Abstract

The contribution of this paper is twofold: First, we prove existence and uniqueness of the weighted maximum likelihood estimator of the multivariate Student-$t$ distribution and propose an efficient algorithm for its computation that we call generalized multivariate myriad filter (GMMF). Second, we use the GMMF in a nonlocal framework for the denoising of images corrupted by different kinds of noise. The resulting method is very flexible and can handle very heavy-tailed noise such as Cauchy noise, but also also Gaussian or wrapped Cauchy noise.

1. Introduction

Besides mean and median filter, myriad filter form an important class of nonlinear filters, in particular in robust signal and image processing. While in a multivariate setting, the mean filter can be defined componentwise, the generalization of the median to higher dimensions is not canonically, but often the geometric median is used, see, e.g., [26, 32, 31]. To the best of our knowledge, a multivariate myriad filter has not been considered yet, and in this paper we make a first attempt based on the multivariate Student-$t$ distribution.

In one dimension, mean, median as well as myriad filters can be derived as maximum likelihood (ML) estimators of the location parameter from a Gaussian, Laplacian respective Cauchy distribution. Concerning a multivariate myriad filter, instead of a multivariate Cauchy distribution we propose to start with the family of more general Student-$t$ distributions, which possesses an additional degrees of freedom parameter $\nu$ that allows to control the robustness of the resulting filter. While the Cauchy distribution is obtained as the special
The multivariate Student $t$-distribution is frequently used in statistics [8], whereas the multivariate Cauchy distribution is far less common and in contrast to the one-dimensional case usually not considered separately from the Student-$t$ distribution. The parameter(s) of a multivariate Student $t$-distribution are usually estimated via the maximum likelihood method in combination with the EM algorithm, since the resulting equations are not solvable in closed form. The EM algorithm has been first derived in [9], for an overview over other estimation methods for the multivariate Student $t$-distribution, in particular the EM algorithm and its variants, we refer to [18] and the references therein.

Recently, the Student-$t$ distribution and closely related Student-$t$ mixture models (SMM) have found interesting applications in different image processing tasks. For instance, in [30] it has been shown that Student-$t$ mixture models are superior to Gaussian mixture models for modeling image patches and the authors proposed an application in image compression. Further applications include robust image segmentation [1, 19, 27] as well as robust registration [6, 33]. In both cases, the SMM is estimated using the EM algorithm that has been derived in [22]. In this work we propose an application to robust denoising of images corrupted by different kinds of noise. The initial motivation for this work were the recent papers [17, 25] and [11] for Cauchy noise removal. In [17, 25] the authors proposed a variational method consisting of a data term that resembles the noise statistics and a total variation regularization term. Based on a maximum likelihood approach the authors of [11] introduced a generalized myriad filter which estimates both the location and the scale parameter of the Cauchy distribution. They used this filter in a nonlocal approach, where for each pixel of the image they chose as samples those pixels that have a similar neighborhood and replaced the initial pixel by its filtered version. Such a pixelwise treatment assumes the pixels of an image to be independent, which is in practice a rather unrealistic assumption; in fact, in natural images they are usually locally highly correlated. Taking the local dependence structure into account may improve the results of image restoration methods. For instance for denoising images corrupted by additive Gaussian noise this led to the state-of-the-art algorithm of Lebrun et al. [12]. In case of a myriad filtering approach designed to denoise images corrupted by additive Cauchy noise this would require to define a multivariate myriad filter. In this work, we derive a multivariate generalized myriad filter (MGMF) based on ML estimation for the family of Student-$t$ distributions, of which the Cauchy distribution forms a special case.

The paper is organized as follows: In Section 2, we introduce the Student-$t$ distribution. Then, in Section 3, we prove existence and uniqueness of weighted maximum likelihood estimators for its parameters. We propose an efficient algorithm for computing the ML estimates in Section 4, prove its convergence and compare it to the classical EM algorithm. In Section 5 we illustrate how the developed algorithm can be applied in the context of
nonlocal (robust) image denoising. Conclusions are given and directions of future research are addressed in Section 6.

2. Multivariate Student-\(t\) Distribution

In this section, we introduce the multivariate Student-\(t\) distribution and collect some of its properties. The probability density function of the \(d\)-dimensional Student \(t\)-distribution \(T_\nu(\mu, \Sigma)\) with \(\nu > 0\) degrees of freedom is given by

\[
f_\nu(x|\mu, \Sigma) := \frac{\Gamma \left( \frac{\nu + d}{2} \right)}{\Gamma \left( \frac{\nu}{2} \right) (\pi \nu)^{\frac{d}{2}} |\Sigma|^{\frac{d}{2}}} \frac{1}{\left[ 1 + \frac{1}{\nu} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]^\frac{\nu + d}{2}},
\]

where \(\Gamma(x) := \int_0^\infty t^{x-1}e^{-t} \, dt\) denotes the Gamma function. The smaller the value of \(\nu\) for fixed location \(\mu\) and scatter matrix \(\Sigma\), the heavier are the tails of the \(T_\nu(\mu, \Sigma)\) distribution, that means, the more robust is the estimation towards outliers. The limiting case \(\nu = 0\) is related to the projected normal distribution on the sphere \(S^{d-1}\), and in particular for \(d = 2\) to the wrapped Cauchy distribution, see also Section 5. Figure 1 illustrates this behavior for the one-dimensional standard Student-\(t\) distribution. For \(\nu \to \infty\), the Student-\(t\) distribution \(T_\nu(\mu, \Sigma)\) converges to the normal distribution \(\mathcal{N}(\mu, \Sigma)\).

![Figure 1: Standard Student-\(t\) distribution \(T_\nu(0, 1)\) for different values of \(\nu\) in comparison with the standard normal distribution \(\mathcal{N}(0, 1)\).](image)

The expectation of the Student-\(t\) distribution is \(E(X) = \mu\) for \(\nu > 1\) and the covariance matrix is given by \(\text{Cov}(X) = \frac{\nu}{\nu - 2} \Sigma\) for \(\nu > 2\), otherwise the quantities are undefined. As the normal distribution, the Student-\(t\) distribution belongs to the class of elliptical distributions. Some important properties that are needed later on are summarized in the next theorem [8]. In the following, we denote by \(\text{SPD}(d)\) the cone of symmetric positive definite matrices.
Theorem 2.1. \[(i)\] Let \(\mu \in \mathbb{R}^d\) and \(\Sigma \in \text{SPD}(d)\). Further, let \(Y \sim \mathcal{N}(0, \Sigma)\) and \(T \sim \chi^2_\nu\) be independent, where \(\chi^2_\nu\) is the \(\chi^2\)-distribution with \(\nu\) degrees of freedom. Then 
\[
X = \mu + \frac{Y}{\sqrt{\frac{T}{\nu}}} \sim T_\nu(\mu, \Sigma).
\]

\[(ii)\] Let \(X \sim T_\nu(\mu, \Sigma), A \in \mathbb{R}^{d \times d}\) be an invertible matrix and \(b \in \mathbb{R}^d\). Then 
\[
AX + b \sim T_\nu(A\mu + b, A \Sigma A^T).
\]

3. Weighted Maximum Likelihood Estimators

In this section, we establish the weighted log-\(\text{ML}\) function and prove the existence and uniqueness of corresponding estimators for the scatter matrix \(\Sigma\) and both the location parameter and the scatter matrix. In [16], sufficient conditions for the existence and uniqueness of a joint minimizer of \(L_\nu\) have been established for uniform weights and \(M\)-estimators whose cost function fulfills certain properties. Similar results can be found in [7]. However, the results in [7, 16] cannot be applied if multiple samples occur. Restricting our attention to the multivariate Student \(t\)-distribution we can give direct proofs of the existence and uniqueness which are simpler than the argumentation in [7, 16]. Moreover, considering different weights in the log-likelihood function we can also allow multiple samples.

3.1. Weighted ML Function

The likelihood function of the \(T_\nu(\mu, \Sigma)\) distribution is given by
\[
\mathcal{L}(\nu, \mu, \Sigma|x_1, \ldots, x_n) := \frac{\Gamma\left(\frac{d + \nu}{2}\right)^n}{\Gamma\left(\nu/2\right)^{\frac{d}{\nu}} |\Sigma|^{\frac{d}{2}} \prod_{i=1}^n \left[1 + \frac{1}{\nu}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right]^{\frac{d+\nu}{2}}}.
\]

and the log-likelihood function by
\[
\ell(\nu, \mu, \Sigma|x_1, \ldots, x_n) := n \log \left(\Gamma\left(\frac{d + \nu}{2}\right)\right) - n \log \left(\Gamma\left(\nu/2\right)\right) - \frac{nd}{2} \log(\pi \nu) - \frac{n}{2} \log(|\Sigma|) - \frac{d + \nu}{2} \sum_{i=1}^n \log \left(1 + \frac{1}{\nu}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right).
\]

Ignoring constants, minimizing \(\ell\) is equivalent to minimizing its negative, weighted version
\[
L(\nu, \mu, \Sigma|x_1, \ldots, x_n) := -2 \log \left(\Gamma\left(\frac{d + \nu}{2}\right)\right) + 2 \log \left(\Gamma\left(\nu/2\right)\right) - \nu \log(\nu) + (d + \nu) \sum_{i=1}^n w_i \log \left(1 + \frac{1}{\nu}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right) + \log(|\Sigma|).
\]
for uniform weights \( w_i = \frac{1}{n} \). We want to allow for different weighting of the summands by introducing weights in

\[
\Delta_n := \left\{ w = (w_1, \ldots, w_n) \in \mathbb{R}_{>0}^n : \sum_{i=1}^n w_i = 1 \right\}.
\]

We omit the dependence of \( L \) on the samples \( x_1, \ldots, x_n \) and indicate by subscripting if one or more of the parameters are fixed, for instance \( L_\nu(\mu, \Sigma) \) means that \( \nu \) is assumed to be known. Furthermore, we denote with \( \delta(x; \mu, \Sigma) = (x - \mu)^T \Sigma^{-1} (x - \mu) \) the Mahalanobis distance between \( x \) and a distribution with mean \( \mu \) and covariance matrix \( \Sigma \) and set \( \delta_i = \delta(x_i; \mu, \Sigma) \) if \( \mu \) and \( \Sigma \) are unambiguous from the context. The derivatives of \( L_\nu \) with respect to \( \mu \) and \( \Sigma \) are given by

\[
\begin{align*}
\frac{\partial L_\nu}{\partial \mu}(\mu, \Sigma) &= -2(d + \nu) \sum_{i=1}^n w_i \frac{\Sigma^{-1}(x_i - \mu)}{\nu + \delta_i}, \\
\frac{\partial L_\nu}{\partial \Sigma}(\mu, \Sigma) &= -(d + \nu) \sum_{i=1}^n \frac{\Sigma^{-1}(x_i - \mu)(x_i - \mu)^T \Sigma^{-T}}{\nu + \delta_i} + \Sigma^{-1},
\end{align*}
\]

where we used the following relations see, for instance, [23]:

\[
\frac{\partial \log(|X|)}{\partial X} = X^{-1}, \quad \frac{\partial a^T X^{-1} b}{\partial X} = -(X^{-T}) ab^T (X^{-T}).
\]

Setting the derivatives to zero results in the equations

\[
\begin{align*}
0 &= \sum_{i=1}^n w_i \frac{x_i - \mu}{\nu + \delta_i}, \\
I &= (d + \nu) \sum_{i=1}^n w_i \frac{\Sigma^{-\frac{1}{2}}(x_i - \mu)(x_i - \mu)^T \Sigma^{-\frac{1}{2}}}{\nu + \delta_i}.
\end{align*}
\]

Computing the trace of both sides of (2) and using the linearity and permutation invariance of the trace operator we see

\[
d = \text{tr}(I) = (d + \nu) \sum_{i=1}^n w_i \frac{\text{tr}(\Sigma^{-\frac{1}{2}}(x_i - \mu)(x_i - \mu)^T \Sigma^{-\frac{1}{2}})}{\nu + \langle x_i - \mu \rangle^T \Sigma^{-1}(x_i - \mu)} = (d + \nu) \sum_{i=1}^n w_i \frac{\delta_i}{\nu + \delta_i},
\]
yielding for any critical point \((\mu, \Sigma)\) of \(L\) the necessary condition
\[
(d + \nu) \sum_{i=1}^{n} w_i \frac{1}{\nu + \delta_i} = 1. \tag{3}
\]

For a critical point \((\hat{\mu}, \hat{\Sigma})\) we reformulate (1) and (2) as fixed-point equations
\[
\hat{\mu} = \sum_{i=1}^{n} w_i \frac{1}{\nu + (x_i - \mu) (x_i - \mu)^T} \nu + (x_i - \hat{\mu})(x_i - \hat{\mu})^T, \tag{4}
\]
\[
\hat{\Sigma} = \sum_{i=1}^{n} w_i \frac{(x_i - \hat{\mu})(x_i - \hat{\mu})^T}{\nu + (x_i - \hat{\mu})(x_i - \hat{\mu})^T} = \sum_{i=1}^{n} w_i \frac{(x_i - \hat{\mu})(x_i - \hat{\mu})^T}{\nu + \delta_i}. \tag{5}
\]

### 3.2. Estimation of Scatter

First, we consider the estimation of only the scatter matrix \(\Sigma\), where the location parameter \(\mu\) is known and fixed, w.l.o.g. \(\mu = 0\). If \(\mu \neq 0\), we might transform the samples to \(y_i = x_i - \mu\), \(i = 1, \ldots, n\). For abbreviation, we set
\[
L_{\nu,0}(\Sigma) = L_{\nu}(0, \Sigma) := (d + \nu) \sum_{i=1}^{n} w_i \log \left( \nu + x_i^T \Sigma^{-1} x_i \right) + \log(|\Sigma|)
\]
and make the following assumption on the samples and weights:

**Assumption 3.1.** Let \(n \geq d\).

(i) Let \(x_1, \ldots, x_n \in \mathbb{R}^d\), be a set of samples such that any subset of samples \(d\) is linearly independent.

(ii) Let \(w \in \Delta_n\) fulfill \((d - 1) w_{\text{max}} < \frac{\nu + d - 1}{\nu + d}\), where \(\nu \geq 0\).

The linear independence assumption holds \(\lambda^d\)-a.s. when sampling from a continuous distribution. The interpretation behind the constraints is that the mass of the (empirical) distribution determined by \(x_1, \ldots, x_n\) is not allowed to be concentrated in lower dimensional subspaces, which would cause the resulting distribution to be degenerated.

**Lemma 3.2.** Let \(x_i \in \mathbb{R}^d\), \(w_i \in \mathbb{R}\), \(i = 1, \ldots, n\), fulfill Assumption 3.1. Further, let \(V \subset \mathbb{R}^d\) be a linear subspace with \(0 \leq \dim(V) \leq d - 1\) and \(\mathcal{I}_V := \{i \in \{1, \ldots, n\} : x_i \in V\}\). Then it holds
\[
\sum_{i \in \mathcal{I}_V} w_i < \frac{\nu + \dim(V)}{\nu + d}. \tag{6}
\]
Proof. First, if $V = \{0\}$, we have $I = \emptyset$, so the statement holds true. Next, let $k = \dim(V) \leq d - 1$. Then, according to the linear independence assumption on the samples $x_1, \ldots, x_n$, it holds $|I_V| \leq k$. Using $w_{\max} < \frac{\nu + d - 1}{(d - 1)(\nu + d)}$, we obtain

$$
\sum_{i \in I_V} w_i < k \frac{\nu + d - 1}{(d - 1)(\nu + d)} = \frac{k}{d - 1} \frac{\nu + k}{\nu + d} \leq \frac{\nu + \dim(V)}{\nu + d}.
$$

Remark 3.3. If (6) holds true, then the linear independence assumption implies $n \geq d$. Indeed, assume that $n < d$ and let $V$ be a linear subspace containing all the samples. Then $\dim(V) \leq d - 1$, but

$$
1 = \sum_{i \in I_V} w_i \leq \frac{\nu + \dim(V)}{d + \nu} \leq \frac{\nu + d - 1}{d + \nu} < 1,
$$

which gives a contradiction.

Next, we show the existence of a minimizer of $L_{\nu,0}$. To this aim, we denote by $\lambda_i(A)$ the $i$-th greatest eigenvalue, $i = 1, \ldots, d$ of a matrix $A \in \text{SPD}(d)$.

Theorem 3.4 (Existence of Scatter). Let $x_i \in \mathbb{R}^d$, $w_i > 0$, $i = 1, \ldots, n$, fulfill Assumption 3.1 and let $\nu > 0$. Then it holds

$$
\argmin_{\Sigma \in \text{SPD}(d)} L_{\nu,0}(\Sigma) \neq \emptyset
$$

and any $\hat{\Sigma} \in \argmin_{\Sigma \in \text{SPD}(d)} L_{\nu,0}(\Sigma)$ is a critical point of $L_{\nu,0}$.

Proof. We show that $L_{\nu,0}(\Sigma)$ tends to infinity if $\Sigma$ approaches the boundary of $\text{SPD}(d)$. Let $\{\Sigma_r\}_{r \in \mathbb{N}} \subseteq \text{SPD}(d)$ be a sequence in $\text{SPD}(d)$, let $\lambda_1 \geq \ldots \geq \lambda_d > 0$ denote the eigenvalues of $\Sigma_r$ and let $e_1, \ldots, e_d$ be corresponding orthonormal eigenvectors of $\Sigma_r$. Observe that if $\Sigma_r$ approaches the boundary of $\text{SPD}(d)$, then $\lambda_1 \xrightarrow{r \to \infty} +\infty$ or $\lambda_d \xrightarrow{r \to \infty} 0$. We prove that $L_{\nu,0}(\Sigma_r) \xrightarrow{r \to \infty} +\infty$ if one of the following situations is met:

i) $\lambda_1 \xrightarrow{r \to \infty} +\infty$ and $\lambda_d \geq c > 0$, $r \in \mathbb{N}$,

ii) $\lambda_d \xrightarrow{r \to \infty} 0$.

First, we have

$$
L_{\nu,0}(\Sigma_r) = (d + \nu) \sum_{i=1}^{n} w_i \log \left( \nu + x_i^T \Sigma_r^{-1} x_i \right) + \log(|\Sigma_r|)
$$

$$
= (d + \nu) \sum_{i=1}^{n} w_i \log \left( \nu + x_i^T \Sigma_r^{-1} x_i \right) + \sum_{j=1}^{d} \log(\lambda_{jr}).
$$

7
Now, in case (i), the first sum is bounded below, while the second one tends to infinity, so that \( L_{\nu,0}(\Sigma_r) \xrightarrow{r \to \infty} +\infty \).

The more involved analysis of case (ii) is inspired by the proof of the Courant-Fischer-Min-Max principle. Let \( 0 \leq p \leq d-1 \) such that \( \lambda_{1r} \geq \ldots \geq \lambda_{pr} \geq c > 0 \) for all \( r \in \mathbb{N} \) and \( \lambda_{p+1r} \geq \ldots \geq \lambda_{dr} \xrightarrow{r \to \infty} 0 \) (if \( p = 0 \), all \( \lambda_{jr} \) tend to zero). Since \( S^d \) is compact, there exist subsequences (w.l.o.g. again denoted by \( e_{jr} \)) such that \( \lim_{r \to \infty} e_{jr} = e_j \) for \( j = 1, \ldots, d \). We introduce the following spaces and sets: For \( k = 1, \ldots, d \), let \( S_k = \text{span}\{e_1, \ldots, e_k\} \), in particular \( S_d = \mathbb{R}^d \) and \( \dim(S_k) = k \). Set \( S_0 = \{0\} \) and define

\[
W_k = S_k \setminus S_{k-1} = \{ y \in \mathbb{R}^d : \langle y, e_k \rangle \neq 0, \langle y, e_l \rangle = 0 \text{ for } l = k + 1, \ldots, d \}, \quad k = 1, \ldots, d.
\]

Let

\[
\tilde{I}_k = \{ i \in \{1, \ldots, n\} : x_i \in S_k \} \quad \text{and} \quad I_k = \{ i \in \{1, \ldots, n\} : x_i \in W_k \}.
\]

Using \( S_k = W_k \cup S_{k-1} \) we obtain \( \tilde{I}_k = I_k \cup \tilde{I}_{k-1} \) for \( k = 1, \ldots, d \). According to Assumption 3.1 it holds \( |I_k| \leq |\tilde{I}_k| \leq \dim(S_k) = k \) for \( k = 1, \ldots, d-1 \).

For \( y \in W_k \) it holds

\[
\liminf_{r \to \infty} y^T \Sigma_r^{-1} y \lambda_{kr}^{-1} \geq \langle y, e_k \rangle^2 > 0
\]

and for \( r \) sufficiently large

\[
y^T \Sigma_r^{-1} y = \sum_{j=1}^d \frac{1}{\lambda_{jr}} \langle y, e_{jr} \rangle^2 \geq \frac{1}{\lambda_{kr}} \langle y, e_{kr} \rangle^2 > 0.
\]

As a consequence, for \( y \in W_k \) we have

\[
\liminf_{r \to \infty} \frac{y^T \Sigma_r^{-1} y}{\lambda_{kr}} \geq \langle y, e_k \rangle^2 > 0. \quad (7)
\]

Introducing the functions

\[
L_j(\Sigma_r) = (d + \nu) \sum_{i \in I_j} w_i \log \left( \nu + x_i^T \Sigma_r^{-1} x_i \right) + \log(\lambda_{jr}),
\]

we can write

\[
L_{\nu,0}(\Sigma_r) = (d + \nu) \sum_{i=1}^n w_i \log(\nu + x_i^T \Sigma_r^{-1} x_i) + \sum_{j=1}^d \log(\lambda_{jr}) = \sum_{j=1}^d L_j(\Sigma_r).
\]

8
Next, we show that for $k \geq p + 1$ it holds

\[ \sum_{j=k}^{d} L_j(\Sigma_r) \geq \sum_{j=k}^{d} c_{jr} + \left( (\nu + k - 1) - (d + \nu) \sum_{i \in \tilde{I}_{k-1}} w_i \right) \log(\lambda_{kr}^{-1}), \]

where $c_{jr} = (d + \nu) \sum_{i \in I_j} w_i \log \left( \frac{\nu + x_i^T \Sigma_r^{-1} x_i}{\lambda_{jr}} \right)$. We write

\[ \sum_{j=k}^{d} L_j(\Sigma_r) = \sum_{j=k}^{d} (d + \nu) \sum_{i \in I_j} w_i \log \left( \frac{\nu + x_i^T \Sigma_r^{-1} x_i}{\lambda_{jr}} \right) + \log(\lambda_{jr}^{-1}) \]

\[ = \left( (d + \nu) \sum_{i \in I_j} w_i - 1 \right) \log(\lambda_{jr}^{-1}) + \left( d + \nu \right) \sum_{i \in I_j} w_i. \]

Thus, it suffices to prove

\[ \sum_{j=k}^{d} \left( (d + \nu) \sum_{i \in I_j} w_i - 1 \right) \log(\lambda_{jr}^{-1}) \geq \left( (\nu + k - 1) - (d + \nu) \sum_{i \in \tilde{I}_{k-1}} w_i \right) \log(\lambda_{kr}^{-1}). \] (8)

Inequality (8) is shown inductively over $k = d, \ldots, 1$. Based on the relation $\tilde{I}_k = I_k \cup \tilde{I}_{k-1}$ we write

\[ \sum_{i \in I_k} w_i - \sum_{i \in \tilde{I}_{k-1}} w_i = \sum_{i \in \tilde{I}_{k-1}} w_i. \] (9)

For the induction basis $k = d$, it suffices to show

\[ (d + \nu) \sum_{i \in I_d} w_i - 1 \geq (\nu + d - 1) - (d + \nu) \sum_{i \in \tilde{I}_{d-1}} w_i. \]

This follows directly from (9), since

\[ (d + \nu) \sum_{i \in I_d} w_i - 1 = (d + \nu) \left( \sum_{i \in I_d} w_i - \sum_{i \in \tilde{I}_{d-1}} w_i \right) - 1 \]

\[ = (d + \nu) \left( 1 - \sum_{i \in \tilde{I}_{d-1}} w_i \right) - 1 \]

\[ = (\nu + d - 1) - (d + \nu) \sum_{i \in \tilde{I}_{d-1}} w_i. \]
Now, assume that (8) holds for some \( k + 1 \) with \( d \geq k + 1 > p + 1 \), that is

\[
\sum_{j=k+1}^{d} \left( (d + \nu) \sum_{i \in I_j} w_i - 1 \right) \log(\lambda^{-1}_{jr}) \geq (d + \nu) \left( \frac{\nu + k}{d + \nu} - \sum_{i \in I_k} w_i \right) \log(\lambda^{-1}_{k+1r}).
\]

We show that it holds for \( k \) as well. We split the sum and estimate

\[
\sum_{j=k}^{d} \left( (d + \nu) \sum_{i \in I_j} w_i - 1 \right) \log(\lambda^{-1}_{jr})
= \sum_{j=k+1}^{d} \left( (d + \nu) \sum_{i \in I_j} w_i - 1 \right) \log(\lambda^{-1}_{jr}) + \left( (d + \nu) \sum_{i \in I_k} w_i - 1 \right) \log(\lambda^{-1}_{kr})
\geq (d + \nu) \left( \frac{\nu + k}{d + \nu} - \sum_{i \in I_k} w_i \right) \log(\lambda^{-1}_{kr}) + (d + \nu) \sum_{i \in I_k} w_i - 1 \log(\lambda^{-1}_{kr}).
\]

Since \( \lambda^{-1}_{k+1r} \geq \lambda^{-1}_{kr} \) we obtain with Lemma 3.2 further

\[
\sum_{j=k}^{d} \left( (d + \nu) \sum_{i \in I_j} w_i - 1 \right) \log(\lambda^{-1}_{jr})
\geq (d + \nu) \left( \frac{\nu + k}{d + \nu} - \sum_{i \in I_k} w_i \right) \log(\lambda^{-1}_{kr}) + \left( (d + \nu) \sum_{i \in I_k} w_i - 1 \right) \log(\lambda^{-1}_{kr})
= \left( \frac{\nu + k}{d + \nu} - \left( 1 + (d + \nu) \sum_{i \in I_k} w_i - (d + \nu) \sum_{i \in I_k} w_i \right) \right) \log(\lambda^{-1}_{kr})
= \left( \frac{\nu + k}{d + \nu} - \left( 1 + (d + \nu) \sum_{i \in I_{k-1}} w_i \right) \right) \log(\lambda^{-1}_{kr})
= \left( (\nu + k - 1) - (d + \nu) \sum_{i \in I_{k-1}} w_i \right) \log(\lambda^{-1}_{kr}).
\]

Finally, we have

\[
L_{\nu,0}(\Sigma_r) = \sum_{j=1}^{d} L_j(\Sigma_r) = \sum_{j=1}^{p} L_j(\Sigma_r) + \sum_{j=p+1}^{d} L_j(\Sigma_r)
\geq \sum_{j=1}^{p} L_j(\Sigma_r) + \sum_{j=p+1}^{d} c_{jr} + \left( (\nu + p) - (d + \nu) \sum_{i \in I_p} w_i \right) \log(\lambda^{-1}_{p+1r}).
\]
By definition of \( p \) and using (7), the first two sums are bounded below for all \( r \in \mathbb{N} \), while with the help of Lemma 3.2 we see that the factor in front of the logarithm is positive, so that the second sum tends to \( +\infty \) for \( r \to \infty \), which yields \( L_{\nu,0}(\Sigma_r) \to +\infty \).

In summary, the minimum of \( L_{\nu,0} \) is attained in \( \text{SPD}(d) \), and since \( L_{\nu,0} \) is continuously differentiable, it is necessarily a critical point fulfilling (2).

The end of the proof of Theorem 3.4 reveals that the condition on the weights stated in (6) is sufficient for existence. The next lemma shows that a strong inequality is necessary.

**Lemma 3.5.** Let \( x_i \in \mathbb{R}^d, w_i > 0, i = 1, \ldots, n \), fulfill Assumption 3.1 and assume there exists a critical point \( \hat{\Sigma} \) of \( L_{\nu,0} \). Then, for all linear subspaces \( V \subseteq \mathbb{R}^d \) with \( 0 \leq \dim(V) \leq d-1 \) it holds

\[
\sum_{i \in I_V} w_i \leq \frac{\nu + \dim(V)}{d + \nu},
\]

where \( I_V = \{ i \in \{1, \ldots, n\} : x_i \in V \} \).

**Proof.** Without loss of generality, we might assume that \( \hat{\Sigma} = I \), otherwise we might transform the samples to \( y_i = R^{-1} x_i \), where \( RR^T = \hat{\Sigma} \). The idea of the proof is to project the samples onto the orthogonal complement of \( V \). More precisely, let \( k = \dim(V) \) and choose an orthonormal basis \( v_1, \ldots, v_k \) of \( V \) such that \( V = \text{span}(v_1, \ldots, v_k) \). Set \( W = (v_1, \ldots, v_k) \) so that \( P = WW^T \) is the orthogonal projection onto \( V \). Now, for \( \Sigma = I \) and \( \mu = 0 \), equation (2) reads as

\[
(d + \nu) \sum_{i=1}^{n} w_i \frac{x_i x_i^T}{\nu + x_i^T x_i} = I.
\]

Multiplying both sides with \( I - P \) and taking the trace afterwards yields

\[
d - k = (d + \nu) \sum_{i=1}^{n} w_i \frac{x_i^T (I - P) x_i}{\nu + x_i^T x_i}.
\]

We split the sum into a sum over \( i \in I_V \) and \( i \in I_V^c = \{1, \ldots, n\} \setminus I_V \) and get

\[
d - k = (d + \nu) \sum_{i \in I_V} w_i \frac{x_i^T (I - P) x_i}{\nu + x_i^T x_i} + (d + \nu) \sum_{i \in I_V^c} w_i \frac{x_i^T (I - P) x_i}{\nu + x_i^T x_i}
\]

\[
= (d + \nu) \sum_{i \in I_V^c} w_i \frac{x_i^T (I - P) x_i}{\nu + x_i^T x_i}.
\]

With \( x_i^T (I - P) x_i \leq x_i^T x_i \) we obtain further

\[
d - k = (d + \nu) \sum_{i \in I_V^c} w_i \frac{x_i^T (I - P) x_i}{\nu + x_i^T x_i} \leq (d + \nu) \sum_{i \in I_V^c} w_i \frac{x_i^T x_i}{\nu + x_i^T x_i}
\]

11
\[
= (d + \nu) \left( \sum_{i \in I} w_i - \sum_{i \in I} w_i \frac{\nu}{\nu + x_i^T x_i} \right) \leq (d + \nu) \sum_{i \in I} w_i \\
= (d + \nu) \left( 1 - \sum_{i \in I} w_i \right) \\
= d + \nu - (d + \nu) \sum_{i \in I} w_i.
\]

Rearranging yields
\[
(d + \nu) \sum_{i \in I} w_i \leq \nu + k
\]
and finally
\[
\sum_{i \in I} w_i \leq \frac{\nu + k}{d + \nu} = \frac{\nu + \dim(V)}{d + \nu}.
\]

Now, we turn to the question whether the minimizer of \( L_{\nu,0} \) is unique. For an alternative proof based on the Mountain pass theorem see Appendix A.1.

**Theorem 3.6 (Uniqueness of Scatter).** Let \( x_i \in \mathbb{R}^d \) and \( w_i \in \mathbb{R} \), \( i = 1, \ldots, n \) fulfill Assumption 3.1. Then there exists a unique critical point \( \Sigma \in \text{SPD}(d) \) of \( L_{\nu,0} \).

**Proof.** Let \( \Sigma_1 \) and \( \Sigma_2 \) be two matrices fulfilling (2). We might w.l.o.g. assume that \( \Sigma_1 = I \), since otherwise we transform the data to \( y_i = \Sigma_1^{-\frac{1}{2}} x_i \). We set \( \lambda_1 = \lambda_1(\Sigma_2) \) and assume that \( \lambda_1 > 1 = \lambda_1(I) \). By (2) it holds
\[
\hat{\Sigma}_2 = (d + \nu) \sum_{i=1}^n w_i \frac{x_i x_i^T}{\nu + x_i^T \Sigma_2^{-1} x_i}. 
\] (10)

Using the Courant-Fisher min-max principle we have
\[
x_i^T \hat{\Sigma}_2^{-1} x_i \geq \lambda_1^{-1} x_i^T x_i,
\]
so that with \( \lambda_1 > 1 \) we estimate
\[
\frac{d + \nu}{\nu + x_i^T \Sigma_2^{-1} x_i} \leq \frac{d + \nu}{\nu + \lambda_1^{-1} x_i^T x_i} = \lambda_1 \frac{d + \nu}{\nu + x_i^T x_i} < \lambda_1 \frac{d + \nu}{\nu + x_i^T x_i}, \quad i = 1, \ldots, n.
\]

Inserting in (10) yields
\[
\hat{\Sigma}_2 = (d + \nu) \sum_{i=1}^n w_i \frac{x_i x_i^T}{\nu + x_i^T \Sigma_2^{-1} x_i} < \lambda_1 (d + \nu) \sum_{i=1}^n w_i \frac{x_i x_i^T}{\nu + x_i^T x_i} = \lambda_1 I.
\]
This is a contradiction to $\lambda_1(\hat{\Sigma}_2) > 1 = \lambda_1(I)$, and consequently we have $\lambda_1 \leq 1$. Similarly one sees $\lambda_d(\hat{\Sigma}_2) \geq 1$, so that finally $\hat{\Sigma}_2 = I = \hat{\Sigma}_1$. 

Finally, we are interested in the limiting case $\nu = 0$. As the next proposition shows, the solution of (2) will no longer be unique. However, there exists a unique solution $\Sigma_0$ with trace $d$, and all other solutions are of the form $\alpha \Sigma_0$ with $\alpha > 0$.

**Proposition 3.7.** Let $x_1, \ldots, x_n$ fulfill Assumption 3.1. Then, it holds

$$
\sum_{i=1}^{n} w_i \frac{\Sigma^{-\frac{1}{2}} x_i x_i^T \Sigma^{-\frac{1}{2}}}{x_i^T \Sigma^{-1} x_i} = \sum_{i=1}^{n} w_i \frac{S^{-\frac{1}{2}} x_i x_i^T S^{-\frac{1}{2}}}{x_i^T \Sigma^{-1} x_i}
$$

if and only if $S = \alpha \Sigma$ for some $\alpha > 0$.

**Proof.** Clearly, if $S = \alpha \Sigma$ then (11) holds true. To show the reverse direction, we may as in the proof of Theorem 3.6 w.l.o.g. assume that $\Sigma = I$. Let $\lambda_1 = \lambda_1(S^{-1})$ denote the largest eigenvalue of $S^{-1}$ with multiplicity $k$ and let $e_1, \ldots, e_k$ be corresponding orthonormal eigenvectors. Further, let

$$
P = \sum_{i=1}^{k} e_i e_i^T
$$

be the associated eigenprojector. The equality (11) means for $\Sigma = I$ that

$$
\sum_{i=1}^{n} w_i \frac{x_i x_i^T}{x_i^T x_i} = \sum_{i=1}^{n} w_i \frac{S^{-\frac{1}{2}} x_i x_i^T S^{-\frac{1}{2}}}{x_i^T S^{-1} x_i}.
$$

Multiplying both sides with $P$ and taking the trace gives

$$
\sum_{i=1}^{n} w_i \frac{x_i^T P x_i}{x_i^T x_i} = \lambda_1 \sum_{i=1}^{n} w_i \frac{x_i^T P x_i}{x_i^T S^{-1} x_i}.
$$

Since $\frac{1}{\lambda_1} x_i^T S^{-1} x_i \leq x_i^T x_i$ with equality only if $P x_i = x_i$, this implies $P x_i = 0$ or $P x_i = x_i$ for all $i = 1, \ldots, n$. This contradicts Assumption 3.1(i) unless $P = I$, that means, $S = \lambda_1 I$. 

**Lemma 3.8.** Let $x_1, \ldots, x_n$ fulfill Assumption 3.1(i) for $\nu = 0$. Then, it holds

$$
\arg\min_{\Sigma \in \text{SPD}(d)} L_{\nu,0}(\Sigma) \neq \emptyset.
$$

**Proof.** First, according to Proposition 3.7 we can restrict the domain of $L_{\nu,0}$ to the bounded set

$$
D_0 = \{\Sigma \in \text{SPD}(d) : \text{tr}(\Sigma) = d\} \subseteq \text{Sym}(d).
$$
Let \( \{\Sigma_r\}_{r \in \mathbb{N}} \subseteq D_0 \) be a sequence in \( D_0 \) with \( \lim_{r \to \infty} \Sigma_r = \Sigma \in \partial D_0 \), we will show that \( L_{\nu,0}(\Sigma_r) \to \infty \). Let \( \lambda_1 \geq \ldots \geq \lambda_{dr} > 0 \) denote the eigenvalues of \( \Sigma_r \). Further, let \( \lambda_1 \geq \ldots \geq \lambda_p > 0, \lambda_{p+1} = \ldots = \lambda_d = 0 \) be the eigenvalues of \( \Sigma \), where \( p = \text{rank}(\Sigma) < d \). Note that \( p \geq 1 \) since \( \text{tr}(\Sigma) = d \). Now, the same argumentation as in the proof of case (ii) of Theorem 3.4 yields

\[
L_{\nu,0}(\Sigma_r) = \sum_{j=1}^{d} L_j(\Sigma_r) = \sum_{j=1}^{p} L_j(\Sigma_r) + \sum_{j=p+1}^{d} L_j(\Sigma_r) \\
\geq \sum_{j=1}^{p} L_j(\Sigma_r) + \sum_{j=p+1}^{d} c_{jr} + \left( p - d \sum_{i \in I_p} \right) \log(\lambda_{p+1}^{-1} r),
\]

which tends to \( +\infty \) for \( r \to \infty \). Note that case (i) of Theorem 3.4 cannot occur when restricting \( L_{\nu,0} \) to \( D_0 \).

Next, let \( \{\Sigma_r\}_{r \in \mathbb{N}} \subseteq D_0 \) be a sequence with \( L_{\nu,0}(\Sigma_r) \to \infty \) \( \inf_{\Sigma \in D_0} L_{\nu,0}(\Sigma) \). Since \( \{\Sigma_r\}_{r \in \mathbb{N}} \) is bounded, it contains a convergent subsequence, whose limit is by the above argumentation an inner point, which completes the proof.

Remark 3.9. The case \( \nu = 0 \) is closely related to the so called projected normal distribution \( \Pi_{\mathcal{N}}(\mu, \Sigma) \) on \( S^d \), which is obtained by projecting a normal distribution onto the sphere, i.e. if \( X \sim \mathcal{N}(\mu, \Sigma) \), then

\[
Y = \frac{X}{||X||_2} \sim \Pi_{\mathcal{N}}(\mu, \Sigma).
\]

This distribution is also called angular Gaussian distribution, see [15]. We focus here on angular centered Gaussian distribution obtained for \( \mu = 0 \). The density of the \( \Pi_{\mathcal{N}}(0, \Sigma) \) distribution (with respect to the Lebesgue measure on \( S^d \)) is given by

\[
f(x|\Sigma) = \frac{\Gamma\left(\frac{d}{2}\right)}{(2\pi)^{\frac{d}{2}}} \frac{1}{||\Sigma||_{\frac{1}{2}}} \frac{1}{\left(x^T \Sigma^{-1} x\right)^{\frac{d}{2}}}. \tag{12}
\]

We have that \( f(x|\Sigma) = f(x|c\Sigma) \) for any \( c > 0 \), so that the positive definite matrix \( \Sigma \) is only identifiable up to a positive factor. Comparing the equations obtained by differentiating the logarithm of the density (12) to the equations (1) and (2) for \( \nu = 0 \) we see that ML estimation of \( \Sigma \) in \( \Pi_{\mathcal{N}}(\mu, \Sigma) \) corresponds exactly to ML estimation of \( \Sigma \) in \( T_0(0, \Sigma) \).

3.3. Estimation of Location and Scatter

In order to show the existence and uniqueness of a joint minimizer of the likelihood function we consider the location and scatter estimation problem as a higher dimensional centered scatter problem (i.e. the higher-dimensional location parameter is zero). This is achieved by
adding the location parameter as an additional column to the scatter matrix and modify
the resulting matrix to make it symmetric, positive definite using Schur complement. The
samples are extended by a known value and we consider the conditional distribution given
the appended value afterwards, thereby making use of a recent result on the conditional
multivariate Student $t$ -distributions [4]. To make our approach self-contained, we first recall
some elementary definitions and results.

The Schur complement (of $D$) of an invertible matrix

$$ M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} $$

is defined as $A - BD^{-1}C$. By the help of the inverse of the Schur complement we can express
the inverse $M^{-1}$ as

$$ M^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}. $$

In our situation, let $(\mu, \Sigma, \lambda) \in \mathbb{R}^d \times \text{GL}(d) \cap \text{Sym}(d) \times \mathbb{R}^*$ and consider the matrix

$$ A = \lambda \begin{pmatrix} \Sigma + \mu \mu^T & \mu \\ \mu^T & 1 \end{pmatrix} \in \mathbb{R}^{(d+1) \times (d+1)}. $$

Note that this parametrization of $A$ is unique, that is, for each $A \in \text{GL}(d+1) \cap \text{Sym}(d+1)$,
there exists a unique triple $(\mu, \Sigma, \lambda) \in \mathbb{R}^d \times \text{GL}(d) \cap \text{Sym}(d) \times \mathbb{R}^*$ and vice versa. With the
help of the Schur complement one sees that the inverse of $A$ is given by

$$ A^{-1} = \frac{1}{\lambda} \begin{pmatrix} \Sigma^{-1} & -\Sigma^{-1} \mu \\ -\mu^T \Sigma^{-1} & 1 + \mu^T \Sigma^{-1} \mu \end{pmatrix} $$

and for $z \in \mathbb{R}^{d+1}$ it holds

$$ z^T A^{-1} z = \frac{1}{\lambda} (1 + (x - \mu)^T \Sigma^{-1} (x - \mu)) \quad (13) $$

and $|A| = |\lambda \Sigma| = \lambda^d |\Sigma|$. Furthermore, we have $A \in \text{SPD}(d+1)$ if and only if $\Sigma \in \text{SPD}(d)$
and $\lambda > 0$.

**Lemma 3.10.** Let $x_1, \ldots, x_n$ be samples of $T_\nu(\mu, \Sigma)$ and associate to each $x_i \in \mathbb{R}^d$ a vector
$z_i = (x_i, 1) \in \mathbb{R}^{d+1}$. Further, let $\nu \geq 1$. Then it holds

$$ \arg\min_{(\mu, \Sigma) \in \mathbb{R}^d \times \text{SPD}(d)} L_\nu(\mu, \Sigma) = \arg\min_{A \in \text{P}(d+1)} \tilde{L}_\nu,0(A), $$

15
where
\[ L_{\tilde{\nu},0}(A) := (\tilde{d} + \tilde{\nu}) \sum_{i=1}^{n} w_i \log(\tilde{\nu} + z_i^T A_1^{-1} z_i) + \log(|A|) \]
with \( \tilde{d} = d + 1 \) and \( \tilde{\nu} = \nu - 1 \), the matrix \( A \) is related to \((\mu, \Sigma)\) via
\[ A = \begin{pmatrix} \Sigma + \mu \mu^T & \mu \\ \mu^T & 1 \end{pmatrix} \in \mathbb{R}^{(d+1) \times (d+1)}, \]
and
\[ P(d) := \{ A \in \text{SPD}(d) : a_{dd} = 1 \}. \]

**Proof.** We write the objective function associated to \( T_\nu(\mu, \Sigma) \) using (13) as
\[ L_\nu(\mu, \Sigma) = \sum_{i=1}^{n} w_i \rho(\delta_i) + \log(|\Sigma|) \]
\[ = (d + \nu) \sum_{i=1}^{n} w_i \log(\nu + \delta_i) + \log(|\Sigma|) \]
\[ = (d + 1 + \nu - 1) \sum_{i=1}^{n} w_i \log(\nu - 1 + z_i^T A_1^{-1} z_i) + \log(|A|) \]
\[ = (\tilde{d} + \tilde{\nu}) \sum_{i=1}^{n} w_i \log(\tilde{\nu} + z_i^T A_1^{-1} z_i) + \log(|A|) = L_{\tilde{\nu},0}(A), \]
where \( \tilde{d} = d + 1 \) and \( \tilde{\nu} = \nu - 1 \). As a consequence, minimizing \( L_\nu \) over \((\mu, \Sigma) \in \mathbb{R}^d \times \text{SPD}(d)\) with \( \nu \) degrees of freedom is equivalent to minimizing \( L_{\tilde{\nu},0}(A) \) with \( \tilde{\nu} = \nu - 1 \) degrees of freedom over \( \text{SPD}(d + 1) \) with the restriction that the \((d + 1)\)-th diagonal element equals one. This requires of course \( \tilde{\nu} \geq 0 \), that is \( \nu \geq 1 \).

In the following, we first state a counterpart of Assumption 3.1 for samples of the special form \( z_i = (x_i, 1) \in \mathbb{R}^{d+1} \), before we analyze critical points of \( L_{\tilde{\nu},0} \) on \( \text{SPD}(d + 1) \). To this aim, we recall that a vector \( x \in \mathbb{R}^d \) is called a **linear combination** of the vectors \( v_1, \ldots, v_k \in \mathbb{R}^d \), if
\[ x = \sum_{i=1}^{k} \lambda_i v_i \quad \text{for some } \lambda_i \in \mathbb{R}, \ i = 1, \ldots, k. \]
If in addition \( \sum_{i=1}^{k} \lambda_i = 1 \), we call \( x \) an **affine combination**. The vectors \( x_1, \ldots, x_k \in \mathbb{R}^d \) are said to be affinely independent if \( \sum_{i=1}^{k} \lambda_i = 0 \) implies \( \lambda_1 = \lambda_2 = \ldots = \lambda_k = 0 \). The next known lemma recalls the notions of affine and linear independence.

**Lemma 3.11.** The following statements are equivalent:

1. \( x \) is a linear combination of \( v_1, \ldots, v_k \).
2. \( x \) is an affine combination of \( v_1, \ldots, v_k \).
3. The vectors \( v_1, \ldots, v_k \) are affinely independent.
4. The vectors \( v_1, \ldots, v_k \) are linearly independent.

16
(i) The vectors \( x_1, \ldots, x_k \in \mathbb{R}^d \) are affinely independent.

(ii) The vectors \( x_2 - x_1, \ldots, x_k - x_1 \in \mathbb{R}^d \) are linearly independent.

(iii) The vectors \( (x_1^1, \ldots, x_k^1) \in \mathbb{R}^{d+1} \) are linearly independent.

Using Lemma 3.11 we directly obtain that the extended samples \( z_1, \ldots, z_k \in \mathbb{R}^{d+1} \) are linearly independent if and only if the vectors \( x_1, \ldots, x_k \in \mathbb{R}^d \) are affinely independent. Further, a set \( \{z_1, \ldots, z_k\} \) span a linear subspace \( V \subseteq \mathbb{R}^{d+1} \) with \( 1 \leq \dim(V) \leq d \) if and only if \( \{x_1, \ldots, x_k\} \) span an affine subspace \( H \subseteq \mathbb{R}^d \) with \( \dim(H) = \dim(V) - 1 \). Note that we can ignore the case \( \dim(V) = 0 \), i.e. \( V = \{0\} \) here, as \( z_i \neq 0, \ i = 1, \ldots, n \). As a consequence, we can modify Assumption 3.1 and Lemma 3.2 as follows:

**Assumption 3.12.** Let \( n \geq d+1 \).

(i) Let \( x_1, \ldots, x_n \in \mathbb{R}^d \) be a set of samples such that any subset of \( d \) samples is affinely independent.

(ii) Let \( w \in \hat{\Delta}_n \) fulfill \( w_{\max} < \frac{\nu + d - 1}{d(\nu + d)} \), where \( \nu \geq 1 \).

In analogy to Lemma 3.2 we have the following result.

**Lemma 3.13.** Let \( x_i \in \mathbb{R}^d, w_i \in \mathbb{R}, i = 1, \ldots, n \), fulfill Assumption 3.12. Further, let \( H \subset \mathbb{R}^d \) be an affine subspace with \( 0 \leq \dim(H) \leq d - 1 \). Then it holds

\[
\sum_{i \in I_H} w_i < \frac{\nu + \dim(H)}{\nu + d},
\]

where \( I_H = \{i \in \{1, \ldots, n\} : x_i \in H\} \).

**Proof.** Let \( 0 \leq k = \dim(H) \leq d - 1 \). Then, according to Assumption (3.12)(i), it holds \( |I_H| \leq k + 1 \). Using \( w_{\max} < \frac{\nu + d - 1}{d(\nu + d)} \) and \( \nu \geq 1 \) we obtain

\[
\sum_{i \in I_H} w_i < (k + 1) \frac{\nu + d - 1}{d(\nu + d)} = \frac{k+1}{d} \frac{(\nu - 1) + k + 1}{\nu + d} \leq \frac{\nu - 1 + k + 1}{\nu + d} = \frac{\nu + \dim(H)}{\nu + d}.
\]

**Remark 3.14.** If (14) holds true, then the affine independence assumption implies \( n \geq d+1 \). Indeed, assume that \( n < d+1 \) and let \( H \) be an affine subspace containing all the samples. Then \( \dim(H) \leq d - 1 \), but

\[
1 = \sum_{i \in I_H} w_i < \frac{\nu + \dim(H)}{\nu + d} \leq \frac{\nu + d - 1}{\nu + d} < 1,
\]

which gives a contradiction.
Theorem 3.15 (Existence of Location and Scatter). Let \( x_i \in \mathbb{R}^d, w_i > 0, i = 1, \ldots, n \), fulfill Assumption 3.12 and let \( \nu > 1 \). Then it holds

\[
\arg\min_{(\mu, \Sigma) \in \mathbb{R}^d \times P_d} L_\nu(\mu, \Sigma) \neq \emptyset
\]

and any \((\hat{\mu}, \hat{\Sigma}) \in \arg\min_{(\mu, \Sigma) \in \mathbb{R}^d \times P_d} L_\nu(\mu, \Sigma)\) is a critical point of \( L_\nu \).

The counterpart of Lemma 3.5 reads as follows:

Lemma 3.16. Let \( x_i \in \mathbb{R}^d, w_i > 0, i = 1, \ldots, n \), fulfill Assumption 3.12 and assume there exists a critical point \((\hat{\mu}, \hat{\Sigma})\) of \( L_\nu \). Then, for all affine subspaces \( H \subseteq \mathbb{R}^d \) with \( 0 \leq \dim(H) \leq d - 1 \) it holds

\[
\sum_{i \in I_H} w_i \leq \frac{\nu + \dim(H)}{d + \nu},
\]

where \( I_H = \{ i \in \{1, \ldots, n\} : x_i \in H \} \).

Proof. W.l.o.g. we might assume that \( \hat{\mu} = 0 \) and \( \hat{\Sigma} = I \), otherwise we might transform the samples to \( y_i = R^{-1}(x_i - \hat{\mu}) \), where \( RR^T = \hat{\Sigma} \). Observing that \( H \) is in this case a linear space, the rest of the proof follows as in the proof of Lemma 3.5.

We have the following existence and uniqueness result.

Theorem 3.17 (Uniqueness of Location and Scatter). Let \( x_i \in \mathbb{R}^d, w_i > 0, i = 1, \ldots, n \), fulfill Assumption 3.12 and let \( \nu > 1 \). Then, there exists a unique critical point \((\hat{\mu}, \hat{\Sigma}) \in \mathbb{R}^d \times \text{SPD}(d)\) of \( L_\nu \).

Proof. By the previous considerations it remains to prove that there exists a unique critical point \( \hat{A} \in \text{P}(d) \) of \( \tilde{L}_{\nu, 0} \). Let \( \hat{A} \) be the unique critical point of \( \tilde{L}_{\nu, 0} \) on \( \text{SPD}(d) \), which is of the general form

\[
\hat{A} = \hat{\lambda} \begin{pmatrix} \hat{\Sigma} + \hat{\mu}\hat{\mu}^T & \hat{\mu} \\ \hat{\mu}^T & 1 \end{pmatrix}.
\]

Then, it suffices to show that \( \hat{\lambda} = 1 \). Now, according to (5), the matrix \( \hat{A} \) fulfills the fixed-point equation

\[
\hat{A} = (\tilde{d} + \tilde{\nu}) \sum_{i=1}^{n} w_i \frac{z_i z_i^T}{\nu + z_i \hat{A}_k^{-1} z_i},
\]

and using

\[
z_i z_i^T = \begin{pmatrix} x_i x_i^T & x_i \\ x_i^T & 1 \end{pmatrix}
\]

18
this can be rewritten as

\[ \hat{A} = (\tilde{d} + \tilde{\nu}) \sum_{i=1}^{n} w_i \frac{z_i z_i^T}{\tilde{\nu} + z_i \hat{A}^{-1} z_i} = (d + \nu) \sum_{i=1}^{n} w_i \frac{x_i x_i^T}{\nu - 1 + \frac{1}{\nu} \left(1+(x_i - \hat{\mu})^T \hat{\Sigma}^{-1} (x_i - \hat{\mu})\right)} \]

By (3) it holds further

\[ 1 = (\tilde{d} + \tilde{\nu}) \sum_{i=1}^{n} \frac{1}{\tilde{\nu} + z_i \hat{A}^{-1} z_i} = (d + \nu) \sum_{i=1}^{n} \frac{1}{\nu - 1 + \frac{1}{\nu} \left(1+(x_i - \hat{\mu})^T \hat{\Sigma}^{-1} (x_i - \hat{\mu})\right)} \]

so that indeed \( \hat{\lambda} = 1 \), which finishes the proof.

The multivariate Cauchy distribution \((\nu = 1)\) requires a special consideration. Using also in this case the reformulation as a centered scatter only problem yields \( \tilde{\nu} = 0 \), so that according to Proposition 3.7 and Theorem 3.8 the matrix \( \hat{A} \) can only be identified up to a positive scalar factor. However, observing that this does not change the values of \( \hat{\mu} \) and \( \hat{\Sigma} \), we conclude that uniqueness holds also in the Cauchy case.

4. Efficient Minimization Algorithm

In this section, we propose an efficient algorithm to compute the maximum likelihood estimates of the multivariate Student-\( t \) distribution and prove its convergence.

4.1. Algorithm

We assume that \( \nu \geq 0 \) in case of estimating only \( \Sigma \) and \( \nu \geq 1 \) when estimating both \( \mu \) and \( \Sigma \). We denote with \( \Sigma = R^2 \) a square root of \( \Sigma \) (alternatively: \( \Sigma = RR^T \) Cholesky decomposition) and define the standardized samples \( y_i = R^{-1}(x_i - \mu) \). Further, we introduce the notation

\[ S_0(\mu, \Sigma) := (d + \nu) \sum_{i=1}^{n} w_i \frac{1}{\nu + (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)} \]

\[ = (d + \nu) \sum_{i=1}^{n} w_i \frac{1}{\nu + \delta_i}, \tag{15} \]
\[ S_1(\mu, \Sigma) := (d + \nu) \sum_{i=1}^{n} w_i \frac{R^{-1}(x_i - \mu)}{\nu + (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)} \]
\[ = (d + \nu) \sum_{i=1}^{n} w_i \frac{y_i}{\nu + \delta_i}, \]
\[ S_2(\mu, \Sigma) := (d + \nu) \sum_{i=1}^{n} w_i \frac{R^{-1}(x_i - \mu)(x_i - \mu)^T R^{-T}}{\nu + (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)} \]
\[ = (d + \nu) \sum_{i=1}^{n} w_i \frac{y_i y_i^T}{\nu + \delta_i}. \quad (16) \]

Recall from (4) and (5) that a point \((\hat{\mu}, \hat{\Sigma})\) is a critical point of \(L_\nu\) if and only if it fulfills the fixed point equations

\[ \hat{\mu} = \sum_{i=1}^{n} w_i \frac{x_i}{\nu + (x_i - \hat{\mu})^T \Sigma^{-1} (x_i - \hat{\mu})} = \hat{\mu} + R \frac{S_1(\hat{\mu}, \hat{\Sigma})}{S_0(\hat{\mu}, \hat{\Sigma})}, \]
\[ \hat{\Sigma} = \sum_{i=1}^{n} w_i \frac{(x_i - \hat{\mu})(x_i - \hat{\mu})^T}{\nu + (x_i - \hat{\mu})^T \Sigma^{-1} (x_i - \hat{\mu})} = \hat{\Sigma} + R \frac{S_2(\hat{\mu}, \hat{\Sigma})}{S_0(\hat{\mu}, \hat{\Sigma})} \hat{R}^T. \]

Based on these fixed point equations we formulate the iterative scheme

\[ \mu_{r+1} = \mu_r + R_r \frac{S_1(\mu_r, \Sigma_r)}{S_0(\mu_r, \Sigma_r)}, \]
\[ \Sigma_{r+1} = R_{r+1} R_r^T = R_r \frac{S_2(\mu_r, \Sigma_r)}{S_0(\mu_r, \Sigma_r)} R_r^T \]
\[ = R_r \frac{\nu S_2(\mu_r, \Sigma_r)}{d + \nu - \text{tr}(S_2(\mu_r, \Sigma_r))} R_r^T. \]

Note that for \(d = 1\), this coincides with the generalized myriad filtering considered in [11]. If one of the parameters is known, it can be held fixed in the above iterative scheme in order to estimate the other parameter. The resulting algorithm is summarized in Algorithm 1. For comparison, we also state the EM algorithm [14] in our notation in Algorithm 2. It turns out that the updates for \(\mu\) coincide in both algorithms, while the update for \(\Sigma\) differ by a division with \(S_0\). The EM algorithm uses a fixed scalar factor \(\frac{1}{\nu + d}\) instead of \(\frac{1}{S_0(\mu_r, \Sigma_r)}\). It turns out that the balancing between \(S_0\) and \(S_2\) leads in particular for small \(\nu\) to a much faster convergence, see the simulation study at the end of this section, whereas it does not increase the computational complexity, since the quantity \(S_0(\mu_r, \Sigma_r)\) has to be computed anyway.
Algorithm 1 Minimization of $L_{\nu}(\cdot, \cdot)$ (Generalized Multivariate Myriad Filter, GMMF)

**Input:** $x_1, \ldots, x_n \in \mathbb{R}^d$, $n \geq d + 1$, $0 < w_i < \frac{\nu + d - 1}{(\nu + d)(d-1)}$, $i = 1, \ldots, n$, $\sum_{i=1}^{n} w_i = 1$

**Initialization:** $\mu_0 = 0$, $\Sigma_0 = I$

for $r = 0, \ldots$ do

$$\mu_{r+1} := \mu_r + R_r \frac{S_1(\mu_r, \Sigma_r)}{S_0(\mu_r, \Sigma_r)}$$

$$\Sigma_{r+1} := R_r \frac{S_2(\mu_r, \Sigma_r)}{S_0(\mu_r, \Sigma_r)} R_r^{T}$$

Algorithm 2 Minimization of $L_{\nu}(\cdot, \cdot)$ (Expectation-Maximization, EM)

**Input:** $x_1, \ldots, x_n \in \mathbb{R}^d$, $n \geq d + 1$, $0 < w_i < \frac{\nu + d - 1}{(\nu + d)(d-1)}$, $i = 1, \ldots, n$, $\sum_{i=1}^{n} w_i = 1$

**Initialization:** $\mu_0 = 0$, $\Sigma_0 = I$

for $r = 0, \ldots$ do

$$\mu_{r+1} := \mu_r + R_r \frac{S_1(\mu_r, \Sigma_r)}{S_0(\mu_r, \Sigma_r)}$$

$$\Sigma_{r+1} := R_r \frac{S_2(\mu_r, \Sigma_r)}{\nu + d} R_r^{T}$$

4.2. Convergence Analysis of the Algorithm

In this subsection, we prove the convergence of the proposed algorithm. Since we will need it at several places we compute the difference of two iterates as

$$L_{\nu}(\mu_{r+1}, \Sigma_{r+1}) - L_{\nu}(\mu_r, \Sigma_r)$$

$$= (d + \nu) \sum_{i=1}^{n} w_i \log \left( \frac{\nu + (x_i - \mu_{r+1})^T \Sigma_{r+1}^{-1} (x_i - \mu_r)}{\nu + (x_i - \mu_r)^T \Sigma_r^{-1} (x_i - \mu_r)} \right)$$

$$= (d + \nu) \sum_{i=1}^{n} w_i \log \left( \frac{\nu + (x_i - \mu_{r+1})^T \Sigma_{r+1}^{-1} (x_i - \mu_r)}{\nu + (x_i - \mu_r)^T \Sigma_r^{-1} (x_i - \mu_r)} \frac{\frac{1}{\omega_{r+1}^2}}{\frac{1}{\omega_r^2}} \right)$$

**Theorem 4.1.** For fixed $\mu \in \mathbb{R}^d$ and $\nu > 0$, let $\{\Sigma_r\}_{r \in \mathbb{N}}$ be defined by Algorithm 1. Then it holds

$$L_{\nu}(\mu, \Sigma_{r+1}) - L_{\nu}(\mu, \Sigma_r) \leq 0.$$ 

**Proof.** By concavity of the logarithm we have

$$L_{\nu}(\mu, \Sigma_{r+1}) - L_{\nu}(\mu, \Sigma_r) = (d + \nu) \sum_{i=1}^{n} w_i \log \left( \frac{\nu + (x_i - \mu)^T \Sigma_{r+1}^{-1} (x_i - \mu)}{\nu + (x_i - \mu)^T \Sigma_r^{-1} (x_i - \mu)} \frac{\frac{1}{\omega_{r+1}^2}}{\frac{1}{\omega_r^2}} \right)$$

21
Thus, we obtain

\[
\frac{\Sigma_{r+1}^{-1/\xi}}{\eta} \leq (d + \nu) \log \left( \sum_{i=1}^{n} w_i \frac{\nu + (x_i - \mu)^T \Sigma_{r+1}^{-1}(x_i - \mu)}{\nu + (x_i - \mu)^T \Sigma_{r}^{-1}(x_i - \mu)} \right),
\]

so that it suffices to show that \( \Upsilon \leq 1 \). We abbreviate \( S_{0r} = S_0(\mu, \Sigma_r) \) and \( S_{2r} = S_2(\mu, \Sigma_r) \) and start with analyzing the factor \( \frac{\Sigma_{r+1}^{-1/\xi}}{\eta} \). Using properties of the determinant it holds

\[
\frac{\Sigma_{r+1}^{-1/\xi}}{\eta} = \frac{|R_r S_{0r} R_r^T|^{1/\xi}}{|R_r R_r^T|^{1/\xi}} = S_{0r}^{-\frac{\eta}{\xi}} |S_{2r}|^{1/\xi}.
\]

Next, we consider the term

\[
\sum_{i=1}^{n} w_i \frac{\nu + (x_i - \mu)^T \Sigma_{r+1}^{-1}(x_i - \mu)}{\nu + (x_i - \mu)^T \Sigma_{r}^{-1}(x_i - \mu)} = \sum_{i=1}^{n} w_i \frac{(x_i - \mu)^T \Sigma_{r}^{-1}(x_i - \mu)}{\nu + (x_i - \mu)^T \Sigma_{r}^{-1}(x_i - \mu)} + \frac{\nu}{d + \nu} S_{0r}.
\]

Using

\[
(x_i - \mu)^T \Sigma_{r+1}^{-1}(x_i - \mu) = \text{tr} \left( (x_i - \mu)^T \Sigma_{r+1}^{-1} \right)
\]

and the linearity of the trace, the sum simplifies to

\[
\sum_{i=1}^{n} w_i \frac{(x_i - \mu)^T \Sigma_{r+1}^{-1}(x_i - \mu)}{\nu + (x_i - \mu)^T \Sigma_{r}^{-1}(x_i - \mu)} = \sum_{i=1}^{n} w_i \frac{\text{tr} \left( (x_i - \mu)^T \Sigma_{r+1}^{-1} \right)}{\nu + (x_i - \mu)^T \Sigma_{r}^{-1}(x_i - \mu)}
\]

\[
= S_{0r} \sum_{i=1}^{n} w_i \frac{\text{tr} \left( S_{2r}^{-1} R_r^{-1}(x_i - \mu)(x_i - \mu)^T R_r^{-T} \right)}{\nu + (x_i - \mu)^T \Sigma_{r}^{-1}(x_i - \mu)}
\]

\[
= S_{0r} \text{tr} \left( S_{2r}^{-1} \sum_{i=1}^{n} w_i \frac{(x_i - \mu)^T \Sigma_{r}^{-1}(x_i - \mu)}{\nu + (x_i - \mu)^T \Sigma_{r}^{-1}(x_i - \mu)} \right) = \frac{1}{d + \nu} S_{0r}.
\]

Thus, we obtain

\[
\Upsilon = \sum_{i=1}^{n} w_i \frac{\nu + (x_i - \mu)^T \Sigma_{r+1}^{-1}(x_i - \mu)}{\nu + (x_i - \mu)^T \Sigma_{r}^{-1}(x_i - \mu)} \frac{1}{\eta} \Sigma_{r+1}^{-1/\xi}. \]

22
\[
= \left( \frac{d}{d + \nu} + \frac{\nu}{d + \nu} \right) S_{0r} - \frac{d + \nu}{d + \nu} |S_{2r}|^{\frac{1}{d + \nu}} = (S_{0r}|S_{2r}|)^{\frac{d}{d + \nu}}.
\]

We have \( \nu S_{0r} + \text{tr}(S_{2r}) = d + \nu \), such that we can express \( S_{0r} \) in terms of \( \text{tr}(S_{2r}) \) as

\[
S_{0r} = \frac{d + \nu}{\nu} - \frac{1}{\nu} \text{tr}(S_{2r}).
\]

Next we consider maximizing the function

\[
g: \ SPD(d) \rightarrow \mathbb{R}, \quad g(X) = \left( \frac{d + \nu}{\nu} - \frac{1}{\nu} \text{tr}(X) \right)^{\nu} |X|
\]

under the constraint \( 0 \leq \text{tr}(X) \leq d + \nu \). Note that for \( X \in SPD(d) \) we always have \( \text{tr}(X), |X| > 0 \). Further, if \( \text{tr}(X) = 0 \) or \( \text{tr}(X) = d + \nu \), we set \( g(X) = 0 \). Since \( g(I) = 1 \), the maximum is not attained at the boundary, but inside the open set. The derivative of \( g \) with respect to \( X \) is given by

\[
\nabla g(X) = -\nu \left( \frac{d + \nu}{\nu} - \frac{1}{\nu} \text{tr}(X) \right)^{\nu - 1} \frac{1}{\nu} |X|^{\nu - 1} \text{tr}(X) X^{-1}
\]

\[
= \left( \frac{d + \nu}{\nu} - \frac{1}{\nu} \text{tr}(X) \right)^{\nu - 1} |X|^{\nu - 1} \left[ \left( \frac{d + \nu}{\nu} - \frac{1}{\nu} \text{tr}(X) \right) X^{-1} - I \right].
\]

The necessary condition for a critical point \( \hat{X} \) of \( g \) reads as

\[
\left( \frac{d + \nu}{\nu} - \frac{1}{\nu} \text{tr}(\hat{X}) \right) I = \hat{X}.
\]

To verify that \( \hat{X} = I \) is a maximizer, we have a look at the Hessian \( \nabla^2 g \) of \( g \) and show that it is negative definite for \( \hat{X} = I \). We compute

\[
D(\nabla g)(X)[H] = |X| \left( \frac{d + \nu}{\nu} - \frac{1}{\nu} \text{tr}(|X|) \right)^{\nu - 2} \left[ \frac{\nu - 1}{\nu} \text{tr}(H) I
\right.
\]

\[
- \left( \frac{d + \nu}{\nu} - \frac{1}{\nu} \text{tr}(|X|) \right) \left( \text{tr}(X^{-1} H) I + \text{tr}(H) X^{-1} \right)
\]

\[
+ \left( \frac{d + \nu}{\nu} - \frac{1}{\nu} \text{tr}(|X|) \right)^2 \text{tr}(X^{-1} H) X^{-1} - X^{-1} H X^{-1}
\]

and further

\[
\langle \nabla^2 g(X)[H], H \rangle = |X| \left( \frac{d + \nu}{\nu} - \frac{1}{\nu} \text{tr}(|X|) \right)^{\nu - 2} \left[ \frac{\nu - 1}{\nu} \text{tr}(H)^2
\right.
\]

\[
- 2 \left( \frac{d + \nu}{\nu} - \frac{1}{\nu} \text{tr}(|X|) \right) \text{tr}(H) \text{tr}(X^{-1} H)
\]

23
Then it holds

\[
\text{For fixed } 
\]

Theorem 4.2. For fixed \( \Sigma \in \text{SPD}(d) \) and \( \nu > 0 \), let \( \{\mu_r\}_{r \in \mathbb{N}} \) be defined by Algorithm 1. Then it holds

\[
L_{\nu}(\mu_{r+1}, \Sigma) - L_{\nu}(\mu_r, \Sigma) \leq 0.
\]

Proof. By concavity of the logarithm we have that

\[
L_{\nu}(\mu_{r+1}, \Sigma) - L_{\nu}(\mu_r, \Sigma) = (d + \nu) \sum_{i=1}^{n} w_i \log \left( \frac{\nu + (x_i - \mu_{r+1})^T \Sigma^{-1}(x_i - \mu_{r+1})}{\nu + (x_i - \mu_r)^T \Sigma^{-1}(x_i - \mu_r)} \right)
\]

so that it suffices to show that \( \Upsilon \leq 1 \). We abbreviate \( S_{0\nu} = S_0(\mu_r, \Sigma) \) and \( S_{1\nu} = S_1(\mu_r, \Sigma) \) and compute

\[
\Upsilon = \sum_{i=1}^{n} w_i \nu + (x_i - \mu_r)^T \Sigma^{-1}(x_i - \mu_r) + 2 \langle R^{-1}(x_i - \mu_r), (R^{-1}(\mu_r - \mu_{r+1})) \rangle + (\mu_r - \mu_{r+1})^T \Sigma^{-1}(\mu_r - \mu_{r+1})
\]

\[
= 1 - 2 \sum_{i=1}^{n} \frac{w_i}{\nu + (x_i - \mu_r)^T \Sigma^{-1}(x_i - \mu_r)} \langle R^{-1}(x_i - \mu_r), \frac{S_{1\nu}}{S_{0\nu}} \rangle + \sum_{i=1}^{n} w_i \frac{\| S_{1\nu} \|_2^2}{\nu + (x_i - \mu_r)^T \Sigma^{-1}(x_i - \mu_r)}
\]

At the critical point \( \hat{X} = I \) we obtain

\[
\langle \nabla^2 g(I)[H], H \rangle = \frac{\nu - 1}{\nu} \text{tr}(H)^2 - 2 \text{tr}(H)^2 + \text{tr}(H)^2 - \text{tr}(H^2)
\]

\[
= -\frac{1}{\nu} \text{tr}(H)^2 - \text{tr}(H^2) < 0,
\]

so that \( \hat{X} = I \) is indeed a maximizer. The corresponding maximum of \( g \) is given by

\[
g(I) = \left( \frac{d+\nu}{\nu} - \frac{1}{\nu} \text{tr}(I) \right)^\nu |I| = 1 \text{ and therewith finally}
\]

\[
\Upsilon = \langle S_{0\nu}^2, S_{2\nu} \rangle^{\frac{1}{2+\nu}} = g(S_{2\nu})^{\frac{1}{2+\nu}} \leq 1.
\]

\[
\square
\]
For we have

\[ \sum_{i=1}^{n} \frac{R^{-1}(x_i - \mu_r)}{\nu + (x_i - \mu_r)^T \Sigma^{-1}(x_i - \mu_r)} \frac{S_{1r}}{S_{0r}} + \| \frac{S_{1r}}{S_{0r}} \|_F^2 \sum_{i=1}^{n} \frac{1}{\nu + (x_i - \mu_r)^T \Sigma^{-1}(x_i - \mu_r)} = \frac{1}{\nu + S_{0r}} \]

\[ = 1 - \frac{1}{d + \nu} \frac{\| S_{1r} \|_F^2}{2} \leq 1, \]

with equality if and only if \( S_{1r} = 0 \), that is, \( \mu_{r+1} = \mu_r \) and \( \mu_r \) is a critical point of \( L_\nu(\cdot, \Sigma) \).

Combining the results of Theorem 4.2 and 4.1 we obtain the following theorem.

**Theorem 4.3.** For \( \nu > 0 \), let \( \{ \mu_r, \Sigma_r \}_{r \in \mathbb{N}} \) be defined by Algorithm 1. Then it holds

\[ L_\nu(\mu_{r+1}, \Sigma_{r+1}) - L_\nu(\mu_r, \Sigma_r) \leq 0. \]

**Proof.** By concavity of the logarithm we have that

\[
L_\nu(\mu_{r+1}, \Sigma_{r+1}) - L_\nu(\mu_r, \Sigma_r)
= (d + \nu) \sum_{i=1}^{n} w_i \log \left( \frac{\nu + (x_i - \mu_{r+1})^T \Sigma_{r+1}^{-1}(x_i - \mu_{r+1})}{\nu + (x_i - \mu_r)^T \Sigma_r^{-1}(x_i - \mu_r)} \right) + \log \left( \frac{|\Sigma_{r+1}|}{|\Sigma_r|} \right)
= (d + \nu) \sum_{i=1}^{n} w_i \log \left( \frac{\nu + (x_i - \mu_{r+1})^T \Sigma_{r+1}^{-1}(x_i - \mu_{r+1})}{\nu + (x_i - \mu_r)^T \Sigma_r^{-1}(x_i - \mu_r)} \frac{|\Sigma_{r+1}|}{|\Sigma_r|} \right)
\leq (d + \nu) \log \left( \sum_{i=1}^{n} w_i \frac{\nu + (x_i - \mu_{r+1})^T \Sigma_{r+1}^{-1}(x_i - \mu_{r+1})}{\nu + (x_i - \mu_r)^T \Sigma_r^{-1}(x_i - \mu_r)} \frac{|\Sigma_{r+1}|}{|\Sigma_r|} \right),
\]

so that it suffices to show that \( \Upsilon \leq 1 \). We abbreviate \( S_{0r} = S_0(\mu_r, \Sigma_r) \), \( S_{1r} = S_1(\mu_r, \Sigma_r) \) and \( S_{2r} = S_2(\mu_r, \Sigma_r) \) and analyze the components of \( \Upsilon \) separately. As in the proof of Theorem 4.1 we have

\[ \frac{|\Sigma_{r+1}|}{|\Sigma_r|} \frac{1}{\nu + S_{0r}} = S_{0r}^{-\frac{d}{\nu + S_{0r}}} |S_{2r}| \frac{1}{\nu + S_{0r}}. \]

Next, we consider the term

\[ \sum_{i=1}^{n} w_i \frac{\nu + (x_i - \mu_{r+1})^T \Sigma_{r+1}^{-1}(x_i - \mu_{r+1})}{\nu + (x_i - \mu_r)^T \Sigma_r^{-1}(x_i - \mu_r)} = \sum_{i=1}^{n} w_i \frac{(x_i - \mu_{r+1})^T \Sigma_{r+1}^{-1}(x_i - \mu_{r+1})}{\nu + (x_i - \mu_r)^T \Sigma_r^{-1}(x_i - \mu_r)} = \frac{\nu}{d + \nu} S_{0r}. \]

Combining the computations in the proofs of Theorem 4.2 and Theorem 4.1 we get for the sum

\[ \sum_{i=1}^{n} w_i \frac{(x_i - \mu_{r+1})^T \Sigma_{r+1}^{-1}(x_i - \mu_{r+1})}{\nu + (x_i - \mu_r)^T \Sigma_r^{-1}(x_i - \mu_r)} \]
\[ \sum_{i=1}^{n} w_i (x_i - \mu_r)^T \Sigma_r^{-1} (x_i - \mu_r) + 2 (R_{r+1}^{-1} (x_i - \mu_r) + (x_i - \mu_{r+1})^T \Sigma_{r+1}^{-1} (x_i - \mu_{r+1}) \]

\[ = \sum_{i=1}^{n} w_i S_{0r} \text{tr} (S_{2r}^{-1} R_r^{-1} (x_i - \mu_r)(x_i - \mu_r)^T R_r^{-T}) - 2 \sum_{i=1}^{n} w_i \frac{(x_i - \mu_r)^T R_r^{-T} S_{0r} S_{2r}^{-1} R_r^{-1} \nu S_{nr}}{\nu + (x_i - \mu_r)^T \Sigma_r^{-1} (x_i - \mu_r)} \]

\[ + \sum_{i=1}^{n} w_i S_{nr}^T R_r^{-T} S_{0r} S_{2r}^{-1} R_r^{-1} \nu S_{nr} = \frac{d}{d + \nu} S_{0r} - \frac{2}{d + \nu} S_{1r}^T S_{2r}^{-1} S_{1r} + \frac{1}{d + \nu} \sum_{i=1}^{n} w_i \frac{1}{\nu + (x_i - \mu_r)^T \Sigma_r^{-1} (x_i - \mu_r)} \]

Since \( S_{2r} \in \text{SPD}(d) \) and consequently also \( S_{2r}^{-1} \in \text{SPD}(d) \) we obtain

\[ Y = \left( \frac{d}{d + \nu} + \frac{\nu}{d + \nu} \right) S_{0r} - \frac{1}{d + \nu} \frac{S_{1r}^T S_{2r}^{-1} S_{1r}}{\geq 0} S_{0r} \frac{\nu}{\Sigma_{r}^{1/2}} |S_{2r}|^{1/2} \leq 1. \]

**Theorem 4.4** (Convergence of Algorithm 1). For any \( \nu > 0 \), the sequence \( \{\mu_r, \Sigma_r\}_{r \in \mathbb{N}} \) generated by Algorithm 1 converges.

**Proof.** Let \( \{\mu_r, \Sigma_r\}_{r \in \mathbb{N}} \) be the sequence of iterates generated by Algorithm 1. Consider the mapping

\[ T(\mu, \Sigma) = \left( \mu + \Sigma^\frac{1}{2} \frac{S_{1}(\mu, \Sigma)}{S_{0}(\mu, \Sigma)} \right) \Sigma^\frac{1}{2} \frac{S_{1}(\mu, \Sigma)}{S_{0}(\mu, \Sigma)} . \]

Then, according to (4), (5) and Theorems 3.15 and 3.17, \( (\mu, \Sigma) = T(\mu, \Sigma) \) is a fixed point of \( T \) if and only if it is the unique minimizer of \( L_\nu \). Consider the case \( (\mu_{r+1}, \Sigma_{r+1}) \neq (\mu_r, \Sigma_r) \) for all \( r \in \mathbb{N} \). We show that the sequence \( \{\mu_r, \Sigma_r\}_{r \in \mathbb{N}} \) is bounded: for \( \mu_r \), the update can be rewritten as

\[ \mu_{r+1} = \frac{1}{\sum_{i=1}^{n} w_i \frac{1}{\nu + (x_i - \mu_r)^T \Sigma_r^{-1} (x_i - \mu_r)}} \]

so it is a convex combination of the samples \( x_1, \ldots, x_n \). Concerning \( \Sigma_r \), we get by (15) and (16)

\[ \Sigma_{r+1} = \frac{1}{\sum_{i=1}^{n} w_i \frac{1}{\nu + (x_i - \mu_r)^T \Sigma_r^{-1} (x_i - \mu_r)}} \]

Thus, \( \Sigma_r \) is a weighted average as well and the sequence remains bounded since \( \nu_r \in \text{conv}(x_1, \ldots, x_n) \) stays bounded, \( (x_i - \mu_r)(x_i - \mu_r)^T \) is bounded as well, \( i = 1, \ldots, n \), and
consequently also $\Sigma_{r+1}$. By Theorem 4.3 we see that the sequence $L_r := L_{\nu}(\mu_r, \Sigma_r)$ is a strictly decreasing, bounded below sequence such that it converges to some $\hat{L}$. Further, $\{(\mu_r, \Sigma_r)\}_{r \in \mathbb{N}}$ contains a convergent subsequence $\{(\mu_{r_s}, \Sigma_{r_s})\}_{s \in \mathbb{N}}$, which converges to some $(\hat{\mu}, \hat{\Sigma})$. By the continuity of $L$ and $T$ we obtain

$$L(\hat{\mu}, \hat{\Sigma}) = \lim_{s \to \infty} L_{\nu}(\mu_{r_s}, \Sigma_{r_s}) = \lim_{s \to \infty} L_{r_s} = \lim_{s \to \infty} L_{r_s+1}$$

$$= \lim_{s \to \infty} L_{\nu}(\mu_{r_s+1}, \Sigma_{r_s+1})$$

$$= \lim_{s \to \infty} L(T(\mu_{r_s}, \Sigma_{r_s})) = L(T(\hat{\mu}, \hat{\Sigma})) .$$

This implies $(\hat{\mu}, \hat{\Sigma}) = T(\hat{\mu}, \hat{\Sigma})$, so that $(\hat{\mu}, \hat{\Sigma})$ is a fixed point of $T$ and consequently the minimizer. Since the minimizer is unique, not only a subsequence, but the whole sequence $\{(\mu_r, \Sigma_r)\}_{r \in \mathbb{N}}$ converges to $(\hat{\mu}, \hat{\Sigma})$, which finishes the proof. \qed

4.3. Simulation Study

In order to evaluate the numerical performance, in particular the speed of convergence, of the proposed algorithm compared to the EM algorithm we did the following Monte Carlo simulation: we draw $n = 100$ i.i.d. samples of a $T_{\nu}(\mu, \Sigma)$ distribution for different degrees of freedom $\nu \in \{1, 5, 10, 100\}$ and run Algorithms 1 respective Algorithm 2 to compute the joint ML-estimate $(\hat{\mu}, \hat{\Sigma})$. Both algorithms are initialized with sample mean and sample covariance and we used the relative difference between two iterates $(\mu_r, \Sigma_r)$ and $(\mu_{r+1}, \Sigma_{r+1})$ as stopping criterion, that is

$$\sqrt{\|\mu_{r+1}-\mu_r\|_F^2 + \|\Sigma_{r+1}-\Sigma_r\|_F^2} < 10^{-6}.$$ 

This experiment is repeated $N = 10,000$ times and afterwards, we calculated the average number of iterations $\bar{\text{iter}}$ and $\bar{\text{iter}}_{\text{EM}}$ needed to reach the tolerance criterion together with their standard deviations. The results are given in Table 1, where we chose $d = 2$, $\mu = 0$ and different values for $\Sigma$. First, we notice that the average number of iterations is in general higher for the EM Algorithm 2, and further, it does merely not depend on $(\mu, \Sigma)$, but only on the degree of freedom $\nu$. Here, the smaller the value of $\nu$, the larger on the one hand the number of iterations for both algorithms, and on the other hand the larger the gain in speed of Algorithm 1 compared to Algorithm 2.

5. Applications in Image Analysis

In this section, we describe how the developed GMMF can be used to denoise images corrupted by different kinds of additive noise. To this aim, let $f : G \to \mathbb{R}$ be a noisy image, where $G = \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$ denotes the image domain.
We assume that each pixel $i = (i_1, i_2) \in G$ is affected by the noise in an independent and identical way, and model the image pixelwise as

$$f_i = u + \frac{\sigma \eta}{\sqrt{\nu}}, \quad \eta \sim \mathcal{N}(0, 1), \ t \sim \chi^2_\nu, \ i \in G,$$

where $\eta \perp t$ are independent, $u$ is the noise-free image we wish to reconstruct and $\sigma > 0, \nu \geq 1$ are assumed to be known. If $\sigma$ and/or $\nu$ are unknown, they might be estimated in constant areas of the image. Constant regions can be found using e.g. the method presented in [29], further details of this method can be found in [11]. The parameter $\nu > 0$ determines the amount of outliers, while $\sigma > 0$ determines their strength. Together with the properties of the Student-$t$ distribution, see Theorem 2.1, this results in independent realizations $f_i$ of $T_\nu(u, \sigma)$ random variables, $i \in G$. Now, for each $i \in G$ we wish to estimate the underlying $u_i$ using a nonlocal generalized myriad filtering approach.

The estimation of the noise-free image requires to select for each $i \in G$ a set of indices of samples $S(i)$ that are interpreted as i.i.d. realizations of $T_\nu(u, \sigma)$. We focus here on a nonlocal approach, which is based on an image self-similarity assumption stating that small

| $\Sigma$ | $\nu$ | $\text{iter} \pm \sigma(\text{iter})$ | $\text{iter}_{EM} \pm \sigma(\text{iter}_{EM})$ |
|---------|------|----------------------------------|-----------------------------------|
| $(0.1, 0)$ | 1 | $20.7552 \pm 1.5430$ | $60.6318 \pm 3.9313$ |
| | 2 | $16.0843 \pm 1.1242$ | $33.5389 \pm 2.0370$ |
| | 5 | $11.166 \pm 0.8121$ | $16.8973 \pm 0.9948$ |
| | 10 | $8.5245 \pm 0.6450$ | $11.1186 \pm 0.6534$ |
| | 100 | $4.1066 \pm 0.3086$ | $4.9072 \pm 0.2915$ |
| $(1, 0)$ | 1 | $20.3536 \pm 1.5899$ | $60.8843 \pm 3.9302$ |
| | 2 | $15.7742 \pm 1.1840$ | $33.6515 \pm 2.0373$ |
| | 5 | $10.9528 \pm 0.8513$ | $16.9305 \pm 0.9957$ |
| | 10 | $8.3487 \pm 0.6646$ | $11.1186 \pm 0.6534$ |
| | 100 | $4.0654 \pm 0.2472$ | $4.9040 \pm 0.2953$ |
| $(5, 0)$ | 1 | $20.2702 \pm 1.6145$ | $60.9139 \pm 3.9326$ |
| | 2 | $15.7136 \pm 1.0099$ | $33.6644 \pm 2.0381$ |
| | 5 | $10.9100 \pm 0.8609$ | $16.9343 \pm 0.9957$ |
| | 10 | $8.3181 \pm 0.6738$ | $11.1191 \pm 0.6540$ |
| | 100 | $4.0627 \pm 0.2424$ | $4.9035 \pm 0.2960$ |
| $(10, 0)$ | 1 | $20.2592 \pm 1.6179$ | $28.0073 \pm 2.0546$ |
| | 2 | $15.7055 \pm 1.2136$ | $33.6662 \pm 2.0386$ |
| | 5 | $10.9050 \pm 0.8725$ | $16.9346 \pm 0.9959$ |
| | 10 | $8.3137 \pm 0.6757$ | $11.1195 \pm 0.6537$ |
| | 100 | $4.0623 \pm 0.2417$ | $4.9036 \pm 0.2958$ |
| $(2, -1)$ | 1 | $20.2091 \pm 1.6384$ | $27.2920 \pm 2.1314$ |
| | 2 | $15.6265 \pm 1.2344$ | $33.6569 \pm 2.0452$ |
| | 5 | $10.8407 \pm 0.8841$ | $16.9230 \pm 0.9954$ |
| | 10 | $8.2607 \pm 0.6844$ | $11.1092 \pm 0.6534$ |
| | 100 | $4.0573 \pm 0.2324$ | $4.8908 \pm 0.3126$ |

### 5.1. Nonlocal Denoising Approach

We assume that each pixel $i = (i_1, i_2) \in G$ is affected by the noise in an independent and identical way, and model the image pixelwise as

$$f_i = u + \frac{\sigma \eta}{\sqrt{\nu}}, \quad \eta \sim \mathcal{N}(0, 1), \ t \sim \chi^2_\nu, \ i \in G,$$
patches of an image can be found several times in the image. Then, the set $S(i)$ constitutes of the indices of the centers of patches that are similar to the patch centered at $i \in \mathcal{G}$. This requires the selection of the patch size and an appropriate similarity measure, which need to be adapted to the noise statistic and the noise level and are detailed later on. Based on the similarity measure, we take as the set $S(i)$ the indices of the centers of the $K$ most similar patches. In order to avoid a computational overload one typically restricts the search zone for similar patches to a $w \times w$ search window around $i \in \mathcal{G}$. Here and in all subsequent cases we extend the image by mirroring at the boundary.

Assuming the pixels of an image to be independent is in practice a rather unrealistic assumption; in fact, in natural images they are locally usually highly correlated. Taking the local dependence structure into account may improve the results of image restoration methods, which motivates to take whole patches (and not only their centers) and estimate their parameters. This can be achieved by applying the GMMF Algorithm 1 with similar patches used as samples. The resulting $\hat{\mu}$ yields the estimated patch values, while the correlation is encoded in the scatter matrix $\hat{\Sigma}$. Proceeding as above gives multiple estimates for each image pixel that are averaged in the end to obtain the final image.

The selection of similar patches constitutes a fundamental step in our nonlocal denoising approach. At this point, the question arises how to compare noisy patches and numerical examples show that an adaptation of the similarity measure to the noise distribution is essential for a robust similarity evaluation. In [3], the authors formulated the similarity between patches as a statistical hypothesis testing problem and proposed among other criteria a similarity measure based on a generalized likelihood test, which we use in the following. Details on this approach can be also found in [11]. In case of the Student-$t$ distribution, the similarity measure between two patches $p = (p_1, \ldots, p_t)$ and $q = (q_1, \ldots, q_t)$ can be computed as

$$S(p, q) = \prod_{i=1}^{t} \left(1 + \frac{(p_i - q_i)^2}{2\sigma} \right)^{-\left(\nu+1\right)}.$$  

In practice, we take the logarithm of $S$ in order to avoid numerical instabilities, resulting in the distance measure

$$d(p, q) = \sum_{i=1}^{t} \log \left(1 + \left(\frac{p_i - q_i}{2\sigma} \right)^2 \right).$$

### 5.2. Cauchy Noise

As mentioned in the introduction, the initial motivation for this work was the consideration of Cauchy noise in [17, 11] and thus we tested our approach on images corrupted by additive Cauchy noise ($\nu = 1$) with noise level $\sigma = 10$. Since the noise level is very high, we chose $n = 50$ patches of size $5 \times 5$ for the denoising. It turns out that the differences in terms of
PSNR or SSIM compared to the current state of the art method [11] are small and nearly not visible in images with much textured regions. However, the improvement is large in images with many constant or smoothly varying areas. This becomes in particular apparent in case of the test image given in Figure 2. Here, the top row displays the original image (left) together with its noisy version (right), which is corrupted by additive Cauchy noise ($\nu = 1$) with noise level $\sigma = 10$. The bottom row shows from left to right the results obtained using the variational method presented in [17], the pixelwise [11] and the patchwise nonlocal myriad filter. While in case of the variational method some of the outliers remain, the result of [11] is rather grainy, which is much improved by our new approach. This is also reflected in the corresponding PSNR and SSIM values stated in the captions of the figure.

5.3. Projected Normal and Wrapped Cauchy Distribution

In our second example, we consider the ML estimation of the scatter matrix $\Sigma$ in the case $\nu = 0$, which is related to the projected normal distribution, see Remark 3.9. For $d = 2$, there is a further relation to the wrapped Cauchy distribution [15]. Since we did not find a reference that provides further details on the relation between the projected normal and the wrapped Cauchy distribution we elaborate it in Appendix A.2. Recall the density of Cauchy $C(a,\gamma)$ distribution with parameters $a \in \mathbb{R}$ and $\gamma > 0$,

$$f(\vartheta|a,\gamma) := \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (\vartheta - a)^2}.$$  

The wrapped Cauchy distribution is obtained by wrapping the Cauchy distribution around the circle, i.e.

$$f_w(\vartheta|a,\gamma) := \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (\vartheta + 2k\pi - a)^2}, \quad \vartheta \in [-\pi, \pi)$$

$$= \frac{1}{2\pi} \frac{1 - \rho^2}{1 + \rho^2 - 2\rho \cos(\vartheta - a)} = \frac{1}{2\pi} \frac{\sinh(\gamma)}{\cosh(\gamma) - \cos(\vartheta - a)},$$

where $\rho := e^{-\gamma}$. We rewrite the density as follows

$$f_w(\vartheta) = \frac{1}{2\pi} \frac{1 - \rho^2}{1 + \rho^2 - 2\rho \cos(\vartheta - a)} = \frac{1}{2\pi} \frac{1 - \rho^2}{1 + \rho^2 - \frac{2\rho}{1+\rho^2}(\cos(a) \cos(\vartheta) + \sin(a) \sin(\vartheta))}$$

$$= \frac{1}{2\pi} \frac{1}{1 + \rho^2 - \frac{2\rho}{1+\rho^2}(\cos(a) \cos(\vartheta) + \sin(a) \sin(\vartheta))}.$$

Setting $\xi_1 = \frac{2\rho}{1+\rho^2} \cos(a)$, $\xi_2 = \frac{2\rho}{1+\rho^2} \sin(a)$ and noting that

$$\frac{1 - \rho^2}{1 + \rho^2} = \sqrt{1 - \xi_1^2 - \xi_2^2}$$

30
Figure 2: Denoising of the test image (top left) corrupted with additive Cauchy noise ($\nu = 1$, $\sigma = 10$) (top right) using the methods proposed in [17] (bottom left), [11] (bottom middle) and our NGMMF (bottom right).
we obtain
\[
f_w(\thet) = \frac{1}{2\pi} \frac{\sqrt{1 - \xi_1^2 - \xi_2^2}}{1 - \xi_1 \cos(\thet) - \xi_2 \sin(\thet)}.
\]

Using this reparametrization, there is a close relation between wrapped Cauchy and projected normal distribution which is detailed in Appendix A.2, see also [15]. Based on Proposition A.6 we can use Algorithm 1 with \( \mu_r \equiv \mu = 0 \) being fixed to do ML estimation for the wrapped Cauchy distribution. The reformulation in terms of \( \xi_1 \) and \( \xi_2 \) and is given in Algorithm 3.

At this point, to have a full equivalence between the parameters \( \Sigma \) respective \( a \) and \( \rho \) the additional variable \( \sqrt{1 - \xi_1^2 - \xi_2^2} \) is needed. It can be ignored if one is only interested in \( \xi_1 \) and \( \xi_2 \).

**Algorithm 3** ML estimation for the wrapped Cauchy distribution

**Input:** \( \vartheta_1, \ldots, \vartheta_n \in S^1, \ n \geq 3, \ 0 < w_i < \frac{1}{2}, \ i = 1, \ldots, n, \ \sum_{i=1}^{n} w_i = 1 \)

**Initialization:** \( \xi_1, 0 = \xi_2, 0 = 0 \)

for \( r = 0, \ldots, \)

\[
\begin{align*}
\xi_{1,r+1} &= \frac{\sum_{i=1}^{n} w_i 1 - \xi_1 - \xi_2 \cos(\vartheta_i) - \xi_2 \sin(\vartheta_i)}{\sum_{i=1}^{n} w_i 1 - \xi_1 - \xi_2 \cos(\vartheta_i) - \xi_2 \sin(\vartheta_i)} \cos(\vartheta_i) \\
\xi_{2,r+1} &= \frac{\sum_{i=1}^{n} w_i 1 - \xi_1 - \xi_2 \cos(\vartheta_i) - \xi_2 \sin(\vartheta_i)}{\sum_{i=1}^{n} w_i 1 - \xi_1 - \xi_2 \cos(\vartheta_i) - \xi_2 \sin(\vartheta_i)} \sin(\vartheta_i) \\
t_{r+1} &= 2 \sum_{i=1}^{n} w_i \sqrt{1 - \xi_1^2 - \xi_2^2} \frac{1 - \xi_1 - \xi_2 \cos(\vartheta_i) - \xi_2 \sin(\vartheta_i)}{1 - \xi_1 \cos(\vartheta_i) - \xi_2 \sin(\vartheta_i)}
\end{align*}
\]

We apply Algorithm 3 to denoise \( S^1 \)-valued images corrupted by wrapped Cauchy noise,

\[
f_i = (u_i + \gamma \eta) \bmod 2\pi, \quad \eta \sim C(0, \gamma), \ \gamma > 0, \ i \in \mathcal{G},
\]

where we chose \( \gamma = 0.1 \), which yields \( \rho = e^{-\gamma} \approx 0.9048 \). The original image as well as the noisy image are given in the top row of Figure 3. The similarity measure to find similar patches is given by

\[
S(p, q) = \frac{(1 - \rho)^4}{(1 + \rho^2 - 2\rho \cos (\frac{p_i - q_i}{2}))^2},
\]
which leads to the distance
\[ d(p, q) = \sum_{i=1}^{t} \log \left( 1 + \left( \frac{p_i - q_i}{2\sigma} \right)^2 \right). \]

We again chose \( n = 50 \) patches of size \( 5 \times 5 \) as samples, but this time extracted only their centers, estimated the parameters and restored the image pixelwise. We compare our approach with the variational method using a first and second order TV-regularizer given in [2] and the nonlocal denoising algorithm based on second order statistic [10]. The results together with the mean-squared reconstruction error are given in the bottom row of Figure 3. Both the variational as well as the second order statistical method cannot cope with the impulsiveness of the wrapped Cauchy noise such that several wrong pixels remain, which is in particular visible in the background. Furthermore, the edges of the color squares and the transitions in the ellipse and in circle are rather fringy. On the contrary, our method restores the image very well, if at all a slight grain can be observed in the color squares which is due to the pixelwise denoising that does not regard neighboring pixels appropriately.

6. Conclusion

We introduced a generalized multivariate myriad filter based on the parameter estimation of the multivariate Student-\( t \) distribution by the weighted maximum likelihood approach. We proposed an efficient algorithm for its computation and illustrated its usage in a nonlocal denoising approach. There are different directions for future work: First, we would like extend our analysis to the case that additionally the degrees of freedom parameter \( \nu \) is unknown and needs to be estimated. Although an EM algorithm has already been derived for this case in [9], there does not exist any result concerning existence and/or uniqueness of the joint ML estimator. Second, it would be interesting to examine whether our approach can be generalized to Student-\( t \) mixture models.

Concerning our denoising approach, fine tuning steps as discussed in [13] such as aggregations of patches [24], the use of an oracle image or a variable patch size to better cope with textured and homogeneous image regions may improve the denoising results. Further, in all our examples we used uniform weights, but weights based for instance on spatial distance or similarity would make sense as well. Another question is how to incorporate linear operators (blur, missing pixels) into the image restoration. Finally, it would be interesting if other types of impulsive noise or spatially varying noise can be treated. Here, first experiments indicate that our filter performs well also for other noise scenarios such as Salt-and-Pepper plus Gaussian noise.
Figure 3: Denoising of an $S^1$-valued image (top left) corrupted with additive wrapped Cauchy noise ($\alpha = 0, \rho = 0.1$) (top right) using the methods proposed in [2] (bottom left), [10] (bottom middle) and our NGMMF (bottom right).
Acknowledgments

Funding by the German Research Foundation (DFG) within the Research Training Group 1932, project area P3, is gratefully acknowledged. Furthermore, we wish to thank Julie Delon for the careful revision of our paper [11] which served as a motivation for this work.

A. Appendix

A.1. Uniqueness of Scatter

We provide an alternative argument for uniqueness of the scatter matrix based on the Hessian $\nabla^2 L_{\nu,0}$ of $L_{\nu,0}$. In order to compute the Hessian of $L_{\nu,0}$ we need the following lemma.

Lemma A.1. For $a \in \mathbb{R}^d$, $c \geq 0$ with $c + \|a\|^2_2 > 0$, the function $F: \text{SPD}(d) \rightarrow \text{Sym}(d)$,

$$F(X) = -\frac{X^{-1}aa^TX^{-1}}{c+a^TX^{-1}a},$$

has the derivative

$$D(F)(X)[H] = \frac{X^{-1}HX^{-1}aa^TX^{-1} - HX^{-1}aa^TX^{-1}X^{-1}X^{-1} - a}{(c+a^TX^{-1}a)^2}.$$

Further, for $a \neq 0$ and $H \in \text{Sym}(d) \setminus \{0\}$ it holds

$$\langle D(f)(X)[H], H \rangle = \frac{2}{c+a^TX^{-1}a}a^TX^{-1}HX^{-1}H^{-1}X^{-1}a - \frac{1}{(c+a^TX^{-1}a)^2}(a^TX^{-1}HX^{-1}a)^2 > 0.$$

Proof. We write $f(X) = F_2(F_1(X))F_3(X)$, where

$$F_1(X) = X^{-1}, \quad DF_1(X)[H] = -X^{-1}HX^{-1},$$
$$F_2(X) = -Xaa^TX, \quad DF_2(X)[H] = -(Haa^TX + Xaa^TH),$$
$$F_3(X) = \frac{1}{c+a^TX^{-1}a}, \quad DF_3(X)[H] = \frac{1}{(c+a^TX^{-1}a)^2}a^TX^{-1}HX^{-1}a.$$

Then, with the help of chain and product rule, we compute

$$D(f)(X)[H] = D(F_2(F_1(X)))[H]F_3(X) + F_2(F_1(X))DF_3(X)[H]$$
$$= D(F_2(F_1(X))) \circ DF_1(X)[H]F_3(X) + F_2(F_1(X))DF_3(X)[H]$$
$$= D(F_2(F_1(X)))[X^{-1}HX^{-1}]F_3(X) + F_2(F_1(X))DF_3(X)[H]$$
$$= \frac{X^{-1}HX^{-1}aa^TX^{-1} - X^{-1}aa^TX^{-1}X^{-1}X^{-1} - a}{(c+a^TX^{-1}a)^2}.$$

35
As a consequence, we have
\[
\langle D(f)(X)[H], H \rangle = \frac{1}{c + a^T X^{-1} a} \left( \langle (X^{-1} H X^{-1} a a^T X^{-1} H X^{-1}), H \rangle + \langle X^{-1} a a^T X^{-1} H X^{-1}, H \rangle \right)
- \frac{1}{(c + a^T X^{-1} a)^2} \langle (X^{-1} a a^T X^{-1} a X^{-1} H X^{-1}), H \rangle
- \frac{1}{c + a^T X^{-1} a} \left( \langle a^T X^{-1} H X^{-1} a + a^T X^{-1} H X^{-1} a, H \rangle \right)
- \frac{1}{(c + a^T X^{-1} a)^2} \langle (a^T X^{-1} H X^{-1} a)(a^T X^{-1} H X^{-1} a), H \rangle
= 2 \frac{a^T X^{-1} H X^{-1} a}{c + a^T X^{-1} a} - \frac{(a^T X^{-1} H X^{-1} a)^2}{(c + a^T X^{-1} a)^2}.
\]

Setting \( b = X^{-\frac{1}{2}} a \) and \( Y = X^{-\frac{1}{2}} H X^{-\frac{1}{2}} \), this can be written as
\[
\langle D(f)(X)[H], H \rangle = 2 \frac{b^T Y b}{c + \|b\|^2} - \frac{(b^T Y b)^2}{(c + \|b\|^2)^2} = 2 \frac{\|Y b\|^2}{c + \|b\|^2} - \frac{(b^T Y b)^2}{(c + \|b\|^2)^2}.
\]

By Cauchy-Schwarz’ inequality it holds \((b^T Y b)^2 \leq \|b\|^2 \|Y b\|^2\) with equality if and only if \( b = \lambda Y b \) for some \( \lambda \neq 0 \), so that we can estimate
\[
\langle D(f)(X)[H], H \rangle \geq 2 \frac{\|Y b\|^2}{c + \|b\|^2} - \frac{\|b\|^2 \|Y b\|^2}{(c + \|b\|^2)^2}
\geq \frac{\|Y b\|^2 (2c + \|b\|^2)}{(c + \|b\|^2)^2} > 0
\]
with equality if and only if \( Y b = 0 = b \), i.e. \( a = 0 \). \( \square \)

With the help of Lemma A.1 we show that \( \nabla^2 L_{\nu,0} \) is positive (semi-)definite at critical points.

**Theorem A.2.** Let \( \hat{\Sigma} \) be a critical point of \( L_{\nu,0} \). Then, the Hessian \( \nabla^2 L_{\nu,0}(\hat{\Sigma}) \) is positive definite for any \( \nu > 0 \) and positive semi-definite for \( \nu = 0 \) so that any critical point is a (strict) minimizer.

**Proof.** First, using Lemma A.1 for \( a = x_i \) and \( c = \nu \) (and the derivative of \( \Sigma^{-1} \)), we compute the Hessian of \( L_{\nu,0} \) as
\[
\langle \nabla^2 L_{\nu,0}(\Sigma)[H], H \rangle = (d+\nu) \sum_{i=1}^n w_i \left( 2x_i^T \Sigma^{-1} H \Sigma^{-1} \Sigma^{-1} x_i - \frac{(x_i^T \Sigma^{-1} H \Sigma^{-1} x_i)^2}{\nu + x_i^T \Sigma^{-1} x_i} \right) - \langle \Sigma^{-1} H \Sigma^{-1}, H \rangle.
\]

36
Set $A = \Sigma^{-\frac{1}{2}}H\Sigma^{-\frac{1}{2}}$ and $y_i = \Sigma^{-\frac{1}{2}}x_i$, then

$$\langle \nabla^2 L_{\nu,0}(\hat{\Sigma})[H], H \rangle = (d + \nu) \sum_{i=1}^{n} w_i \left( \frac{2y_i^TA^2y_i}{\nu + \|y_i\|^2_2} - \frac{(y_i^T Ay_i)^2}{(\nu + \|y_i\|^2_2)^2} \right) - (A, A)$$

$$= (d + \nu) \sum_{i=1}^{n} w_i \left( \frac{2\|Ay_i\|^2_2}{\nu + \|y_i\|^2_2} - \frac{(y_i^T Ay_i)^2}{(\nu + \|y_i\|^2_2)^2} \right) - \|A\|^2_F$$ (17)

Now, since $\hat{\Sigma}$ is a critical point of $L_{\nu,0}$, it holds according to (2)

$$(d + \nu) \sum_{i=1}^{n} w_i \frac{y_i^Ty_i^T}{\nu + \|y_i\|^2_2} = I,$$

and multiplying both sides from the left and right with $A$ we obtain

$$(d + \nu) \sum_{i=1}^{n} w_i \frac{Ay_iy_i^TA}{\nu + \|y_i\|^2_2} = A^2.$$

Taking the trace results in

$$\|A\|^2_F = (d + \nu) \sum_{i=1}^{n} w_i \frac{\|Ay_i\|^2_2}{\nu + \|y_i\|^2_2} = (d + \nu) \sum_{i=1}^{n} w_i \frac{\|Ay_i\|^2_2}{\nu + \|y_i\|^2_2}$$

and plugging this into (17) we have

$$\langle \nabla^2 L_{\nu,0}(\hat{\Sigma})[H], H \rangle = (d + \nu) \sum_{i=1}^{n} w_i \left( \frac{\|Ay_i\|^2_2}{\nu + \|y_i\|^2_2} - \frac{(y_i^T Ay_i)^2}{(\nu + \|y_i\|^2_2)^2} \right).$$

With the help of the Cauchy-Schwarz inequality we estimate $(y_i^T Ay_i)^2 \leq \|y_i\|^2_2 \|Ay_i\|^2_2$, and therewith

$$\langle \nabla^2 L_{\nu,0}(\hat{\Sigma})[H], H \rangle \geq (d + \nu) \sum_{i=1}^{n} w_i \left( \frac{\|Ay_i\|^2_2}{\nu + \|y_i\|^2_2} - \frac{\|y_i\|^2_2 \|Ay_i\|^2_2}{(\nu + \|y_i\|^2_2)^2} \right)$$

$$= (d + \nu) \sum_{i=1}^{n} w_i \left( \frac{\|Ay_i\|^2_2 (\nu + \|y_i\|^2_2) - \|y_i\|^2_2 \|Ay_i\|^2_2}{(\nu + \|y_i\|^2_2)^2} \right)$$

$$= (d + \nu) \sum_{i=1}^{n} w_i \frac{\nu \|Ay_i\|^2_2}{(\nu + \|y_i\|^2_2)^2} \geq 0,$$

where the inequality is strict for $\nu > 0$.

In order to show that $L_{\nu,0}$ possesses only one critical point we make use of the *Mountain Pass Theorem*. To formulate it, we first introduce a notion going back to works of Palais.
and Smale [21, 20, 28].

**Definition A.1.** Let \( H \) be a Hilbert space and \( J \) be a real-valued functional on \( H \).

(i) A sequence \( \{x_n\}_{n\in\mathbb{N}} \subseteq H \) is called a Palais-Smale sequence for \( J \), if \( \{J(x_n)\}_{n\in\mathbb{N}} \) is uniformly bounded in \( n \) and \( \|\nabla J(x_n)\| \to 0 \) as \( n \to \infty \).

(ii) A continuously differentiable, real-valued functional \( J \in C^1(H, \mathbb{R}) \) on a Hilbert space \( H \) satisfies the Palais-Smale condition, if every Palais-Smale sequence has a convergent subsequence in \( H \).

A simple criterion that implies the Palais-Smale condition is given in the next lemma.

**Lemma A.3.** Let \( J \in C^1(H, \mathbb{R}) \) be a functional such that \( |J| \) or \( \|\nabla J\|_2 \) (or equivalently, \( |J| + \|\nabla J\|_2 \)) is coercive. Then, \( J \) fulfills the Palais-Smale condition.

**Proof.** Let \( |J| + \|\nabla J\|_2 \) be coercive and \( \{x_n\}_{n\in\mathbb{N}} \subseteq H \) be a Palais-Smale sequence. Then, \( \{x_n\}_{n\in\mathbb{N}} \) has to be bounded, and consequently, according to the Bolzano-Weierstraß theorem it possesses a convergent subsequence. \( \square \)

The Mountain Pass Theorem guarantees the existence of critical points and is an important tool for proving existence of solutions to both ordinary and partial differential equations. It can be found for instance in [5, Section 8.5].

**Theorem A.4 (Mountain Pass Theorem).** Let \( J \in C^1(H, \mathbb{R}) \) be a functional satisfying the Palais-Smale condition. Assume further

(i) \( J(0) = 0 \),

(ii) there exist constants \( r, a > 0 \) such that \( J(x) \geq a \) if \( \|x\| = r \),

(iii) there exists an element \( y \in H \) with \( \|y\| > r \) and \( J(y) \leq 0 \).

Define

\[
\Gamma := \{ \gamma \in C([0, 1], H) : \gamma(0) = 0, \gamma(1) = y \}.
\]

Then,

\[
e = \inf_{\gamma \in \Gamma} \max_{t \in [0,1]} J(\gamma(t))
\]

is a critical value of \( J \).

In order to apply the Mountain Pass Theorem A.4 we extend the functional \( L_{\nu,0} \) for fixed \( \nu > 0 \) to the Hilbert space \( \text{Sym}(d) \) by

\[
L_{\nu,0}(\Sigma) = \begin{cases} 
(d + \nu) \sum_{i=1}^{n} w_i \log \left( \nu + x_i^T \Sigma^{-1} x_i \right) + \log(|\Sigma|) & \text{if } \Sigma \in \text{SPD}(d), \\
+\infty & \text{otherwise}.
\end{cases}
\]
Theorem A.5. Let \( x_1, \ldots, x_n \in \mathbb{R}^d \) fulfill Assumption 3.1(i). For fixed \( \nu > 0 \), the functional \( L_{\nu,0} \) admits a unique local minimum, which is also a global one.

Proof. Assume towards a contradiction that \( L_{\nu,0} \) admits at least two strict local minimizer \( \Sigma_1 \) and \( \Sigma_2 \) for some fixed \( \nu > 0 \), where w.l.o.g. \( \infty > L_{\nu,0}(\Sigma_1) \geq L_{\nu,0}(\Sigma_2) \). We define the functional

\[
J(X) = L_{\nu,0}(X + \Sigma_1) - L_{\nu,0}(\Sigma_1).
\]

Then \( J \) is coercive, so it fulfills the Palais-Smale condition. Further, it holds \( J(0) = 0 \), and since \( \Sigma_1 \) is a strict local minimum, there exists \( \varepsilon > 0 \) such that

\[
L_{\nu,0}(\Sigma_1) < L_{\nu,0}(X) \quad \text{for all } X \in \overline{B_\varepsilon(\Sigma_1)}.
\]

Thus, if we define

\[
a := \min_{X \in \partial B_\varepsilon(\Sigma_1)} \{ L_{\nu,0}(X) - L_{\nu,0}(\Sigma_1) \},
\]

and set \( r = \varepsilon \), we have for all \( X \in \text{Sym}(d) \) with \( \|X - \Sigma_1\|_F = r \)

\[
J(X - \Sigma_1) = L_{\nu,0}(X) - L_{\nu,0}(\Sigma_1) \geq a > 0.
\]

Furthermore, since \( L_{\nu,0}(\Sigma_2) \leq L_{\nu,0}(\Sigma_1) \), \( \Sigma_2 \notin \overline{B_\varepsilon(\Sigma_1)} \) so that \( \|\Sigma_2 - \Sigma_1\| > \varepsilon \) and

\[
J(\Sigma_2 - \Sigma_1) = L_{\nu,0}(\Sigma_2) - L_{\nu,0}(\Sigma_1) \leq 0.
\]

According to Theorem A.4, the functional \( J \) admits a critical point which is not a minimum, a contradiction to Lemma A.2.

\[ \square \]

A.2. Wrapped Cauchy and Projected Normal Distributions

In the following we detail the relation between the two-dimensional projected normal distribution and the wrapped Cauchy distribution.

Proposition A.6. Let \( \vartheta \sim \Pi_N(\Sigma) \) be a random variable following a projected normal distribution, then \( \varphi = (2\vartheta) \mod 2\pi \sim C_w(a, \rho) \) has a wrapped Cauchy distribution with parameters

\[
a = \arctan \left( \frac{2\sigma_{12}}{\sigma_{11} - \sigma_{22}} \right) \quad \text{and} \quad \rho = \left( \frac{\text{tr}(\Sigma) - 2\sqrt{\Sigma}}{\text{tr}(\Sigma) + 2\sqrt{\Sigma}} \right)^{\frac{1}{2}}.
\]

Conversely, a projected normal distributed random variable can be obtained by a random
phase unwrapping’ as follows: Let \( \vartheta \sim C_w(a, \rho) \) and set
\[
\varphi = \frac{\vartheta}{2} - \pi \text{sgn}(\vartheta) \xi,
\]
where \( \xi \) is a Bernoulli random variable with \( \mathbb{P}(\xi = 0) = \mathbb{P}(\xi = 1) = \frac{1}{2} \) that is independent from \( \vartheta \).

**Proof.** We show that the densities of the corresponding random variables coincide. To this aim, we parametrize \( x \in S^1 \) as \( x = \begin{pmatrix} \cos(\vartheta) \\ \sin(\vartheta) \end{pmatrix} \) and let
\[
\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}, \quad \Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{pmatrix}.
\]
We compute
\[
x^T \Sigma^{-1} x = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \left( \sigma_{22}x_1^2 - 2\sigma_{12}x_1x_2 + \sigma_{11}x_2^2 \right)
\]
\[
= \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \left( \sigma_{22} \left( \frac{1}{2} + \frac{1}{2} \cos(2\vartheta) \right) + \sigma_{11} \left( \frac{1}{2} - \frac{1}{2} \cos(2\vartheta) \right) - \sigma_{12} \sin(2\vartheta) \right)
\]
\[
= \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \left( \frac{1}{2}(\sigma_{11} + \sigma_{22}) - \frac{1}{2}(\sigma_{11} - \sigma_{22}) \cos(2\vartheta) - \sigma_{12} \sin(2\vartheta) \right)
\]
\[
= \frac{1}{2}(\sigma_{11} + \sigma_{22}) \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \left( 1 - \frac{\sigma_{11} - \sigma_{22}}{\sigma_{11} + \sigma_{22}} \cos(2\vartheta) - \frac{2\sigma_{12}}{\sigma_{11} + \sigma_{22}} \sin(2\vartheta) \right)
\]
\[
= \frac{1}{2} \text{tr}(\Sigma) \left( 1 - \frac{\sigma_{11} - \sigma_{22}}{\sigma_{11} + \sigma_{22}} \cos(2\vartheta) - \frac{2\sigma_{12}}{\sigma_{11} + \sigma_{22}} \sin(2\vartheta) \right)
\]
and identifying \( \xi_1 = \frac{\sigma_{11} - \sigma_{22}}{\sigma_{11} + \sigma_{22}} \) and \( \xi_2 = \frac{2\sigma_{12}}{\sigma_{11} + \sigma_{22}} \) such that
\[
\sqrt{1 - \xi_1^2 - \xi_2^2} = \frac{2\sqrt{|\sigma_{11}\sigma_{22} - \sigma_{12}^2|}}{\sigma_{11} + \sigma_{22}} = \frac{\sqrt{|\Sigma|}}{2 \text{tr}(\Sigma)}
\]
yields the first claim. Observe that the wrapped Cauchy distribution is unimodular, whereas the projected normal distribution is antipodally symmetric, which causes the mod operation.

Concerning the second statement, observe that
\[
\mathbb{P}(\text{sgn}(\vartheta) = 1) = \frac{\pi - a}{2\pi}, \quad \mathbb{P}(\text{sgn}(\vartheta) = 0) = 0, \quad \mathbb{P}(\text{sgn}(\vartheta) = -1) = \frac{\pi + a}{2\pi}.
\]
Define \( Y = -\text{sgn}(\vartheta)\xi \pi \), then we may write \( \varphi = \frac{\vartheta}{2} + Y \) with

\[
\begin{align*}
\mathbb{P}(Y = \pi) &= \mathbb{P}(\text{sgn}(\vartheta) = -1, \xi = 1) = \frac{\pi + a}{4\pi}, \\
\mathbb{P}(Y = 0) &= \mathbb{P}(\xi = 0) = \frac{1}{2}, \\
\mathbb{P}(Y = -\pi) &= \mathbb{P}(\text{sgn}(\vartheta) = 1, \xi = 1) = \frac{\pi - a}{4\pi}.
\end{align*}
\]

Now, the density of \( \varphi \) is given by

\[
f_\varphi(u) = f_{\vartheta/2} \ast f_Y(u) = \int 2f_\vartheta(2v)f_Y(u-v)\,dv
\]

\[
= 2f_\vartheta(2(u-\pi)) \frac{\pi + a}{4\pi} + 2f_\vartheta(2u) \frac{1}{2} + 2f_\vartheta(2(u+\pi)) \frac{\pi - a}{4\pi}
\]

\[
= 2f_\vartheta(2u) = f_{\vartheta/2}(u),
\]

which finishes the proof.

References

[1] A. Banerjee and P. Maji. Spatially constrained Student’s \( t \)-distribution based mixture model for robust image segmentation. *Journal of Mathematical Imaging Vision*, 60(3):355–381, 2018.

[2] R. Bergmann, F. Laus, G. Steidl, and A. Weinmann. Second order differences of cyclic data and applications in variational denoising. *SIAM Journal on Imaging Sciences*, 7(4):2916–2953, 2014.

[3] C.-A. Deledalle, L. Denis, and F. Tupin. How to compare noisy patches? patch similarity beyond Gaussian noise. *International Journal of Computer Vision*, 99(1):86–102, 2012.

[4] P. Ding. On the conditional distribution of the multivariate \( t \) distribution. *The American Statistician*, 70(3):293–295, 2016.

[5] L. Evans. *Partial Differential Equations*, volume 19 of *Graduate studies in mathematics*. American Mathematical Society, Providence, 1998.

[6] D. Gerogiannis, C. Nikou, and A. Likas. The mixtures of Student’s \( t \)-distributions as a robust framework for rigid registration. *Image and Vision Computing*, 27(9):1285–1294, 2009.

[7] J. T. Kent and D. E. Tyler. Redescending \( M \)-estimates of multivariate location and scatter. *The Annals of Statistics*, pages 2102–2119, 1991.
[8] S. Kotz and S. Nadarajah. *Multivariate t-Distributions and Their Applications*. Cambridge University Press, 2004.

[9] K. L. Lange, R. J. Little, and J. M. Taylor. Robust statistical modeling using the $t$ distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

[10] F. Laus, M. Nikolova, J. Persch, and G. Steidl. A nonlocal denoising algorithm for manifold-valued images using second order statistics. *SIAM Journal on Imaging Sciences*, 10(1):416–448, March 2017.

[11] F. Laus, F. Pierre, and G. Steidl. Nonlocal myriad filters for Cauchy noise removal. *Journal of Mathematical Imaging and Vision*, 60(8):1324–1354, 2018.

[12] M. Lebrun, A. Buades, and J.-M. Morel. A nonlocal Bayesian image denoising algorithm. *SIAM Journal on Imaging Sciences*, 6(3):1665–1688, 2013.

[13] M. Lebrun, M. Colom, A. Buades, and J. Morel. Secrets of image denoising cuisine. *Acta Numerica*, 21:475–576, 2012.

[14] C. Liu and D. B. Rubin. ML estimation of the $t$ distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, pages 19–39, 1995.

[15] K. V. Mardia and P. E. Jupp. *Directional Statistics*, volume 494. John Wiley & Sons, 2009.

[16] R. A. Maronna. Robust $M$-estimators of multivariate location and scatter. *Annals in Statistics*, 4(1):51–67, 01 1976.

[17] J.-J. Mei, Y. Dong, T.-Z. Huang, and W. Yin. Cauchy noise removal by nonconvex ADMM with convergence guarantees. *Journal of Scientific Computing*, 74(2):743–766, 2018.

[18] S. Nadarajah and S. Kotz. Estimation methods for the multivariate $t$-distribution. *Acta Applicandae Mathematicae*, 102(1):99–118, 2008.

[19] T. M. Nguyen and Q. J. Wu. Robust Student’s-$t$ mixture model with spatial constraints and its application in medical image segmentation. *IEEE Transactions on Medical Imaging*, 31(1):103–116, 2012.

[20] R. S. Palais. Morse theory on Hilbert manifolds. *Topology*, 2(4):299–340, 1963.

[21] R. S. Palais and S. Smale. A generalized Morse theory. *Bulletin of the American Mathematical Society*, 70(1):165–172, 1964.
[22] D. Peel and G. J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, Oct 2000.

[23] K. B. Petersen, M. S. Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7:15, 2008.

[24] A. Saint-Dizier, J. Delon, and C. Bouveyron. A unified view on patch aggregation. hal-preprint hal-01865340.

[25] F. Sciacchitano, Y. Dong, and T. Zeng. Variational approach for restoring blurred images with Cauchy noise. *SIAM Journal on Imaging Sciences*, 8(3):1894–1922, 2015.

[26] S. Setzer, G. Steidl, and T. Teuber. On vector and matrix median computation. *Journal of Computational and Applied Mathematics*, 236:2200–2222, 2012.

[27] G. Sfikas, C. Nikou, and N. Galatsanos. Robust image segmentation with mixtures of student’s t-distributions. In *2007 IEEE International Conference on Image Processing*, volume 1, pages I – 273–I – 276, 2007.

[28] S. Smale. Morse theory and a non-linear generalization of the Dirichlet problem. In *The Collected Papers of Stephen Smale: Volume 2*, pages 511–525. World Scientific, 2000.

[29] C. Sutour, C.-A. Deledalle, and J.-F. Aujol. Estimation of the noise level function based on a nonparametric detection of homogeneous image regions. *SIAM Journal on Imaging Sciences*, 8(4):2622–2661, 2015.

[30] A. Van Den Oord and B. Schrauwen. The Student-t mixture as a natural image patch prior with application to image compression. *Journal of Machine Learning Research*, 15(1):2061–2086, 2014.

[31] E. Weiszfeld. Sur le point pour lequel les sommes des distances de n points donnés et minimum. *Tôhoku Mathematical Journal*, 43:355 – 386, 1937.

[32] M. Welk, J. Weickert, F. Becker, C. Schnörr, C. Feddern, and B. Burgeth. Median and related local filters for tensor-valued images. *Signal Processing*, 87(2):291 – 308, 2007.

[33] Z. Zhou, J. Zheng, Y. Dai, Z. Zhou, and S. Chen. Robust non-rigid point set registration using Student’s-t mixture model. *PloS one*, 9(3):e91381, 2014.