x_d) \text{ where } f_i \in \mathcal{S}.$$

[2] The formula is the direct consequence of the fact that the orthogonal projection is exactly the solution of least squares solution of the equations $v\alpha = x - m$, where $\alpha = (\alpha_1, \ldots, \alpha_d)^T$

Notation $\mathrm{aff}(S^{\otimes d})$ – will denote the space of affine. If we are given a density $f$ on $\mathbb{R}^d$, and an affine map $A : \mathbb{R}^d \ni \lambda \to m + \mathrm{V}\lambda$, then the degenerate density on the space $V$ with respect to the $d$-dimensional Lebesgue (Haar) measure $\lambda_d$ is given by

$$f_V : V \ni x \to \frac{1}{|A|} f(A^{-1}x)$$

where $|A|$ is the generalization of determinant given by ... The formula for the KL is therefore given by

$$\sum_x \ln f(A^{-1}p_V x) - \ln N(A^{-1}p_V x).$$

Observe that $\Sigma\Lambda_V = (A^{-1}p_V)\Sigma(A^{-1}p_V)^T$. Consequently, the minus part equals

$$\mathrm{card}X(-\frac{1}{2}\ln|(A^{-1}p_V)\Sigma(A^{-1}p_V)^T| - \frac{d}{2}\ln(2\pi e)).$$

PROCEDURE to compute $\mathrm{KL}^d_{m,\mathrm{V}}(X, \mathcal{S})$:

- data $X$ and family of one-dimensional densities on $\mathcal{S}$ given,
- fix $m, \mathrm{V}$,
- put $\Lambda = (\lambda^x_{m,\mathrm{V}})_{x \in X} \subset \mathbb{R}^d$,
- by $\Lambda_i$ we denote the set consisting of $i$-th coordinate of $\Lambda$,
- compute[3]

$$\mathrm{KL}(\Lambda, \mathcal{S}^d, \mathcal{G}) = \sum_{i=1}^{d} \mathrm{mle}(\Lambda_i, \mathcal{S}) - \mathrm{mle}(\Lambda, \mathcal{G}).$$

We put

$$\mathrm{KL}^d(X, \mathcal{S}) = \inf \mathrm{KL}^d_{m,\mathrm{V}}(X, \mathcal{S}).$$

**Theorem 3.2.** *a) Independent of affine transformations b) czy mozemy sie zawezic do popdrzestrzeni*

**Problem 3.1.** czy jest znany wzor dla mle przy split gaussian?

**Theorem 3.3.**

Now suppose that we have found a base $m, \mathrm{V}$ which minimizes (3). Denote by $(\alpha)_i$ $i$-th coordinate of $\alpha$, then we can rank the vectors according to the non-gaussianity of the $i$-th coordinate of the projection:

$$i \to \mathrm{KL}((X_{m,\mathrm{V}})_i, \mathcal{F}^1, \mathcal{G}).$$

We want to introduce a new measure to see if the subspace we found is correct. The model has to be affine independent. To do so, assume that we are given data $X = (x_i)$ and the transformed/obtained data $\tilde{X} = (\tilde{x}_i)$. We define the measure between the best affine transformation between data, to do so by mean squares we solve the problem

$$A\tilde{x}_i + b = x_i.$$

The mean squarred error is the desired value:

$$i(X, \tilde{X}) = \frac{1}{N} \sum_{i=1}^{N} \|x_i - (A\tilde{x}_i + b)\|^2.$$

[3] sometimes we need optimization

**Example 3.1.** Take the real-data $X \subset \mathbb{R}^d$, add the next $D-d$ coordinates by some normal density – we obtain new data set $\tilde{X}$. Try to find the first $d$ coordinates.

Measure the value of

$$i(X, \tilde{X}).$$

**Example 3.2.** Take the real-data $X \subset \mathbb{R}^d$, add the next $D-d$ coordinates with zeros. Next perturb all coordinates by some normal density. Try to find the first $d$ coordinates. Come back by least squares between the original coordinates and the projection.

# 4 Przemek

The density of the one-dimensional Split Gaussian distribution is given by the formula

$$SN(x; m, \sigma^2, \tau^2) = \begin{cases} c \cdot \exp[-\frac{1}{2\sigma^2}(x-m)^2], & \text{where } x \le m \\ c \cdot \exp[-\frac{1}{2\tau^2\sigma^2}(x-m)^2], & \text{where } x > m \end{cases}$$

where $c = \sqrt{\frac{2}{\pi}}\sigma^{-1}(1+\tau)^{-1}$.

A natural generalization of the univariate split normal distribution to the multivariate settings was presented by [?]. Roughly speaking, authors assume that a vector $\mathrm{x} \in \mathbb{R}^d$ follows the multivariate Split Normal distribution, if its principal components are orthogonal and follow the one-dimensional Split Normal distribution.

**Definition 4.1.** A density of the multivariate Split Normal distribution is given by

$$SN_d(\mathrm{x}; \mathrm{m}, \sigma, \tau) = \prod_{j=1}^{d} SN(x_j; m_j, \sigma_j^2, \tau_j^2),$$

where $\mathrm{m} = [m_1, \dots, m_d]^T$, $\sigma = [\sigma_1^2, \dots, \sigma_d^2]^T$ and $\tau = [\tau_1^2, \dots, \tau_d^2]^T$.

In our case we will use density on projection on $d < D$ subspaces. Therefore we need a density $d$-subspace Split Normal distribution.

**Definition 4.2.** A density of the multivariate $d$-subspace Split Normal distribution is given by

$$SN_{d<D}(\mathrm{x}; \mathrm{m}, W, \sigma^2, \tau^2) = SN_d((W^TW)^{-1}W^T(\mathrm{x}-\mathrm{m}); 0, \sigma^2, \tau^2),$$

where $(W^TW)^{-1}W^T(x-m) \in \mathbb{R}^d$ $\omega_j \in \mathbb{R}^D$ is the $j$-th column of non-singular matrix $W = [w_1, \dots, w_d]$, $\mathrm{m} = [m_1, \dots, m_D]^T$, $\sigma = [\sigma_1, \dots, \sigma_d]^T$ and $\tau = [\tau_1, \dots, \tau_d]^T$.

Let us recall that the standard Gaussian density in $\mathbb{R}^d$ is defined by

$$N(\mathrm{x}; \mathrm{m}, \Sigma) = \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \exp\left(-\tfrac{1}{2}(\mathrm{x}-\mathrm{m})^T\Sigma^{-1}(\mathrm{x}-\mathrm{m})\right),$$

where m denotes the mean, $\Sigma$ is the covariance matrix.

**Definition 4.3.** A density of the multivariate $d$-subspace Normal distribution is given by

$$N_{d<D}(\mathrm{x}; \mathrm{m}, \Sigma, W) = N((W^TW)^{-1}W^T(\mathrm{x}-\mathrm{m}); 0, \Sigma),$$

where $(W^TW)^{-1}W^T(\mathrm{x}-\mathrm{m}) \in \mathbb{R}^d$ $\omega_j \in \mathbb{R}^D$ is the $j$-th column of non-singular matrix $W = [w_1, \dots, w_d]$, $\mathrm{m} = [m_1, \dots, m_D]^T$, $\Sigma = \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2)$.

Our goal is to minimize

$$\mathrm{KL}(X, \mathcal{F}, \mathcal{G}) = \mathrm{mle}(X, \mathcal{F}) - \mathrm{mle}(X, \mathcal{G})$$

In our language

$$\begin{aligned} &\mathrm{KL}_{d<D}(X; \mathrm{m}, W, \sigma, \tau, \Sigma) = \\ &= \sum_{\mathrm{x} \in X} \ln(SN_{d<D}(\mathrm{x}; \mathrm{m}, W, \sigma, \tau)) - \sum_{\mathrm{x} \in X} \ln(N_{d<D}(\mathrm{x}; \mathrm{m}, \Sigma, W)) \end{aligned} \tag{4}$$

We known

$$\sum_{\mathrm{x} \in X} \ln(N_{d<D}(\mathrm{x}; \mathrm{m}, \Sigma, W)) = -\frac{d}{2} \ln(2\pi e) - \frac{1}{2} \ln \det(\Sigma_W),$$

where

$$\Sigma_W = \mathrm{cov}(\{(W^T W)^{-1} W^T (\mathrm{x} - \mathrm{m}) \colon \mathrm{x} \in \mathbb{R}^D\})$$

## 4.1 Optimization problem

The density of the multivariate d-subspace Normal distribution depends on four parameters $\mathrm{m} \in \mathbb{R}^d$, $W \in \mathcal{M}(\mathbb{R}^d)$, $\sigma \in \mathbb{R}^d$, $\tau \in \mathbb{R}^d$. We can find them by minimizing the simpler function, which depends on only $m \in \mathbb{R}^d$ and $W \in \mathcal{M}(\mathbb{R}^d)$. Other parameters are given by explicit formulas. Let us notice that in this case our minimization problem simplifies to minimizing the function $\mathrm{mle}(X, \mathcal{F}) = \sum_{\mathrm{x} \in X} \ln(SN_{d<D}(\mathrm{x}; \mathrm{m}, W, \sigma, \tau))$

**Theorem 4.1.** *Let $\mathrm{x}_1, \ldots, \mathrm{x}_n$ be given. Then the likelihood maximized w.r.t. $\sigma$ and $\tau$ is*

$$\hat{L}(X; \mathrm{m}, W) = \left(\frac{2n}{\pi e}\right)^{dn/2} \left(\prod_{j=1}^{d} g_j(\mathrm{m}, W)\right)^{-3n/2}, \tag{5}$$

*where*

$$\begin{aligned} g_j(\mathrm{m}, W) &= s_{1j}^{1/3} + s_{2j}^{1/3}, \\ s_{1j} &= \sum_{i \in I_j} [\omega_j^T (\mathrm{x}_i - \mathrm{m})]^2, I_j = \{i = 1, \ldots, n \colon \omega_j^T (\mathrm{x}_i - \mathrm{m}) \leq 0\}, \\ s_{2j} &= \sum_{i \in I_j^c} [\omega_j^T (\mathrm{x}_i - \mathrm{m})]^2, I_j^c = \{i = 1, \ldots, n \colon \omega_j^T (\mathrm{x}_i - \mathrm{m}) > 0\}, \end{aligned}$$

*where $\omega_j$ is the $j$-th column of non-singular matrix $(W^T W)^{-1} W^T$ and the maximum likelihood estimators of $\sigma_j^2$ and $\tau_j$ are*

$$\hat{\sigma}_j^2(\mathrm{m}, W) = \tfrac{1}{n} s_{1j}^{2/3} g_j(\mathrm{m}, W), \quad \hat{\tau}_j(\mathrm{m}, W) = \left(\frac{s_{2j}}{s_{1j}}\right)^{1/3}.$$

*Proof of Theorem 4.1.* Let $X = \{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ and $W_\omega = (W^T W)^{-1} W^T$. We write

$$\mathrm{z}_i = W_\omega(\mathrm{x}_i - m), \quad \mathrm{z}_{ij} = \omega_j^T (\mathrm{x}_i - m),$$

for observation $i$, where $i = 1, \ldots, n$ and coordinates $j = 1, \ldots, d$.

Let us consider the likelihood function, i.e.

$$\begin{aligned} L(X; \mathrm{m}, W, \sigma, \tau) = \prod_{i=1}^{n} SN_{d<D}(\mathrm{x}_i; \mathrm{m}, W, \sigma, \tau) == \sum_{\mathrm{x} \in X} \ln(SN_{d<D}(\mathrm{x}; \mathrm{m}, W, \sigma, \tau)) \prod_{i=1}^{n} \prod_{j=1}^{d} SN(\omega_j^T (\mathrm{x}_i - \mathrm{m}); 0, \sigma^2, \tau^2) \\ = c_1^n \big( \prod_{j=1}^{d} \sigma_j (1 + \tau_j) \big)^{-n} \prod_{i=1}^{n} \prod_{j=1}^{d} \exp[-\tfrac{1}{2\sigma_j^2} z_{ij}^2 (\mathbb{1}_{\{z_{ij} > 0\}})], \end{aligned}$$

where $c_1 = \left(\sqrt{\frac{2}{\pi}}\right)^d$. Now we take the log-likelihood function, i.e.

$$\begin{aligned}
&\ln(L(X;\mathrm{m},W,\sigma,\tau)) \\
&= \ln\left(c_1^n\Big(\prod_{j=1}^{d}\sigma_j(1+\tau_j)\Big)^{-n}\right) + \sum_{i=1}^{n}\sum_{j=1}^{d}\Big[-\tfrac{1}{2\sigma_j^2}z_{ij}^2(\mathbb{1}_{\{z_{ij}\leq 0\}} + \tau_j^{-2}\mathbb{1}_{\{z_{ij}>0\}})\Big] \\
&= \ln\left(c_1^n\Big(\prod_{j=1}^{d}\sigma_j(1+\tau_j)\Big)^{-n}\right) - \tfrac{1}{2}\sum_{j=1}^{d}\Big(\sigma_j^{-2}\sum_{i\in I_j} z_{ij}^2 + \tfrac{\sigma_j^{-2}}{\tau_j^2}\sum_{i\in I_j^c} z_{ij}^2\Big) \\
&= \ln\left(c_1^n\Big(\prod_{j=1}^{d}\sigma_j(1+\tau_j)\Big)^{-n}\right) - \sum_{j=1}^{d}\tfrac{1}{2\sigma_j^2}\Big(s_{1j} + \tfrac{1}{\tau_j^2}s_{2j}\Big).
\end{aligned}$$

We fix m, $W$ and maximize the log-likelihood function over $\tau$ and $\sigma$. In such a case we have to solve the following system of equations

$$\begin{aligned}
\tfrac{\partial \ln(L(X;\mathrm{m},W,\sigma,\tau))}{\partial \sigma_j} &= -\tfrac{n}{\sigma_j} + \sigma_j^{-3}(s_{1j} + \tau_j^{-2}s_{2j}) = 0, \\
\tfrac{\partial \ln(L(X;\mathrm{m},W,\sigma,\tau))}{\partial \tau_j} &= -\tfrac{n}{1+\tau_j} + \tfrac{s_{2j}}{\tau_j^3\sigma_j^2} = 0,
\end{aligned}$$

for $j = 1, \ldots, d$. By simple calculations we obtain the expressions for the estimators

$$\hat{\sigma}_j^2(\mathrm{m},W) = \tfrac{1}{n}s_{1j}^{2/3}g_j(\mathrm{m},W), \qquad \hat{\tau}_j(\mathrm{m},W) = \left(\frac{s_{2j}}{s_{1j}}\right)^{1/3}.$$

Substituting it into the log-likelihood function, we get

$$\begin{aligned}
\hat{L}(\mathrm{m},W) &= \left(\tfrac{2}{\pi}\right)^{\frac{dn}{2}}\Big(\prod_{j=1}^{d}\tfrac{1}{\sqrt{n}}g_j(\mathrm{m},W)^{\frac{3}{2}}\Big)^{-n}e^{-\frac{dn}{2}} \\
&= \left(\tfrac{2n}{\pi e}\right)^{\frac{dn}{2}}\Big(\prod_{j=1}^{d}g_j(\mathrm{m},W)\Big)^{-\frac{3n}{2}}.
\end{aligned}$$

□

Thanks to the above theorem, instead of looking for the maximum of the likelihood function, it is enough to obtain the maximum of the simpler function (5) which depends on two parameters $\mathrm{m} \in \mathbb{R}^d$ and $W \in \mathcal{M}(\mathbb{R}^d)$

$$l(X;\mathrm{m},W) = \prod_{j=1}^{d} g_j(\mathrm{m},W) \tag{6}$$

where $\omega_j$ stands for the $j$-th column of matrix $W$. Consequently, maximization of (5) is equivalent to minimization of (6), see the following corollary.

**Corollary 4.1.** *Let $X \subset \mathbb{R}^d$, $\mathrm{m} \in \mathbb{R}^d$, $W \in \mathcal{M}(\mathbb{R}^d)$ be given, then*

$$\mathit{argmax}_{\mathrm{m},W}\hat{L}(X;\mathrm{m},W) = \operatorname*{argmin}_{\mathrm{m},W} l(X;\mathrm{m},W).$$

## 4.2 Gradient

One of the possible methods of optimization is the gradient method. Since the minimum of $l$ is equal to the minimum of $\ln(l)$, in this subsection we calculate the gradient of $\ln(l)$. Before we prove suitable Theorem 4.2, we recall the following lemma.

**Lemma 4.1.** *Let $A = (a_{ij})_{1\leq i,j\leq d}$ be a differentiable map from real numbers to $d \times d$ matrices then*

$$\frac{\partial \det(A)}{\partial a_{ij}} = \mathrm{adj}^T(A)_{ij}, \tag{7}$$

*where* $\mathrm{adj}(A)$ *stands for the adjugate of* $A$*, i.e. the transpose of the cofactor matrix.*

*Proof.* By the Laplace expansion $\det A = \sum\limits_{j=1}^{d}(-1)^{i+j}a_{ij}M_{ij}$ where $M_{ij}$ is the minor of the entry in the $i$-th row and $j$-th column. Hence

$$\frac{\partial \det A}{\partial a_{ij}} = (-1)^{i+j}M_{ij} = \mathrm{adj}^T(A)_{ij}.$$

□

Now we are ready to calculate gradient of our cost function.

**Theorem 4.2.** *Let $X \subset \mathbb{R}^d$, $\mathrm{m} = (\mathrm{m}_1, \dots, \mathrm{m}_d)^T \in \mathbb{R}^d$, $W = (\omega_{ij})_{1\leq i,j\leq d}$ non-singular be given. Then $\nabla_{\mathrm{m}} \ln l(X;\mathrm{m},W) = \left(\frac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_1}, \dots, \frac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_d}\right)^T$, where*

$$\tfrac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_k} = \sum_{j=1}^{d} \tfrac{-1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}}\left(\tfrac{1}{3s_{1j}^{\frac{2}{3}}}\sum_{i\in I_j} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} + \tfrac{1}{3s_{2j}^{\frac{2}{3}}}\sum_{i\in I_j^c} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk}\right).$$

*Moreover, $\nabla_W \ln l(X;\mathrm{m},W) = \left[\frac{\partial \ln l(X;\mathrm{m},W)}{\partial \omega_{pk}}\right]_{1\leq p,k\leq d}$, where*

$$\tfrac{\partial \ln l(X;\mathrm{m},W)}{\partial \omega_{pk}} = \tfrac{1}{s_{1p}^{\frac{1}{3}}+s_{2p}^{\frac{1}{3}}}\left(\tfrac{1}{3}s_{1p}^{-\frac{2}{3}}\sum_{i\in I_p} 2\omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) + +\tfrac{1}{3}s_{2p}^{-\frac{2}{3}}\sum_{i\in I_p^c} 2\omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k)\right).$$

*and*

$$\begin{aligned} s_{1j} &= \sum_{i\in I_j}[\omega_j^T(\mathrm{x}_i - \mathrm{m})]^2, I_j = \{i = 1, \dots, n \colon \omega_j^T(\mathrm{x}_i - \mathrm{m}) \leq 0\},\\ s_{2j} &= \sum_{i\in I_j^c}[\omega_j^T(\mathrm{x}_i - \mathrm{m})]^2, I_j^c = \{i = 1, \dots, n \colon \omega_j^T(\mathrm{x}_i - \mathrm{m}) > 0\}.\end{aligned}$$

*Proof of Theorem 4.2.* Let us start with the partial derivative of $\ln(l)$ with respect to m. We have

$$\tfrac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_k} = \sum_{j=1}^{d} \tfrac{\partial \ln(g_j(\mathrm{m},W))}{\partial \mathrm{m}_k} = \sum_{j=1}^{d} \tfrac{1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}} \tfrac{\partial(s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}})}{\partial \mathrm{m}_k} \sum_{j=1}^{d} \tfrac{1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}}\left(\tfrac{1}{3s_{1j}^{\frac{2}{3}}}\tfrac{\partial s_{1j}}{\partial \mathrm{m}_k} + \tfrac{1}{3s_{2j}^{\frac{2}{3}}}\tfrac{\partial s_{2j}}{\partial \mathrm{m}_k}\right).$$

Now, we need $\frac{\partial s_{1j}}{\partial \mathrm{m}_k}$ and $\frac{\partial s_{2j}}{\partial \mathrm{m}_k}$, therefore

$$\tfrac{\partial s_{1j}}{\partial \mathrm{m}_k} = \sum_{i\in I_j} \tfrac{\partial[\omega_j^T(\mathrm{x}_i - \mathrm{m})]^2}{\partial \mathrm{m}_k} = \sum_{i\in I_j} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\tfrac{\partial \omega_j^T(\mathrm{x}_i - \mathrm{m})}{\partial \mathrm{m}_k} = \sum_{i\in I_j} -2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk}.$$

Analogously we get

$$\tfrac{\partial s_{2j}}{\partial \mathrm{m}_k} = \sum_{i\in I_j^c} -2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk}.$$

Hence

$$\tfrac{\partial \ln l}{\partial \mathrm{m}_k} = \sum_{j=1}^{d} \tfrac{-1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}}\left(\tfrac{1}{3s_{1j}^{\frac{2}{3}}}\sum_{i\in I_j} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} + \tfrac{1}{3s_{2j}^{\frac{2}{3}}}\sum_{i\in I_j^c} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk}\right).$$

Now we calculate the partial derivative of $\ln l(X;\mathrm{m},W)$ with respect to the matrix $W$. We have

$$\tfrac{\partial \ln l(X;\mathrm{m},W)}{\partial \omega_{pk}} = \sum_{j=1}^{d} \tfrac{\partial \ln(g_j(\mathrm{m},W))}{\partial \omega_{pk}}.$$

Now we calculate

$$\tfrac{\partial \ln(g_j(\mathrm{m},W))}{\partial \omega_{pk}} = \tfrac{1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}} \tfrac{\partial(s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}})}{\partial \omega_{pk}} = \tfrac{1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}} \left( \tfrac{1}{3s_{1j}^{\frac{2}{3}}} \tfrac{\partial s_{1j}}{\partial \omega_{pk}} + \tfrac{1}{3s_{2j}^{\frac{2}{3}}} \tfrac{\partial s_{2j}}{\partial \omega_{pk}} \right),$$

where

$$\tfrac{\partial s_{1j}}{\partial \omega_{pk}} = \sum_{i \in I_j} \tfrac{\partial [\omega_j^T(\mathrm{x}_i - \mathrm{m})]^2}{\partial \omega_{pk}} = \sum_{i \in I_j} 2\omega_j^T(\mathrm{x}_i - \mathrm{m}) \tfrac{\partial \omega_j^T(\mathrm{x}_i - \mathrm{m})}{\partial \omega_{pk}} =$$
$$\begin{cases} 0, & \text{if } j \neq p \\ \sum\limits_{i \in I_p} 2\omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k), & \text{if } j = p \end{cases}$$

and $\mathrm{x}_{ik}$ is the $k$-th element of the vector $\mathrm{x}_i$. Analogously we get

$$\frac{\partial s_{2j}}{\partial \omega_{pk}} = \begin{cases} 0, & \text{if } j \neq p \\ \sum\limits_{i \in I_p^c} 2\omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k), & \text{if } j = p. \end{cases}$$

Hence we obtain

$$\tfrac{\partial \ln l}{\partial \omega_{pk}} = \tfrac{1}{s_{1p}^{\frac{1}{3}}+s_{2p}^{\frac{1}{3}}} \left( \tfrac{1}{3} s_{1p}^{-\frac{2}{3}} \sum_{i \in I_p} 2\omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) + \tfrac{1}{3} s_{2p}^{-\frac{2}{3}} \sum_{i \in I_p^c} 2\omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) \right).$$

□

Hence, we get

$$\tfrac{\partial \ln \hat{L}}{\partial \mathrm{m}_k} = -\tfrac{3n}{2} \sum_{j=1}^{d} \tfrac{-1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}} \left( \tfrac{1}{3s_{1j}^{\frac{2}{3}}} \sum_{i \in I_j} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} + \tfrac{1}{3s_{2j}^{\frac{2}{3}}} \sum_{i \in I_j^c} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} \right).$$

and

$$\tfrac{\partial \ln \hat{L}}{\partial \omega_{pk}} = -\tfrac{3n}{2} \tfrac{1}{s_{1p}^{\frac{1}{3}}+s_{2p}^{\frac{1}{3}}} \left( \tfrac{1}{3} s_{1p}^{-\frac{2}{3}} \sum_{i \in I_p} 2\omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) + \tfrac{1}{3} s_{2p}^{-\frac{2}{3}} \sum_{i \in I_p^c} 2\omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) \right).$$

Now we can calculate gradients of the function $\sum\limits_{\mathrm{x} \in X} \ln(N_{d<D}(\mathrm{x};\mathrm{m},\Sigma,W)) = -\frac{d}{2}\ln(2\pi e) - \frac{1}{2}\ln\det(\Sigma_W)$.

**Lemma 4.2.** *Let $X \subset \mathbb{R}^d$, $\mathrm{m} = (\mathrm{m}_1, \ldots, \mathrm{m}_d)^T \in \mathbb{R}^d$, $W = (\omega_{ij})_{1 \leq i,j \leq d}$ non-singular be given. Let $C(\mathrm{m},W) = \ln\det(\Sigma_W)$ then $\nabla_{\mathrm{m}} C(X;\mathrm{m},W) = \left(\frac{\partial C(X;\mathrm{m},W)}{\partial \mathrm{m}_1}, \ldots, \frac{\partial C(X;\mathrm{m},W)}{\partial \mathrm{m}_d}\right)^T$, where*

$$\tfrac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_k} = \sum_{j=1}^{d}$$

*Moreover, $\nabla_W C(X;\mathrm{m},W) = \left[\frac{\partial C(X;\mathrm{m},W)}{\partial \omega_{pk}}\right]_{1 \leq p,k \leq d}$, where*

*Proof.*

$$C(X;\mathrm{m},W) = \ln\det(\Sigma_W) =$$

Let us notice that

$$\Sigma_W = (W^TW)^{-1}W^T\mathrm{cov}(X)\left((W^TW)^{-1}W^T\right)^T = (W^TW)^{-1}W^T\mathrm{cov}(X)W\left((W^TW)^{-1}\right)^T$$
$$= (W^TW)^{-1}\mathrm{cov}(W^TX)((W^TW)^{-1})^T$$

Hence

$$\det(\Sigma_W) = \det((W^TW)^{-1})\det(W^TX)\det(((W^TW)^{-1})^T)$$

and

$$\ln\det(\Sigma_W) = \ln\det((W^TW)^{-1}) + \ln\det(\mathrm{cov}(W^TX)) + \ln\det(((W^TW)^{-1})^T)$$
$$= 2\ln\det((W^TW)^{-1}) + \ln\det(W^TX) = 2\ln\frac{1}{\det(W^TW)} + \ln\det(\mathrm{cov}(W^TX))$$

Moreover,

$$\frac{\partial\mathrm{cov}(W^TX)}{\partial W} = \frac{\partial W^T\mathrm{cov}(X)W}{\partial W} = (\mathrm{cov}(X)+\mathrm{cov}(X)^T)W = 2\mathrm{cov}(X)W$$

and

$$\frac{\partial(W^TW)}{\partial W} = 2W$$

$$\frac{\partial C(X;\mathrm{m},W)}{\partial\omega_{pk}} = 2\det(W^TW)\frac{-1}{(\det(W^TW))^2}\mathrm{adj}^T(W^TW)2W + \frac{1}{\det(\mathrm{cov}(W^TX))}\mathrm{adj}^T(\mathrm{cov}(W^TX))2\mathrm{cov}(X)W$$
$$= \frac{-2}{\det(W^TW)}\mathrm{adj}^T(W^TW)2W + \frac{1}{\det(\mathrm{cov}(W^TX))}\mathrm{adj}^T(\mathrm{cov}(W^TX))2\mathrm{cov}(X)W$$
$$= -4(W^TW)^{-1}W + 2(\mathrm{cov}(W^TX))^{-1}\mathrm{cov}(X)W$$

□

Summing up,

$$\frac{\partial\frac{1}{2}\ln\det(\Sigma_W)}{\partial\mathrm{m}_k} = 0$$

$$\frac{\partial\frac{1}{2}\ln\det(\Sigma_W)}{\partial\omega_{pk}} = -2(W^TW)^{-1}W + (\mathrm{cov}(W^TX))^{-1}\mathrm{cov}(X)W$$

**Theorem 4.3.** *Let $X\subset\mathbb{R}^d$, $\mathrm{m}=(\mathrm{m}_1,\ldots,\mathrm{m}_d)^T\in\mathbb{R}^d$, $W=(\omega_{ij})_{1\le i,j\le d}$ non-singular be given. Then $\nabla_{\mathrm{m}}KL(X;\mathrm{m},W) = \left(\frac{\partial KL(X;\mathrm{m},W)}{\partial\mathrm{m}_1},\ldots,\frac{\partial KL(X;\mathrm{m},W)}{\partial\mathrm{m}_d}\right)^T$, where*

$$\tfrac{\partial KL(X;\mathrm{m},W)}{\partial\mathrm{m}_k} = -\tfrac{3n}{2}\sum_{j=1}^{d}\tfrac{-1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}}\left(\tfrac{1}{3s_{1j}^{\frac{2}{3}}}\sum_{i\in I_j}2\omega_j^T(\mathrm{x}_i-\mathrm{m})\omega_{jk} + \tfrac{1}{3s_{2j}^{\frac{2}{3}}}\sum_{i\in I_j^c}2\omega_j^T(\mathrm{x}_i-\mathrm{m})\omega_{jk}\right).$$

*Moreover, $\nabla_W KL(X;\mathrm{m},W) = \left[\frac{\partial KL(X;\mathrm{m},W)}{\partial\omega_{pk}}\right]_{1\le p,k\le d}$, where*

$$\tfrac{\partial KL(X;\mathrm{m},W)}{\partial\omega_{pk}} = -\tfrac{n}{2}\tfrac{1}{s_{1p}^{\frac{1}{3}}+s_{2p}^{\frac{1}{3}}}\left(s_{1p}^{-\frac{2}{3}}\sum_{i\in I_p}2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k) + s_{2p}^{-\frac{2}{3}}\sum_{i\in I_p^c}2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k)\right)+$$
$$+(\mathrm{cov}(WX))^{-1}\mathrm{cov}(X)W^T.$$

*and*

$$s_{1j} = \sum_{i\in I_j}[\omega_j^T(\mathrm{x}_i-\mathrm{m})]^2, I_j = \{i=1,\ldots,n\colon\omega_j^T(\mathrm{x}_i-\mathrm{m})\le 0\},$$
$$s_{2j} = \sum_{i\in I_j^c}[\omega_j^T(\mathrm{x}_i-\mathrm{m})]^2, I_j^c = \{i=1,\ldots,n\colon\omega_j^T(\mathrm{x}_i-\mathrm{m})>0\}.$$

# 5 MODEL II

**Definition 5.1.** A density of the multivariate Split Normal $d$ and Normal $D-d$ distribution is given by

$$SN_d N_{D-d}(\mathrm{x}; \mathrm{m}, W, \sigma^2, \tau^2) = \det(W) \prod_{j=1}^{d} SN(\omega_j^T(\mathrm{x}-\mathrm{m}); 0, \sigma_j^2, \tau_j^2) \prod_{j=d+1}^{D} N(\omega_j^T(\mathrm{x}-\mathrm{m}); 0, \sigma_j^2),$$

where $\omega_j$ is the $j$-th column of non-singular matrix $W$, $\mathrm{m} = (m_1, \ldots, m_d)^T$, $\sigma = (\sigma_1, \ldots, \sigma_d)$ and $\tau = (\tau_1, \ldots, \tau_{D-d})$.

The density of the multivariate d-subspace Normal distribution depends on four parameters $\mathrm{m} \in \mathbb{R}^d$, $W \in \mathcal{M}(\mathbb{R}^d)$, $\sigma \in \mathbb{R}^d$, $\tau \in \mathbb{R}^d$. We can find them by minimizing the simpler function, which depends on only $m \in \mathbb{R}^d$ and $W \in \mathcal{M}(\mathbb{R}^d)$. Other parameters are given by explicit formulas. Let us notice that in this case our minimization problem simplifies to minimizing the function $\mathrm{mle}(X, \mathcal{F}) = \sum\limits_{\mathrm{x} \in X} \ln(SN_{d<D}(\mathrm{x}; \mathrm{m}, W, \sigma, \tau))$

The density of the GSN distribution depends on four parameters $\mathrm{m} \in \mathbb{R}^d$, $W \in \mathcal{M}(\mathbb{R}^d)$, $\sigma \in \mathbb{R}^d$, $\tau \in \mathbb{R}^d$. We can find them by minimizing the simpler function, which depends on only $m \in \mathbb{R}^d$ and $W \in \mathcal{M}(\mathbb{R}^d)$. Other parameters are given by explicit formulas.

**Theorem 5.1.** *Let* $\mathrm{x}_1, \ldots, \mathrm{x}_n$ *be given. Then the likelihood maximized w.r.t.* $\sigma$ *and* $\tau$ *is*

$$\hat{L}(X; \mathrm{m}, W) = C\Big(\frac{1}{|\det(W)|^{\frac{2}{3}}} \prod_{j=1}^{d} g_j(\mathrm{m}, W) \prod_{j=d+1}^{D} (s_{1j} + s_{2j})^{\frac{1}{3}}\Big)^{-3n/2}, \tag{8}$$

*where*

$$\begin{aligned} g_j(\mathrm{m}, W) &= s_{1j}^{1/3} + s_{2j}^{1/3}, \\ s_{1j} = &\sum_{i \in I_j} [\omega_j^T(\mathrm{x}_i - \mathrm{m})]^2, I_j = \{i = 1, \ldots, n \colon \omega_j^T(\mathrm{x}_i - \mathrm{m}) \le 0\}, \\ s_{2j} = &\sum_{i \in I_j^c} [\omega_j^T(\mathrm{x}_i - \mathrm{m})]^2, I_j^c = \{i = 1, \ldots, n \colon \omega_j^T(\mathrm{x}_i - \mathrm{m}) > 0\}, \end{aligned}$$

*and the maximum likelihood estimators of* $\sigma_j^2$ *and* $\tau_j$ *are*

$$\hat{\sigma}_j^2(\mathrm{m}, W) = \tfrac{1}{n} s_{1j}^{2/3} g_j(\mathrm{m}, W), \quad \hat{\tau}_j(\mathrm{m}, W) = \left(\frac{s_{2j}}{s_{1j}}\right)^{1/3}.$$

*Proof of Theorem 5.1.* Let $X = \{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$. We write

$$\mathrm{z}_i = W(\mathrm{x}_i - m), \quad \mathrm{z}_{ij} = \omega_j^T(\mathrm{x}_i - m),$$

for observation $i$, where $i = 1, \ldots, n$ and coordinates $j = 1, \ldots, d$.

Let us consider the likelihood function, i.e.

$$\begin{aligned} &L(X; \mathrm{m}, W, \sigma, \tau) = \prod_{i=1}^{n} SN_d N_{D-d}(\mathrm{x}_i; \mathrm{m}, W, \sigma^2, \tau^2) \\ &= \prod_{i=1}^{n} |\det(W)| \prod_{j=1}^{d} SN(\omega_j^T(\mathrm{x}_i - \mathrm{m}); 0, \sigma_j^2, \tau_j^2) \prod_{j=d+1}^{D} N(\omega_j^T(\mathrm{x}_i - \mathrm{m}); 0, \sigma_j^2) \\ &= \Big(c_1 |\det(W)|\Big)^n \Big(\prod_{j=1}^{d} \sigma_j (1 + \tau_j)\Big)^{-n} \prod_{i=1}^{n} \\ &\prod_{j=1}^{d} \exp\Big[-\tfrac{1}{2\sigma_j^2} z_{ij}^2 (\mathbb{1}_{\{z_{ij} \le 0\}} + \tau_j^{-2} \mathbb{1}_{\{z_{ij} > 0\}})\Big] \prod_{j=d+1}^{D} \exp\Big[-\tfrac{1}{2\sigma_j^2} z_{ij}^2\Big], \end{aligned}$$

where $c_1 = \left(\sqrt{\frac{2}{\pi}}\right)^d$. Now we take the log-likelihood function, i.e.

$$\begin{aligned}
&\ln(L(X;\mathrm{m},W,\sigma,\tau)) \\
&= \ln\left(\left(c_1|\det(W)|\right)^n\Big(\prod_{j=1}^{d}\sigma_j(1+\tau_j)\Big)^{-n}\right) + \sum_{i=1}^{n}\sum_{j=1}^{d}\Big[-\tfrac{1}{2\sigma_j^2}z_{ij}^2(\mathbb{1}_{\{z_{ij}\le 0\}} + \tau_j^{-2}\mathbb{1}_{\{z_{ij}>0\}})\Big] + \sum_{i=1}^{n}\sum_{j=d+1}^{D}\Big[-\tfrac{1}{2\sigma_j^2}z_{ij}^2\Big] \\
&= \ln\left(\left(c_1|\det(W)|\right)^n\Big(\prod_{j=1}^{d}\sigma_j(1+\tau_j)\Big)^{-n}\right) - \tfrac{1}{2}\sum_{j=1}^{d}\Big(\sigma_j^{-2}\sum_{i\in I_j}z_{ij}^2 + \tfrac{\sigma_j^{-2}}{\tau_j^2}\sum_{i\in I_j^c}z_{ij}^2\Big) - \tfrac{1}{2}\sum_{j=d+1}^{D}\sigma_j^{-2}\Big(\sum_{i\in I_j}z_{ij}^2 + \sum_{i\in I_j^c}z_{ij}^2\Big) \\
&= \ln\left(\left(c_1|\det(W)|\right)^n\Big(\prod_{j=1}^{d}\sigma_j(1+\tau_j)\Big)^{-n}\right) - \sum_{j=1}^{d}\tfrac{1}{2\sigma_j^2}\Big(s_{1j} + \tfrac{1}{\tau_j^2}s_{2j}\Big) - \sum_{j=d+1}^{D}\tfrac{1}{2\sigma_j^2}\Big(s_{1j} + s_{2j}\Big).
\end{aligned}$$

We fix m, $W$ and maximize the log-likelihood function over $\tau$ and $\sigma$. In such a case we have to solve the following system of equations

$$\begin{aligned}
\tfrac{\partial \ln(L(X;\mathrm{m},W,\sigma,\tau))}{\partial \sigma_j} &= -\tfrac{n}{\sigma_j} + \sigma_j^{-3}(s_{1j} + \tau_j^{-2}s_{2j}) + \sigma_j^{-3}(s_{1j} + s_{2j}) = 0, \\
\tfrac{\partial \ln(L(X;\mathrm{m},W,\sigma,\tau))}{\partial \tau_j} &= -\tfrac{n}{1+\tau_j} + \tfrac{s_{2j}}{\tau_j^3\sigma_j^2} = 0,
\end{aligned}$$

for $j = 1,\ldots,d$. By simple calculations we obtain the expressions for the estimators

$$\hat{\sigma}_j^2(\mathrm{m},W) = \tfrac{1}{n}s_{1j}^{2/3}g_j(\mathrm{m},W), \qquad \hat{\tau}_j(\mathrm{m},W) = \left(\frac{s_{2j}}{s_{1j}}\right)^{1/3}.$$

Substituting it into the log-likelihood function, we get

$$\begin{aligned}
\hat{L}(\mathrm{m},W) &= \left(\tfrac{2}{\pi}\right)^{\frac{dn}{2}}|\det(W)|^n \cdot \Big(\prod_{j=1}^{d}\tfrac{1}{\sqrt{n}}g_j(\mathrm{m},W)^{\frac{3}{2}}\Big)^{-n}e^{-\frac{dn}{2}} \cdot \Big(\prod_{j=d+1}^{D}\tfrac{1}{\sqrt{n}}(\tfrac{s_{1j}+s_{2j}}{n})^{\frac{3}{2}}\Big)^{-n} \\
&= \left(\tfrac{2n}{\pi e}\right)^{\frac{dn}{2}}\Big(\tfrac{1}{|\det(W)|^{\frac{2}{3}}}\prod_{j=1}^{d}g_j(\mathrm{m},W)\prod_{j=d+1}^{D}(s_{1j}+s_{2j})^{\frac{1}{3}}\Big)^{-\frac{3n}{2}}.
\end{aligned}$$

□

Thanks to the above theorem, instead of looking for the maximum of the likelihood function, it is enough to obtain the maximum of the simpler function (5) which depends on two parameters $\mathrm{m} \in \mathbb{R}^d$ and $W \in \mathcal{M}(\mathbb{R}^d)$

$$l(X;\mathrm{m},W) = \frac{1}{|\det(W)|^{\frac{2}{3}}}\prod_{j=1}^{d}g_j(\mathrm{m},W)\prod_{j=d+1}^{D}(s_{1j}+s_{2j})^{\frac{1}{3}} \tag{9}$$

where $\omega_j$ stands for the $j$-th column of matrix $W$. Consequently, maximization of (5) is equivalent to minimization of (6), see the following corollary.

**Corollary 5.1.** *Let $X \subset \mathbb{R}^d$, $\mathrm{m} \in \mathbb{R}^d$, $W \in \mathcal{M}(\mathbb{R}^d)$ be given, then*

$$\mathit{argmax}_{\mathrm{m},W}\hat{L}(X;\mathrm{m},W) = \operatorname*{argmin}_{\mathrm{m},W} l(X;\mathrm{m},W).$$

## 5.1 Gradient II

**Theorem 5.2.** *Let $X \subset \mathbb{R}^d$, $\mathrm{m} = (\mathrm{m}_1, \ldots, \mathrm{m}_d)^T \in \mathbb{R}^d$, $W = (\omega_{ij})_{1 \le i,j \le d}$ non-singular be given. Then $\nabla_{\mathrm{m}} \ln l(X; \mathrm{m}, W) = \left(\frac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_1}, \ldots, \frac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_d}\right)^T$, where*

$$\frac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_k} = \sum_{j=1}^{d} \frac{-1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}} \left( \frac{1}{3s_{1j}^{\frac{2}{3}}} \sum_{i \in I_j} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} + \frac{1}{3s_{2j}^{\frac{2}{3}}} \sum_{i \in I_j^c} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} \right) +$$
$$\sum_{j=d+1}^{D} \frac{-1}{3(s_{1j}+s_{2j})} \left( \sum_{i \in I_j} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} + \sum_{i \in I_j^c} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} \right).$$

*Moreover, $\nabla_W \ln l(X; \mathrm{m}, W) = \left[\frac{\partial \ln \tilde{l}(X;\mathrm{m},W)}{\partial \omega_{pk}}\right]_{1 \le p,k \le d}$, where*

$$\frac{\partial \ln l(X;\mathrm{m},W)}{\partial \omega_{pk}} = -\frac{2}{3}(\omega^{-1})_{pk}^T + \frac{1}{s_{1p}^{\frac{1}{3}}+s_{2p}^{\frac{1}{3}}} \left( \frac{1}{3} s_{1p}^{-\frac{2}{3}} \sum_{i \in I_p} 2\omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) \right.$$
$$\left. + \frac{1}{3} s_{2p}^{-\frac{2}{3}} \sum_{i \in I_p^c} 2\omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) \right) +$$
$$\frac{1}{3(s_{1j}+s_{2j})} \left( \sum_{i \in I_p} 2\omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) + \sum_{i \in I_p^c} 2\omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) \right).$$

*and*

$$s_{1j} = \sum_{i \in I_j} [\omega_j^T(\mathrm{x}_i - \mathrm{m})]^2, I_j = \{i = 1, \ldots, n \colon \omega_j^T(\mathrm{x}_i - \mathrm{m}) \le 0\},$$
$$s_{2j} = \sum_{i \in I_j^c} [\omega_j^T(\mathrm{x}_i - \mathrm{m})]^2, I_j^c = \{i = 1, \ldots, n \colon \omega_j^T(\mathrm{x}_i - \mathrm{m}) > 0\}.$$

*Proof of Theorem 5.2.* Let us start with the partial derivative of $\ln(l)$ with respect to m. We have

$$\frac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_k} = \sum_{j=1}^{d} \frac{\partial \ln(g_j(\mathrm{m},W))}{\partial \mathrm{m}_k} + \sum_{j=d+1}^{D} \frac{\partial \ln((s_{1j}+s_{2j})^{\frac{1}{3}})}{\partial \mathrm{m}_k} = \sum_{j=1}^{d} \frac{1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}} \frac{\partial (s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}})}{\partial \mathrm{m}_k} + \sum_{j=d+1}^{D} \frac{1}{(s_{1j}+s_{2j})^{\frac{1}{3}}} \frac{\partial ((s_{1j}+s_{2j})^{\frac{1}{3}})}{\partial \mathrm{m}_k} =$$
$$\sum_{j=1}^{d} \frac{1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}} \left( \frac{1}{3s_{1j}^{\frac{2}{3}}} \frac{\partial s_{1j}}{\partial \mathrm{m}_k} + \frac{1}{3s_{2j}^{\frac{2}{3}}} \frac{\partial s_{2j}}{\partial \mathrm{m}_k} \right) + \sum_{j=d+1}^{D} \frac{1}{(s_{1j}+s_{2j})^{\frac{1}{3}}} \frac{1}{3} \frac{1}{(s_{1j}+s_{2j})^{\frac{2}{3}}} \left( \frac{\partial s_{1j}}{\partial \mathrm{m}_k} + \frac{\partial s_{2j}}{\partial \mathrm{m}_k} \right).$$

Now, we need $\frac{\partial s_{1j}}{\partial \mathrm{m}_k}$ and $\frac{\partial s_{2j}}{\partial \mathrm{m}_k}$, therefore

$$\frac{\partial s_{1j}}{\partial \mathrm{m}_k} = \sum_{i \in I_j} \frac{\partial [\omega_j^T(\mathrm{x}_i - \mathrm{m})]^2}{\partial \mathrm{m}_k} = \sum_{i \in I_j} 2\omega_j^T(\mathrm{x}_i - \mathrm{m}) \frac{\partial \omega_j^T(\mathrm{x}_i - \mathrm{m})}{\partial \mathrm{m}_k} = \sum_{i \in I_j} -2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk}.$$

Analogously we get

$$\frac{\partial s_{2j}}{\partial \mathrm{m}_k} = \sum_{i \in I_j^c} -2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk}.$$

Hence

$$\frac{\partial \ln l}{\partial \mathrm{m}_k} = \sum_{j=1}^{d} \frac{-1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}} \left( \frac{1}{3s_{1j}^{\frac{2}{3}}} \sum_{i \in I_j} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} + \frac{1}{3s_{2j}^{\frac{2}{3}}} \sum_{i \in I_j^c} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} \right) +$$
$$\sum_{j=d+1}^{D} \frac{-1}{3(s_{1j}+s_{2j})} \left( \sum_{i \in I_j} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} + \sum_{i \in I_j^c} 2\omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} \right).$$

Now we calculate the partial derivative of $\ln l(X; \mathrm{m}, W)$ with respect to the matrix $W$. We have

$$\frac{\partial \ln l(X;\mathrm{m},W)}{\partial \omega_{pk}} = \frac{\partial \ln |\det(W)|^{-\frac{2}{3}}}{\partial \omega_{pk}} + \sum_{j=1}^{d} \frac{\partial \ln(g_j(\mathrm{m},W))}{\partial \omega_{pk}} + \sum_{j=d+1}^{D} \frac{\partial \ln((s_{1j}+s_{2j})^{\frac{1}{3}})}{\partial \omega_{pk}}.$$

To calculate the derivative of the determinant we use Jacobi's formula (see Lemma 4.1). Hence

$$\begin{aligned}&\frac{\partial \ln(\det(W)^{-\frac{2}{3}})}{\partial \omega_{pk}} = \det(W)^{\frac{2}{3}}\Big(-\frac{2}{3}\Big)\det(W)^{-\frac{5}{3}}\frac{\partial \det(W)}{\partial \omega_{pk}} = -\frac{2}{3}\det(W)^{-1}\mathrm{adj}^T(W)_{pk}\\ &= -\frac{2}{3}\frac{1}{\det(W)}\left[\det(W)(W^{-1})^T_{pk}\right] = -\frac{2}{3}(\omega^{-1})^T_{pk},\end{aligned}$$

where $(\omega^{-1})^T_{pk}$ is the element in the $p$-th row and $k$-th column of the matrix $(W^{-1})^T$. Now we calculate

$$\frac{\partial \ln(g_j(\mathrm{m},W))}{\partial \omega_{pk}} = \frac{1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}}\frac{\partial(s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}})}{\partial \omega_{pk}} = \frac{1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}}\left(\frac{1}{3s_{1j}^{\frac{2}{3}}}\frac{\partial s_{1j}}{\partial \omega_{pk}} + \frac{1}{3s_{2j}^{\frac{2}{3}}}\frac{\partial s_{2j}}{\partial \omega_{pk}}\right),$$

where

$$\begin{aligned}&\frac{\partial s_{1j}}{\partial \omega_{pk}} = \sum_{i\in I_j}\frac{\partial[\omega_j^T(\mathrm{x}_i-\mathrm{m})]^2}{\partial \omega_{pk}} = \sum_{i\in I_j} 2\omega_j^T(\mathrm{x}_i-\mathrm{m})\frac{\partial \omega_j^T(\mathrm{x}_i-\mathrm{m})}{\partial \omega_{pk}} =\\ &\begin{cases} 0, & \text{if } j\neq p\\ \sum\limits_{i\in I_p} 2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k), & \text{if } j=p\end{cases}\end{aligned}$$

and $\mathrm{x}_{ik}$ is the $k$-th element of the vector $\mathrm{x}_i$. Analogously we get

$$\frac{\partial s_{2j}}{\partial \omega_{pk}} = \begin{cases} 0, & \text{if } j\neq p\\ \sum\limits_{i\in I_p^c} 2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k), & \text{if } j=p.\end{cases}$$

Moreover,

$$\frac{\partial \ln((s_{1j}+s_{2j})^{\frac{1}{3}})}{\partial \omega_{pk}} = \frac{1}{(s_{1j}+s_{2j})^{\frac{1}{3}}}\frac{\partial((s_{1j}+s_{2j})^{\frac{1}{3}})}{\partial \omega_{pk}} = \frac{1}{(s_{1j}+s_{2j})^{\frac{1}{3}}}\frac{1}{3}\frac{1}{(s_{1j}+s_{2j})^{\frac{2}{3}}}\left(\frac{\partial s_{1j}}{\partial \omega_{pk}} + \frac{\partial s_{2j}}{\partial \omega_{pk}}\right),$$

Hence we obtain

$$\begin{aligned}&\frac{\partial \ln l}{\partial \omega_{pk}} = -\frac{2}{3}(\omega^{-1})^T_{pk} + \frac{1}{s_{1p}^{\frac{1}{3}}+s_{2p}^{\frac{1}{3}}}\Bigg(\frac{1}{3}s_{1p}^{-\frac{2}{3}}\sum_{i\in I_p} 2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k)\\ &+\frac{1}{3}s_{2p}^{-\frac{2}{3}}\sum_{i\in I_p^c} 2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k)\Bigg)+\\ &\frac{1}{3(s_{1j}+s_{2j})}\Bigg(\sum_{i\in I_p} 2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k) + \sum_{i\in I_p^c} 2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k)\Bigg).\end{aligned}$$

□

# References