A Survey on Surrogate Approaches to Non-negative Matrix Factorization

Pascal Fernsel · Peter Maass

Abstract Motivated by applications in hyperspectral imaging we investigate methods for approximating a high-dimensional non-negative matrix $Y$ by a product of two lower-dimensional, non-negative matrices $K$ and $X$. This so-called non-negative matrix factorization is based on defining suitable Tikhonov functionals, which combine a discrepancy measure for $Y \approx KX$ with penalty terms for enforcing additional properties of $K$ and $X$.

The minimization is based on alternating minimization with respect to $K$ or $X$, where in each iteration step one replaces the original Tikhonov functional by a locally defined surrogate functional. The choice of surrogate functionals is crucial: It should allow a comparatively simple minimization and simultaneously its first order optimality condition should lead to multiplicative update rules, which automatically preserve non-negativity of the iterates.

We review the most standard construction principles for surrogate functionals for Frobenius-norm and Kullback-Leibler discrepancy measures. We extend the known surrogate constructions by a general framework, which allows to add a large variety of penalty terms.

The paper finishes by deriving the corresponding alternating minimization schemes explicitly and by applying these methods to MALDI imaging data.

Keywords Non-negative matrix factorization · Multi-parameter regularization · Surrogate approaches · MM algorithms · Orthogonal NMF · Supervised NMF · MALDI mass spectrometry
1 Introduction

Matrix factorization methods for large scale data sets have seen increasing scientific interest recently due to their central role for a large variety of machine learning tasks. The main aim of such approaches is to obtain a low-rank approximation of a typically large data matrix by factorizing it into two smaller matrices. One of the most widely used matrix factorization method is the principal component analysis (PCA), which uses the singular value decomposition (SVD) of the given data matrix.

In this work, we review the particular case of non-negative matrix factorization (NMF), which is favorable for a range of applications where the data under investigation naturally satisfies a non-negativity constraint. These include dimension reduction, data compression, basis learning, feature extraction as well as higher level tasks such as classification or clustering [11,26,27,30]. PCA based approaches without any non-negativity constraints would not lead to satisfactory results in this case since possible negative entries of the computed matrices cannot be easily interpreted for naturally non-negative datasets.

Typically, the NMF problem is formulated as a minimization problem. The corresponding cost function includes a suitable discrepancy term, which measures the difference between the data matrix and the calculated factorization, as well as penalty terms to tackle the non-uniqueness of the NMF, to deal with numerical instabilities but also to provide the matrices with desirable properties depending on the application task. The NMF cost functions are commonly non-convex and require tailored minimization techniques to ensure the minimization but also the non-negativity of the matrix iterates. This leads us to the so called surrogate minimization approaches, which are also known as majorize-minimization algorithms [18,23,36]. Such surrogate methods have been investigated intensively for some of the most interesting discrepancy measures and penalty terms [11,13,16,23,25,33,36]. The idea is to replace the original cost function by a so called surrogate functional, such that its minimization induces a monotonic decrease of the objective function. It should be constructed in such a way that it is easier to minimize and that the deduced update rules should preserve the non-negativity of the iterates, which typically leads to alternating, multiplicative update rules.

It appears, that these constructions are obtained case-by-case employing different analytical approaches and different motivations for their derivation. The purpose of this paper, first of all, is to give a unified approach to surrogate constructions for NMF discrepancy functionals. This general construction principle is then applied to a wide class of functionals obtained by different combinations of divergence measures and penalty terms, thus extending the present state of the art for surrogate based NMF-constructions.

Secondly, one needs to develop minimization schemes for these functionals. Here we develop concepts for obtaining multiplicative minimization schemes, which automatically preserve non-negativity without the need for further projections.

Finally, we exemplify some characteristic properties of the different functionals with MALDI imaging data, which are particularly high-dimensional and challenging hyperspectral data sets.

The paper is organized as follows. Section 2 introduces the basic definition of the considered NMF problems. Section 3 gives an overview about the theory of surrogate functionals as well as the construction principles. This is then exemplified in
Section 4 for the most important cases of discrepancy terms, namely the Frobenius norm and the Kullback-Leibler divergence as well as for a variety of penalty terms. Section 5 discusses alternating minimization schemes for these general functionals with the aim to obtain non-negativity-preserving, multiplicative iterations. Finally, Section 6 contains numerical results for MALDI imaging data.

1.1 Notation

Throughout this work, we will denote matrices in bold capital Latin or Greek letters (e.g. $Y$, $K$, $\Psi$, $A$) while vectors will be written in small bold Latin or Greek letters (e.g. $c$, $d$, $\beta$, $\zeta$). The entries of matrices and vectors will be indicated in a non-bold format to distinguish between the $i$-th entry $x_i$ of a vector $x$ and $n$ different vectors $x_j$ for $j = 1, \ldots, n$. In doing so, we write for the entry of a matrix $M$ in the $i$-th row and the $j$-th column $M_{ij}$ and the $i$-th entry of a vector $x$ the symbol $x_i$. The same holds for an entry of a matrix product: the $ij$-th entry of the matrix product $MN$ will be indicated as $(MN)_{ij}$.

Furthermore, we will use a dot notation to indicate rows and columns of matrices. For a matrix $M$ we will write $M_{\cdot j}$ for the $j$-th column and $M_{i, \cdot}$ for the $i$-th row of the matrix.

What is more, we will use $\| \cdot \|$ for the usual Euclidean norm, $\| M \|_1 := \sum_{ij} |M_{ij}|$ for the 1-norm and $\| M \|_F$ for the Frobenius norm of a matrix $M$.

Besides that, we will use equivalently the terms function and functional for a mapping into the real numbers.

Finally, the dimensions of the matrices in the considered NMF problem are reused in this work and will be introduced in the following section.

2 Non-negative Matrix Factorization

Before we introduce the basic NMF problem, we give the following definition to clarify the meaning of a non-negative matrix.

**Definition 1** A matrix $M \in \mathbb{R}^{m \times n}$ is called non-negative if $M \in \mathbb{R}_{\geq 0}^{m \times n}$, where $\mathbb{R}_{\geq 0} := \{ x \in \mathbb{R} : x \geq 0 \}$.

The non-negativity of an arbitrary matrix $M$ will be abbreviated for simplicity as $M \geq 0$ in the later sections of this work.

The basic NMF problem requires to approximately decompose a given non-negative matrix $Y \in \mathbb{R}_{\geq 0}^{n \times m}$ into two smaller non-negative matrix factors $K \in \mathbb{R}_{\geq 0}^{n \times p}$ and $X \in \mathbb{R}_{\geq 0}^{p \times m}$, such that $p \ll \min(n, m)$ and

$$Y \approx KX.$$
the column vectors $Y_{*,j}$ as well as the row vectors $Y_{i,*}$:

\[
Y_{*,j} \approx \sum_{k=1}^{p} K_{*,k} X_{kj},
\]

\[
Y_{i,*} \approx \sum_{k=1}^{p} K_{ik} X_{k,*},
\]

\[
Y \approx KX = \sum_{k=1}^{p} K_{*,k} X_{k,*}.
\]

Note that the product $K_{*,k} X_{k,*}$ on the right hand side of the third equation yields rank-one matrices for every $k$.

By these representations, we can regard the rows $X_{k,*}$ as a low-dimensional set of basis vectors, which are tailored for approximating the high-dimensional data vectors, i.e. NMF solves the task of basis learning with non-negativity constraints. Following the interpretation given above, we can also regard NMF as a basis for compression. $K$ and $X$ are determined by storing $(n+m) \cdot p$ coefficients, as opposed to $n \cdot m$ coefficients for $Y$. The columns of $K$ can be regarded as characteristic components of the given data set $\{ Y_{*,j} \}_j$. If these data vectors are input for a classification task, one can use the $p$ correlation values with the column vectors of $K$ as features for constructing the classification scheme, which yields efficient and qualitatively excellent classifications, see [26,30,34].

The standard variational approach for constructing an NMF is to define a suitable discrepancy measure $D(\cdot, \cdot)$ between $Y$ and $KX$ and to minimize the resulting functional. Despite their seemingly simple structure NMF problems are ill-posed, non-linear and non-convex, i.e. they require stabilization techniques as well as tailored approaches for minimization. In this paper, we consider discrepancy measures based on divergences [17].

**Definition 2 (Divergence)** Let $\Omega$ be an arbitrary set. A divergence $D$ is a map $D: \Omega \times \Omega \to \mathbb{R}$, which fulfills the following properties:

(i) $D(x, y) \geq 0 \quad \forall (x, y) \in \Omega \times \Omega$

(ii) $D(x, y) = 0 \iff x = y$

**Definition 3 ($\beta$-divergence)** The $\beta$-divergence $d_\beta : \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ for $\beta \in \mathbb{R}$ is defined as

\[
d_\beta(x, y) := \begin{cases} 
\frac{x^\beta}{\beta(\beta - 1)} + \frac{y^\beta}{\beta} - \frac{xy^{\beta-1}}{\beta - 1} & \text{for } \beta \in \mathbb{R} \setminus \{0, 1\}, \\
x \log \frac{x}{y} - x + y & \text{for } \beta = 1, \\
\frac{x}{y} - \log \left( \frac{x}{y} \right) - 1 & \text{for } \beta = 0.
\end{cases}
\] (1)

Furthermore, we define accordingly $D_\beta : \mathbb{R}_{>0}^{n \times m} \times \mathbb{R}_{>0}^{n \times m} \to \mathbb{R}$ for arbitrary $m, n \in \mathbb{N}$ as

\[
D_\beta(M, N) = \sum_{i=1}^{n} \sum_{j=1}^{m} d_\beta(M_{ij}, N_{ij}).
\] (2)
The corresponding matrix divergences are defined componentwise, i.e. $\beta = 2$ yields the Frobenius norm and $\beta = 1$ the Kullback-Leibler divergence. These discrepancy measures are typically amended by so called penalty terms for stabilization and for enforcing additional properties such as sparsity or orthogonality. This yields the following general minimization task.

**Definition 4 (NMF Minimization Problem)** For a data matrix $Y \in \mathbb{R}^{n \times m}_{\geq 0}$, we consider the following generalized NMF minimization task

$$
\min_{K \geq 0, X \geq 0} D_\beta(Y, KX) + \sum_{\ell=1}^L \alpha_\ell \varphi_\ell(K, X).
$$

(3)

The functional

$$
F(K, X) := D_\beta(Y, KX) + \sum_{\ell=1}^L \alpha_\ell \varphi_\ell(K, X)
$$

(4)

is called the **cost functional**. Furthermore, we call

(i) $D_\beta(Y, KX)$ the **discrepancy term**,  
(ii) $\alpha_\ell$ the **regularization parameters** or weights  
(iii) and $\varphi_\ell(K, X)$ the **penalty terms**.

The functional in (4) is typically non-convex in $(K, X)$. Hence, algorithms based on alternating minimization with respect to $K$ or $X$ are favourable, i.e.

$$
K^{[d+1]} = \arg \min_{K \geq 0} F(K, X^{[d]}),
$$

(5)

$$
X^{[d+1]} = \arg \min_{X \geq 0} F(K^{[d+1]}, X),
$$

(6)

where the index $d$ denotes the iteration index of the corresponding matrices.

This yields simpler, often convex restricted problems with respect to either $K$ or $X$. Considering for example the minimization of the NMF functional with Frobenius norm and without any penalty term, yields a high dimensional linear system $K^\top Y = K^\top K X$, which, however, would need to be solved iteratively.

Instead, so called surrogate methods for computing NMF decompositions have been proposed recently and are introduced in the next section. They also consider alternating minimization steps for $K$ and $X$, but they replace the restricted minimization problems in (5) and (6) by simpler minimization tasks, which are obtained by locally replacing $F$ by surrogate functionals for $K$ and $X$ separately.

### 3 Surrogate Functionals

In this section, we discuss general surrogate approaches for minimizing general non-convex functionals, which are then exemplified for specific NMF functionals in later sections.

Let us consider a general functional $F : \Omega \to \mathbb{R}$ where $\Omega \subset \mathbb{R}^N$ and the minimization problem

$$
\min_{x \in \Omega} F(x).
$$
We will later add suitable conditions guaranteeing the existence of minimizers or at least the existence of stationary points. Surrogate concepts replace this task by solving a sequence of comparatively simpler and convex surrogate functionals, which can be minimized efficiently. These methods are also commonly referred to as \textit{surrogate minimization} (or \textit{maximization}) algorithms (SM) or also as MM algorithms, where the first M stands for majorize and the second M for minimize (see also \cite{15,22,36}). Such approaches have been demonstrated to be very useful in many fields of inverse problems, in particular for hyperspectral imaging \cite{11}, medical imaging applications such as transmission tomography \cite{13,14} as well as MALDI imaging and tumor typing applications \cite{26}.

Replacing a non-convex functional by a series of convex problems is the main motivation for such surrogate approaches. However, if constructed appropriately, they can also be used to replace non-differentiable functionals by a series of differentiable problems and they can be tailored such that gradient-descent methods for minimization yield multiplicative update rules which automatically incorporate non-negativity constraints without further projections.

From this point on it is important to note that possible zero denominators during the derivation of the NMF algorithms as well as in the multiplicative update rules themselves will not be discussed explicitly throughout this work. Usually, this issue is handled in practice by adding a small positive constant in the denominator during the iteration scheme. In fact, the instability of NMF algorithms due to the convergence of some entries in the matrices to zero is not sufficiently discussed in the literature and still needs proper solution techniques. We will not focus on this problem and turn now to the basic definition and properties of surrogate functionals.

3.1 Definitions and Basic Properties

As in \cite{25}, we use the following definition of a surrogate functional.

**Definition 5 (Surrogate Functional)** Let $\Omega \subseteq \mathbb{R}^N$ denote an open set and $F : \Omega \to \mathbb{R}$ a functional defined on $\Omega$. Then $Q_F : \Omega \times \Omega \to \mathbb{R}$ is called a \textit{surrogate functional} or a \textit{surrogate} for $F$, if it satisfies the following conditions:

(i) $Q_F(x, a) \geq F(x)$ for all $x, a \in \Omega$

(ii) $Q_F(x, x) = F(x)$ for all $x \in \Omega$

This is the most basic definition, which does not require any convexity or differentiability of the functional. However, it already allows to prove that the iteration

$$x^{[d+1]} := \arg \min_{x \in \Omega} Q_F(x, x^{[d]})$$

yields a sequence which monotonically decreases $F$.

**Lemma 1 (Monotonic Decrease by Surrogate Functionals)** Let $\Omega \subseteq \mathbb{R}^N$ denote an open set, $F : \Omega \to \mathbb{R}$ a given function and $Q_F$ a surrogate functional for $F$. Assume that $\arg \min_{x \in \Omega} Q_F(x, a)$ is well defined for all $a \in \Omega$. Define the iterated updates by

$$x^{[d+1]} := \arg \min_{x \in \Omega} Q_F(x, x^{[d]})$$

...
with $x^{[0]} = \arg \min \alpha \in \Omega Q_F(x, \alpha)$ for an arbitrary $\alpha \in \Omega$. Then, $F(x^{[d]})$ is a monotonically decreasing sequence, i.e.

$$F(x^{[d+1]}) \leq F(x^{[d]}). \quad (9)$$

**Proof** The monotone decrease (9) follows directly from the defining properties of surrogate functionals, see Definition 5. We obtain

$$F(x^{[d+1]}) \leq Q_F(x^{[d+1]}, x^{[d]}) \overset{(*)}{=} Q_F(x^{[d]}, x^{[d]}) = F(x^{[d]}),$$

where $(*)$ follows from the definition of $x^{[d+1]}$ in (8).

**Remark 1 (Addition of Surrogate Functionals)** Let $\Omega \subseteq \mathbb{R}^n$ be an open set, $F, G : \Omega \rightarrow \mathbb{R}$ pointwise defined functionals and $Q_F, Q_G$ corresponding surrogates. Then $Q_F + Q_G$ is a surrogate functional for $F + G$.

For each functional $F$ there typically exist a large variety of surrogate functionals and we can aim at optimizing the structure of surrogate functionals. The following additional property is the key to simple and efficient minimization schemes for surrogate functionals.

**Definition 6 (Separability of a Surrogate Functional)** Let $\Omega \subseteq \mathbb{R}^N$ denote an open set, $F : \Omega \rightarrow \mathbb{R}$ a functional and $Q_F$ a surrogate for $F$. The surrogate $Q_F$ is called **separable**, if there exist functions $g_i : \mathbb{R} \times \Omega \rightarrow \mathbb{R}$, such that

$$Q_F(x, \alpha) = \sum_{i=1}^{N} g_i(x_i, \alpha) \quad \forall x, \alpha \in \Omega. \quad (10)$$
Lemma 1 above only ensures the monotonic decrease of the cost functional, which is not sufficient to guarantee convergence of the sequence \( \{x^{[d]}\} \) to a minimizer of \( F \) or at least to a stationary point of \( F \). The convergence theory for surrogate functionals is far from being complete (see also the works [23] and [36]). Despite this lack of theoretical foundation, surrogate based minimization yields strictly decreasing sequences for a large variety of applications. In particular, surrogate based methods can be constructed such that first order optimality conditions lead to multiplicative update rules, which - in view of applications to NMF constructions - is a very desirable property.

We now turn to discussing three different construction principles for surrogate functionals.

### 3.2 Jensen’s Inequality

The starting point is the well known Jensen’s inequality for convex functions, see [10].

**Lemma 2 (Jensen’s Inequality)** Let \( \Omega \subseteq \mathbb{R}^N \) denote a convex set, \( F : \Omega \to \mathbb{R} \) a convex function and \( \lambda_i \in [0,1] \) non-negative numbers for \( i \in \{1, \ldots, k\} \) with \( \sum_{i=1}^{k} \lambda_i = 1 \). Then for all \( x_i \in \Omega \), it holds that

\[
F\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leq \sum_{i=1}^{k} \lambda_i F(x_i).
\]  

(11)

In this subsection we consider functionals \( F \) which are derived from continuously differentiable and convex functions \( f : \mathbb{R}_{>0} \to \mathbb{R} \) via

\[
F : \Omega \to \mathbb{R} \\
\quad \quad \quad v \mapsto f(c^\top v)
\]  

(12)

for \( \Omega \subseteq \mathbb{R}^N_{\geq 0} \) and some auxiliary variable \( c \in \Omega \). This also implies, that \( F \) is convex, since

\[
F(v) \geq F(\tilde{v}) + \nabla F(\tilde{v})^\top (v - \tilde{v})
\]

\[
\iff f(c^\top v) \geq f(c^\top \tilde{v}) + f'(c^\top \tilde{v})(c^\top v - c^\top \tilde{v}).
\]

We now choose \( \lambda_i \in [0,1] \) with \( \sum_{i=1}^{N} \lambda_i = 1 \) and \( \alpha \in \mathbb{R}^N \) and define

\[
\lambda_i := \frac{c_i b_i}{c^\top b} \]

(13)

\[
\alpha_i := \frac{c_i v_i}{\lambda_i} = \frac{v_i c^\top b}{b_i}
\]

(14)

for some \( b \in \Omega \). This implies

\[
F(v) = f(c^\top v) = f\left(\sum_{i=1}^{N} \lambda_i \alpha_i \right) \leq \sum_{i=1}^{N} \frac{c_i b_i}{c^\top b} f\left(\frac{c^\top b}{b_i} v_i\right) =: Q_F(v, b).
\]  

(15)
The functional $Q_F : \Omega \times \Omega \rightarrow \mathbb{R}$ defines a surrogate for $F$, which can be seen by the inequality above and by observing

$$Q_F(v, v) = \sum_{i=1}^N c_i^T f(c_i^T v) = f(c_i^T v) = F(v).$$

### 3.3 Low Quadratic Bound Principle

This concept is based on a Taylor expansion of $F$ in combination with a majorization of the quadratic term. This so called low quadratic bound principle (LQBP) has been introduced in [5] and was used in particularly for the computation of maximum-likelihood estimators. These methods do not require that $F$ itself is convex and its construction is based on the following lemma.

**Lemma 3 (Low Quadratic Bound Principle)** Let $\Omega \subseteq \mathbb{R}^N$ denote an open and convex set and $f : \Omega \rightarrow \mathbb{R}$ a twice differentiable functional. Assume that a matrix $A(x) \in \mathbb{R}^{N \times N}$ exists, such that $A(x) - \nabla^2 f(x)$ is positive semi-definite for all $x \in \Omega$. We then obtain a quadratic majorization

$$f(x) \leq f(a) + \nabla f(a)^T(x - a) + \frac{1}{2}(x - a)^T A(x - a) \quad \forall x, a \in \Omega \quad (16)$$

$$=: Q_f(x, a),$$

and $Q_f$ is a surrogate functional for $f$.

**Proof** The proof of this classical result is based on the second-order Taylor polynomial of $f$ and shall be left to the reader.

The related update rule for surrogate minimization can be stated explicitly under natural assumptions on the matrix $A$.

**Corollary 1** Assume that the assumptions of Lemma 3 hold. In addition, assume that $A$ is a positive definite and symmetric matrix. Then, the corresponding surrogate $Q_f$ is strictly convex in its first variable and we have from (16)

$$x^{[d+1]} = \arg \min_{x \in \Omega} Q_f(x, x^{[d]}) = x^{[d]} - A^{-1}(x^{[d]}) \nabla f(x^{[d]}). \quad (17)$$

**Proof** For an arbitrary $\alpha \in \{1, \ldots, N\}$, we have that

$$\frac{\partial Q_f}{\partial x_\alpha}(x, a) = \frac{\partial}{\partial x_\alpha} \left( \sum_{i=1}^N \frac{\partial f}{\partial x_i}(a)(x_i - a_i) + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N A(a)_{ij} (x_i - a_i)(x_j - a_j) \right)$$

$$= \frac{\partial f}{\partial x_\alpha}(a) + \frac{1}{2} \sum_{i=1}^N \sum_{j \neq \alpha} A(a)_{ij} (x_i - a_i)(x_j - a_j) + \frac{1}{2} \sum_{i=1}^N A(a)_{i\alpha}(x_i - a_i) + A(a)_{\alpha\alpha}(x_\alpha - a_\alpha)$$

$$\overset{(*)}{=} \frac{\partial f}{\partial x_\alpha}(a) + A(a)_{\alpha\alpha} x_\alpha - a_\alpha,$$
where \((\ast)\) utilizes the symmetry of \(A\). Hence, it holds that \(\nabla_\ast Q_f(x, a) = \nabla f(a) + A(a)(x - a)\). The Hessian matrix of \(Q_f\) then satisfies
\[
\nabla_\ast^2 Q_f(x, a) = A(a).
\]

This implies the positive definiteness of the functional, hence, it has a unique minimizer, which is given by
\[
\nabla_\ast x^*_a Q_f(x, a) = \nabla f(a) + A(a)(x^*_a - a)
\]
\[\iff x^*_a = a - A^{-1}(a)\nabla f(a).
\]
This is the update rule above.

The computation of the inverse of \(A\) is particularly simple if \(A\) is a diagonal matrix. Furthermore, the diagonal structure ensures the separability of the surrogate functional mentioned in Definition 6. Therefore, we consider matrices of the form
\[
A(a)_{ii} := \left(\frac{\nabla^2 f(a)}{a_i} + \kappa_i\right) a_i, \tag{18}
\]
where \(\kappa_i \geq 0\) has to be chosen individually depending on the considered cost function. We will see that an appropriate choice of \(\kappa_i\) will lead finally to the desired multiplicative update rules of the NMF algorithm.

The matrix \(A(a)\) in \([18]\) fulfills the conditions in Corollary 1 as it can be seen by the following lemma. Therefore, if \(A\) is constructed as in \([18]\), the update rule in \([17]\) can be applied immediately.

**Lemma 4** Let \(M \in \mathbb{R}^{N \times N}_{\geq 0}\) denote a symmetric matrix. With \(a \in \mathbb{R}^N > 0\) and \(\kappa_i \geq 0\), we define the diagonal matrix \(A\), such that
\[
A_{ii} := \frac{(Ma)_i + \kappa_i}{a_i}, \tag{19}
\]
for \(i = 1, \ldots, N\). Then \(A\) and \(A - M\) are positive semi-definite.

**Proof** Let \(\zeta \in \mathbb{R}^N\) denote an arbitrary vector and let \(\delta\) denote the Kronecker symbol. Then
\[
\zeta^\top (A-M)\zeta = \sum_{i,j=1}^N \zeta_i \delta_{ij} \left(\frac{(Ma)_i + \kappa_i}{a_i} - \sum_{i,j=1}^N \zeta_i M_{ij} \zeta_j\right)
\]
\[
= \sum_{i=1}^N \zeta_i^2 \frac{(Ma)_i}{a_i} + \zeta_i^2 \frac{\kappa_i}{a_i} - \sum_{i,j=1}^N \zeta_i M_{ij} \zeta_j
\]
\[
\geq \sum_{i,j=1}^N \zeta_i^2 \frac{a_j}{a_i} M_{ij} - \sum_{i,j=1}^N \zeta_i M_{ij} \zeta_j
\]
\[
= \sum_{i=1}^N \zeta_i^2 M_{ij} + \sum_{i,j=1}^N \left(\zeta_i^2 \frac{a_j}{a_i} + \zeta_i^2 \frac{a_i}{a_j}\right) M_{ij} - \sum_{i,j=1}^N \zeta_i M_{ij} \zeta_j
\]
= \sum_{i,j=1}^{N} \left[ \frac{1}{2} \zeta^2_i \frac{a_i}{a_j} M_{ij} + \frac{1}{2} \zeta^2_j \frac{a_j}{a_i} M_{ij} - \zeta_i M_{ij} \zeta_j \right] \\
= \frac{1}{2} \sum_{i,j=1}^{N} \zeta^2_i \frac{a_i}{a_j} M_{ij} + \zeta^2_j \frac{a_j}{a_i} M_{ij} - 2 \sqrt{\frac{a_j}{a_i}} \sqrt{\frac{a_i}{a_j}} M_{ij} \zeta_j \\
= \frac{1}{2} \sum_{i,j=1}^{N} \left( \sqrt{\frac{a_j}{a_i}} \zeta_i - \sqrt{\frac{a_i}{a_j}} \zeta_j \right)^2 M_{ij} \geq 0.

The positive semi-definiteness of $\Lambda$ follows from its diagonal structure. \hfill \Box

### 3.4 Further Construction Principles

So far we have discussed two major construction principles based on either Jensen’s inequality or on upper bounds for the quadratic term in Taylor expansions. [23] lists further construction principles, which however will not be used for NMF constructions in the subsequent sections of this paper. For completeness we shortly list their main properties.

A relaxation of the approach based on Jensen’s inequality is achieved by choosing $\alpha_i \geq 0, i \in \{1, \ldots, N\}$ such that $\sum_{i=1}^{N} \alpha_i = 1$ and $\alpha_i > 0$ if $c_i \neq 0$, which yields

$$
F(v) = f(c^T v) \leq \sum_{i=1}^{N} \alpha_i f \left( \frac{c_i}{\alpha_i} (v_i - b_i) + c^T b \right) =: Q_F(v, b).
$$

A typical choice is

$$
\alpha_i := \frac{|c_i|^p}{\sum_{j=1}^{N} |c_j|^p}
$$

which leads to surrogate functionals for $p \geq 0$. This type of surrogate was originally introduced in the context of medical imaging, see [12], for positron emission tomography.

Another approach is based on combining arithmetic with geometric means and can be used for constructing surrogates for posynomial functions. For $\alpha, v, a \in \mathbb{R}^N_{>0}$, we obtain

$$
F(v) = \prod_{i=1}^{N} v_i^{\alpha_i} \leq \left( \prod_{i=1}^{N} a_i^{\alpha_i} \right) \sum_{i=1}^{N} \frac{\alpha_i}{\sum_{k=1}^{N} \alpha_k} \left( \frac{v_i}{a_i} \right)^{\sum_{k=1}^{N} \alpha_i} \right) =: Q_F(v, a)
$$

### 4 Surrogates for NMF Functionals

In this section we apply the general construction principles of Section 3 to the NMF problem as stated in (3). The resulting functional $F(K, X)$ depends on both factors of the matrix decomposition and minimization is attempted by alternating minimization with respect to $K$ and $X$ as described in (5) and (6). However, we replace the functional $F$ in each iteration by suitable surrogate functionals, which allow an explicit minimization. Hence, we avoid the minimization of $F$ itself, which even for the most simple quadratic formulation requires to solve...
a high dimensional linear system.

We start by considering the discrepancy terms for $\beta = 2$ (Frobenius norm) and $\beta = 1$ (Kullback-Leibler divergence) and determine surrogate functionals with respect to $X$ and $K$. We then add several penalty terms and develop surrogate functionals accordingly. With regard to the construction of surrogates for the case of $\beta = 0$, which leads to the so-called Itakura-Saito divergence, we refer to the works [15,16,32].

4.1 Frobenius Discrepancy and Low Quadratic Bound Principle

We start by constructing a surrogate for the minimization with respect to $X$ for the Frobenius discrepancy

$$F(X) := \frac{1}{2} \| Y - KX \|^2_F.$$  \hspace{1cm} (20)

Let $Y_{\bullet,j}$, resp. $X_{\bullet,j}$, denote the column vectors of $Y$, resp. $X$. The separability of $F$ yields

$$F(X) = \frac{1}{2} \sum_{j=1}^m \| Y_{\bullet,j} - KX_{\bullet,j} \|^2 =: \sum_{j=1}^m f_{Y_{\bullet,j}}(X_{\bullet,j}).$$  \hspace{1cm} (21)

Hence, the minimization separates for the different $f_{Y_{\bullet,j}}$ terms. The Hessian of these terms is given by

$$\nabla^2 f_{Y_{\bullet,j}}(a) = K^\top K$$

and the LQBP construction principle of the previous section with $\kappa_k = 0$ yields

$$A_{f_{Y_{\bullet,j}}}(a)_{kk} = \frac{(K^\top Ka)_k}{a_k},$$

leading to the surrogate functionals

$$Q_{f_{Y_{\bullet,j}}}(x,a) = f_{Y_{\bullet,j}}(a) + \nabla f_{Y_{\bullet,j}}(a)^\top(x-a) + \frac{1}{2}(x-a)^\top A_{f_{Y_{\bullet,j}}}(a)(x-a).$$  \hspace{1cm} (22)

An appropriate choice of $\kappa_k$ ensures the multiplicativity of the final NMF algorithm. In the case of the Frobenius discrepancy term, we will see that suitable $\kappa_k$ can be chosen dependent on $\ell_1$ regularization terms in the cost function, which are not included up to now (see Subsection 4.4 and Appendix A.1 for more details on this issue). Due to the absent $\ell_1$ terms, we set $\kappa_k = 0$ to get the desired multiplicative update rules. Summing up the contributions of the columns of $X$ yields the final surrogate

$$Q_F : \mathbb{R}^{p \times m} \times \mathbb{R}^{p \times m} \to \mathbb{R}, \quad (X,A) \mapsto \sum_{j=1}^m Q_{f_{Y_{\bullet,j}}}(X_{\bullet,j}, A_{\bullet,j})$$

The equivalent construction for $K$ can be obtained by regarding the rows of $K$ separately, which for

$$g_y : \mathbb{R}^p_{\geq 0} \to \mathbb{R}, \quad k \mapsto \frac{1}{2} \| y - kX \|^2$$
yields $\nabla^2 g_y(a) = XX^\top$. Putting
\[
A_{g_y}(a)_{kk} = \frac{(aXX^\top)_{kk}}{a_k}
\]
leads to the surrogate
\[
Q_{g_y}(k, a) = g_y(a) + (k - a)\nabla g_y(a) + \frac{1}{2}(k - a)A_{g_y}(a)(k - a)^\top.
\]
We summarize this surrogate construction in the following theorem.

**Theorem 1 (Surrogate Functional for the Frobenius Norm with LQBP)**
We consider the cost functionals
\[
F(X) := \frac{1}{2}\|Y - KX\|_F^2 \quad \text{and} \quad G(K) := \frac{1}{2}\|Y - KX\|_F^2.
\]
Then
\[
Q_{F,1}(X, A) = \sum_{j=1}^m Q_{f_{v_{x,j}}}(X_{\bullet,j}, A_{\bullet,j}), \quad (23)
\]
\[
Q_{G,1}(K, A) = \sum_{i=1}^n Q_{g_{Y_i,\bullet}}(K_{i,\bullet}, A_{i,\bullet}) \quad (24)
\]
define separable and convex surrogate functionals.

### 4.2 Frobenius Discrepancy and Jensen’s Inequality

Again we focus on deriving a surrogate functional for $X$, the construction for $K$ will be very similar. Expanding the Frobenius discrepancy yields
\[
F(X) := \frac{1}{2}\|Y - KX\|_F^2 = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - (KX)_{ij})^2.
\]
Putting $v := X_{\bullet,j} \in \mathbb{R}_{\geq 0}^p$ and $c := K_{i,\bullet}^\top \in \mathbb{R}_{\geq 0}^p$ allows us to define
\[
f : \mathbb{R}_{\geq 0} \to \mathbb{R} \quad \text{with} \quad f(t) := (Y_{ij} - t)^2, \quad (25)
\]
such that
\[
f(c^\top v) = (Y_{ij} - (KX)_{ij})^2. \quad (26)
\]
Hence we have separated the Forbenius discrepancy suitably for applying Jensen’s inequality. Following the construction principle in Subsection 3.2, we define
\[
\lambda_k = \frac{K_{ik}A_{kj}}{(KA)_{ij}}, \quad (27)
\]
\[
\alpha_k = \frac{K_{ik}X_{kj}}{\lambda_k}, \quad (28)
\]
with the auxiliary variable $A \in \mathbb{R}_{\geq 0}^{p \times m}$ and $b := A_{\bullet,j} \in \mathbb{R}_{\geq 0}^p$, which yields the inequality
\[
(Y_{ij} - (KX)_{ij})^2 \leq \sum_{k=1}^p \frac{K_{ik}A_{kj}}{(KA)_{ij}} \left( Y_{ij} - \frac{X_{kj}}{A_{kj}}(KA)_{ij} \right)^2.
\]
Inserting this into the decomposition of the Frobenius discrepancy yields the surrogate $Q_{F,2}(X, A)$ by

$$F(X) \leq \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{(KA)_{ij}} \sum_{k=1}^{p} K_{ik} A_{kj} \left( Y_{ij} - \frac{X_{kj}}{A_{kj}} (KA)_{ij} \right)^2 := Q_{F,2}(X, A),$$

The construction of a surrogate for $K$ proceeds in the same way. We summarize the results in the following theorem.

**Theorem 2 (Surrogate Functional for the Frobenius Norm with Jensen’s Inequality)** We consider the cost functionals $F(X) := \frac{1}{2} \| Y - KX \|_F^2$ and $G(K) := \frac{1}{2} \| Y - KX \|_F^2$. Then

$$Q_{F,2}(X, A) := \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{(KA)_{ij}} \sum_{k=1}^{p} K_{ik} A_{kj} \left( Y_{ij} - \frac{X_{kj}}{A_{kj}} (KA)_{ij} \right)^2$$

$$Q_{G,2}(K, A) := \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{(AX)_{ij}} \sum_{k=1}^{p} A_{ik} X_{kj} \left( Y_{ij} - \frac{K_{ik}}{A_{ik}} (AX)_{ij} \right)^2$$

define separable and convex surrogate functionals.

These surrogates are equal to the ones proposed in [11]. We will later use first order necessary conditions of the surrogate functionals for obtaining algorithms for minimization. We already note

$$\frac{\partial Q_{F,1}}{\partial X_{\alpha\beta}} = \frac{\partial Q_{F,2}}{\partial X_{\alpha\beta}} \quad \text{and} \quad \frac{\partial Q_{G,1}}{\partial K_{\alpha\beta}} = \frac{\partial Q_{G,2}}{\partial K_{\alpha\beta}},$$

i.e. despite the rather different derivations, the update rules for the surrogates obtained by LQBP and Jensen’s inequality will be identical.

### 4.3 Surrogates for Kullback-Leibler Divergence

The case $\beta = 1$ in Definition 3 yields the so-called Kullback-Leibler divergence (KLD). For matrices $M, N \in \mathbb{R}_{>0}^{n \times m}$, it is defined as

$$\text{KL}(M, N) := D_1(M, N) = \sum_{i=1}^{n} \sum_{j=1}^{m} M_{ij} \log \left( \frac{M_{ij}}{N_{ij}} \right) - M_{ij} + N_{ij}$$

and has been investigated intensively in connection with non-negative matrix factorization methods [11,16,24,25]. In our context, we define the cost functional for the NMF construction by

$$F(X, K) := \text{KL}(Y, KX).$$

We will focus in this subsection on Jensen’s inequality for constructing surrogates for the KLD since they will lead to the known classical NMF algorithms (see also [11,24,25]). However, it is also possible to use the LQBP principle to construct a suitable surrogate functional for the KLD which leads to different, multiplicative update rules (see Appendix B).
We start by deriving a surrogate for the minimization with respect to \( \mathbf{X} \), i.e. we consider

\[
F(\mathbf{X}) := \text{KL}(\mathbf{Y}, \mathbf{KX}) = \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} \log(Y_{ij}) - Y_{ij} \log((\mathbf{KX})_{ij}) - Y_{ij} + (\mathbf{KX})_{ij}.
\]

Using the same \( \lambda_k \) and \( \alpha_k \) as in the section above and applying it to the convex function \( f(t) := -\ln(t) \), we obtain

\[
-\ln((\mathbf{KX})_{ij}) \leq -\sum_{k=1}^{p} \frac{K_{ik}A_{kj}}{(KA)_{ij}} \ln \left( \frac{X_{kj}}{A_{kj}(KA)_{ij}} \right).
\]

Multiplication with \( Y_{ij} \geq 0 \) and the addition of appropriate terms yields

\[
F(\mathbf{X}) \leq \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} \ln(Y_{ij}) - Y_{ij} + (\mathbf{KX})_{ij} - \frac{Y_{ij}}{(KA)_{ij}} \sum_{k=1}^{p} K_{ik}A_{kj} \ln \left( \frac{X_{kj}}{A_{kj}(KA)_{ij}} \right)
\]

\[
=: Q_{F,1}(\mathbf{X}, \mathbf{A}).
\]

The condition \( Q_{F,1}(\mathbf{X}, \mathbf{X}) = F(\mathbf{X}) \) follows by simple algebraic manipulations, such that \( Q_{F,1} \) is a valid surrogate functional for \( F \).

The approach by Jensen’s inequality is very flexible and we obtain different surrogates \( Q_{F,2} \) and \( Q_{F,3} \) by using i.e. \( f_1(t) = Y_{ij} \ln(Y_{ij}/t) - Y_{ij} + t \) or \( f_2(t) = -Y_{ij} \ln(t) + t \) instead of \( f \). Inserting the same \( \lambda_k \) and \( \alpha_k \) as before in Equation (15), we obtain immediately the surrogates

\[
Q_{F,2}(\mathbf{X}, \mathbf{A}) = \left[ \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{(KA)_{ij}} \sum_{k=1}^{p} K_{ik}A_{kj} \left( Y_{ij} \ln \left( \frac{Y_{ij}}{X_{kj}(KA)_{ij}} \right) - Y_{ij} + \frac{X_{kj}}{A_{kj}(KA)_{ij}} \right),
\]

\[
Q_{F,3}(\mathbf{X}, \mathbf{A}) = \left[ \sum_{i=1}^{n} \sum_{j=1}^{m} \left( Y_{ij} \ln(Y_{ij}) - Y_{ij}
\right.
\]

\[
+ \frac{1}{(KA)_{ij}} \sum_{k=1}^{p} K_{ik}A_{kj} \left( -Y_{ij} \ln \left( \frac{X_{kj}}{A_{kj}(KA)_{ij}} \right) + \frac{X_{kj}}{A_{kj}(KA)_{ij}} \right).
\]

It easy to check, that the partial derivatives for all three variants are the same, hence, the update rules obtained in the next section based on first order optimality conditions will be identical. Applying the same approach for obtaining a surrogate for \( \mathbf{K} \) yields the following theorem.

**Theorem 3 (Surrogate Functional for the Kullback-Leibler Divergence with Jensen’s Inequality)** We consider the cost functionals \( F(\mathbf{X}) := \text{KL}(\mathbf{Y}, \mathbf{KX}) \) and \( G(\mathbf{K}) := \text{KL}(\mathbf{Y}, \mathbf{KX}) \). Then

\[
Q_F(\mathbf{X}, \mathbf{A}) := \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} \ln(Y_{ij}) - Y_{ij} + (\mathbf{KX})_{ij} - \frac{Y_{ij}}{(AX)_{ij}} \sum_{k=1}^{p} K_{ik}A_{kj} \ln \left( \frac{X_{kj}}{A_{kj}(AX)_{ij}} \right)
\]

\[
Q_G(\mathbf{K}, \mathbf{A}) := \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} \ln(Y_{ij}) - Y_{ij} + (\mathbf{KX})_{ij} - \frac{Y_{ij}}{(AX)_{ij}} \sum_{k=1}^{p} K_{ik}A_{kj} \ln \left( \frac{X_{kj}}{A_{kj}(AX)_{ij}} \right)
\]

define separable and convex surrogate functionals.
4.4 Surrogates for $\ell_1$- and $\ell_2$-Norm Penalties

Computing an NMF is an ill-posed problem, see [11], hence one needs to add stabilizing penalty terms for obtaining reliable matrix decompositions. The most standard penalties are $\ell_1$- and $\ell_2$-terms for the matrix factors leading to

$$\min_{K, X \geq 0} D_\beta(Y, KX) + \lambda \|X\|_1 + \frac{\mu}{2}\|K\|^2_F + \frac{\nu}{2}\|X\|^2_F + \omega\|K\|_1$$

(32)

for $\beta \in \{1, 2\}$.

The $\ell_2$-penalty prohibits exploding norms for each matrix factor and the $\ell_1$-term promotes sparsity in the minimizing factors, see [21,28] for a general exposition.

Combinations of $\ell_1$- and $\ell_2$-norms are sometimes called elastic net regularization, [20], due to there importance in medical imaging.

These penalty terms are convex and they separate, hence, they can be used as surrogates themselves. For the case of Kullback-Leibler divergences this leads to the following surrogate for minimization with respect to $X$:

$$Q_F(X, A) := Q_{KL}(X, A) + \lambda \|X\|_1 + \frac{\nu}{2}\|X\|^2_F,$$

where $Q_{KL}$ is the surrogate for the Kullback-Leibler divergence of Theorem 3 for $X$.

The Frobenius case cannot be treated in the same way. If we use the penalty terms as surrogates themselves and obtain the standard minimization algorithm by first order optimality conditions, then this does not lead to a multiplicative algorithm, which preserves the non-negativity of the iterates. It can be easily seen that the $\ell_1$-penalty term causes this difficulty. For a more extended discussion on this, see Appendix A.1.

Hence, we have to construct a different surrogate. Similar to the discussion in Subsection 4.1, we consider here $f_y : \mathbb{R}_{\geq 0}^p \rightarrow \mathbb{R}$ with

$$f_y(x) := \frac{1}{2}\|y - Kx\|^2 + \lambda \|x\|_1 + \frac{\nu}{2}\|x\|^2,$$

which yields the Hessian $\nabla^2 f_y(a) = K^T K + \nu I$. The choice of $\kappa_k$ is done dependent on the $\ell_1$ regularization term of the cost function $f_y$ as already described in Subsection 4.1. It can be shown in the derivation of the NMF algorithm that $\kappa_k = \lambda$ for all $k$ leads to multiplicative update rules. A more general cost function is considered in Appendix A.1 where the concrete effect of $\kappa_k$ is described in more detail.

This yields the following diagonal matrix $A_{f_y}(a)$:

$$A_{f_y}(a)_{kk} = ((K^T K + \nu I)_{xx}a)_k + \lambda a_k.$$
4.5 Surrogates for Orthogonality Constraints

The observation that a non-negative matrix with pairwise orthogonal rows has at most one non-zero entry per column is the motivation for introducing orthogonality constraints for $K$ or $X$. This will lead to strictly uncorrelated feature vectors, which is desirable in several applications e.g. for obtaining discriminating biomarkers from mass spectra, see Section 4.4 on MALDI Imaging.

We could add the orthogonality constraint $K^T K = I$ as an additional penalty term $\sigma_K \|K^T K - I\|^2$. However, this would introduce fourth order terms. Hence we introduce additional variables $V$ and $W$ and split the orthogonality condition into two second order terms leading to

$$
\min_{K, X, V, W \geq 0} \left\{ D_p(Y, KX) + \frac{\sigma_{K,1}}{2} \|I - V^TK\|^2_F + \frac{\sigma_{K,2}}{2} \|V-K\|^2_F \\
+ \frac{\sigma_{X,1}}{2} \|I - XW^T\|^2_F + \frac{\sigma_{X,2}}{2} \|W - X\|^2_F \right\}.
$$

(33)

Surrogates for the terms $\|I - V^T K\|^2_F$ and $\|I - XW^T\|^2_F$ can be calculated via Jensen’s inequality (see Subsection 4.2). The other penalties can be used as surrogates themselves and therefore, we obtain

**Theorem 4 (Surrogate Functionals for Orthogonality Constraints)** We consider the cost functionals

$$
F(X) := \frac{\sigma_{X,1}}{2} \|I - XW^T\|^2_F + \frac{\sigma_{X,2}}{2} \|W - X\|^2_F := G(W),
$$

$$
H(K) := \frac{\sigma_{K,1}}{2} \|I - V^T K\|^2_F + \frac{\sigma_{K,2}}{2} \|V - K\|^2_F := J(V)
$$

with $\sigma_{X,1}, \sigma_{X,2}, \sigma_{K,1}, \sigma_{K,2} \geq 0$. Then

$$
Q_F(X, A) := \frac{\sigma_{X,1}}{2} \sum_{k=1}^p \sum_{l=1}^n \sum_{j=1}^m \frac{1}{(AW')_{kj}} A_{kj} W_{lj} \left( \delta_{kl} - \frac{X_{kj}}{A_{kj}} (AW')_{kl} \right)^2 + \frac{\sigma_{X,2}}{2} \|W - X\|^2_F
$$

$$
Q_G(W, A) := \frac{\sigma_{X,1}}{2} \sum_{k=1}^p \sum_{l=1}^n \sum_{j=1}^m \frac{1}{(X'A)_{kj}} X_{kj} A_{lj} \left( \delta_{kl} - \frac{W_{lj}}{A_{lj}} (X'A)_{kl} \right)^2 + \frac{\sigma_{X,2}}{2} \|W - X\|^2_F
$$

$$
Q_H(K, A) := \frac{\sigma_{K,1}}{2} \sum_{k=1}^p \sum_{l=1}^n \sum_{i=1}^m \frac{1}{(V'A_{kl})} V_{ik} A_{kl} \left( \delta_{kl} - \frac{K_{kl}}{A_{kl}} (V'A)_{kl} \right)^2 + \frac{\sigma_{K,2}}{2} \|V - K\|^2_F
$$

$$
Q_J(V, A) := \frac{\sigma_{K,1}}{2} \sum_{k=1}^p \sum_{l=1}^n \sum_{i=1}^m \frac{1}{(A'K)_{kl}} A_{kl} K_{ik} \left( \delta_{kl} - \frac{V_{ik}}{A_{ik}} (A'K)_{kl} \right)^2 + \frac{\sigma_{K,2}}{2} \|V - K\|^2_F
$$

define separable and convex surrogate functionals.

4.6 Surrogates for Total Variation Penalties

Total variation (TV) penalty terms are the second important class of regularization terms besides $\ell_p$-penalty terms. TV-penalties aim at smooth or even piecewise constant minimizers, hence they are defined in terms of first order or higher order derivatives [7].

Originally, they were introduced for denoising applications in image processing.
but have since been applied to inpainting, deconvolution and other inverse problems, see e.g. [8]. The precise mathematical formulation of the total variation in the continuous case is described in the following definition.

**Definition 7 (Total Variation (Continuous))** Let $\Omega \subset \mathbb{R}^N$ be open and bounded. The **total variation** of a function $u \in L^1_{\text{loc}}(\Omega)$ is defined as

$$TV(u) := \sup \left\{ -\int_{\Omega} u \text{div} \phi \, dx : \phi \in C^\infty_c(\Omega, \mathbb{R}^N) \text{ mit } \|\phi(x)\| \leq 1 \ \forall x \in \Omega \right\}.$$ 

There exist several analytic relaxations of TV based on $\ell_1$-norms of the gradient, which are more tractable for analytical investigations. For numerical implementations one rather uses the $L_1$-norm of the gradient $\|\nabla f\|_{L_1}$ as a more computationally tractable substitute. For discretization the gradient is typically replaced by a finite difference approximation [9].

For applying TV-norms to data, we assume that the row index in the data matrix $Y$ refers to spatial locations and the column index to so-called channels. In this case, we consider the most frequently used isotropic TV for applying it to measured, discretized hyperspectral data.

**Definition 8 (Total Variation (Discrete))** For fixed $\varepsilon_{TV} > 0$, the **total variation** of a matrix $K \in \mathbb{R}^{n \times p}$ is defined as

$$TV(K) := \sum_{k=1}^{p} \psi_k \sum_{i=1}^{n} \sqrt{\varepsilon_{TV}^2 + \sum_{\ell \in N_i} (K_{ik} - K_{\ell k})^2},$$

where $\psi_k \in \mathbb{R}_{\geq 0}$ is a weighting of the $k$-th data channel and $N_i \subseteq \{1, \ldots, n\} \setminus \{i\}$ denotes the index set referring to spatially neighboring pixels.

We will use the following short hand notation

$$|\nabla_{ik} K| := \sqrt{\varepsilon_{TV}^2 + \sum_{\ell \in N_i} (K_{ik} - K_{\ell k})^2},$$

which can be seen as a finite difference approximation of the gradient magnitude of the image $K_{*,k}$ at pixel $K_{ik}$ for some neighbourhood pixels defined by $N_i$. A typical choice for neighbourhood pixels in two dimensions for the pixel $(0, 0)$ is $N_{(0,0)} := \{(1, 0), (0, 1)\}$ to get an estimate of the gradient components along both axes. Finally, by introducing the positive constant $\varepsilon_{TV} > 0$, we get a differentiable approximation of the total variation penalty.

In Section 6 we will discuss the application of NMF methods to hyperspectral MALDI imaging datasets, which has a natural ‘spatial structure’ in its columns. In this section we stay with a generic choice of $N_i$ as well as of the $\psi_k$ and we construct a surrogate following the approach of the groundbreaking works of [13] and [29].

For $t \geq 0$ and $s > 0$ we use the inequality (linear majorization)

$$\sqrt{t} \leq \sqrt{s} + \frac{t - s}{2\sqrt{s}}$$

and apply it in order to compare $\nabla_{ik} K$ with values obtained by an arbitrary non-negative matrix $A$:
\begin{align*}
|\nabla_{ik} K| & \leq |\nabla_{ik} A| + \frac{|\nabla_{ik} K|^2 - |\nabla_{ik} A|^2}{2|\nabla_{ik} A|} \\
& \leq \frac{|\nabla_{ik} K|^2 + |\nabla_{ik} A|^2}{2|\nabla_{ik} A|} \\
& = \frac{2\varepsilon^2_{TV} + \sum_{\ell \in N_i} \left( (K_{ik} - K_{ik})^2 + (A_{ik} - A_{ik})^2 \right)}{2|\nabla_{ik} A|}.
\end{align*}

Summation with respect to $i$, multiplication with $\psi_k$ and summation with respect to $k$ leads to

$$TV(K) \leq \sum_{k=1}^p \psi_k \sum_{i=1}^n \frac{2\varepsilon^2_{TV} + \sum_{\ell \in N_i} \left( (K_{ik} - K_{ik})^2 + (A_{ik} - A_{ik})^2 \right)}{2|\nabla_{ik} A|} =: Q_{TV}^{Oli}(K, A).$$

This yields a candidate for a surrogate $Q_{TV}^{Oli}$ for the TV-penalty term, which is the same as the one used in [29]. However, it is not separable, hence we aim at a second, separable approximation. For arbitrary $a, b, c, d \in \mathbb{R}$ we have

$$\frac{1}{2} \left( (a - b)^2 + (c - d)^2 \right) \leq (a - b)(c - d) + (b - d)^2 + (a - c)^2. \quad (37)$$

This leads to

$$Q_{TV}^{Oli}(K, A) = \sum_{k=1}^p \psi_k \sum_{i=1}^n \frac{\varepsilon^2_{TV} + \sum_{\ell \in N_i} \left( (K_{ik} - K_{ik})^2 + (A_{ik} - A_{ik})^2 \right)}{|\nabla_{ik} A|} \leq \sum_{k=1}^p \psi_k \sum_{i=1}^n \frac{\varepsilon^2_{TV} + \sum_{\ell \in N_i} \left( (K_{ik} - K_{ik})(A_{ik} - A_{ik}) + (K_{ik} - A_{ik})^2 \right)}{|\nabla_{ik} A|} =: Q_{TV}(K, A).$$

Therefore, we have the following

**Theorem 5 (Surrogate Functional for TV Penalty Term)** We consider the cost functional $F(K) := TV(K)$ with the total variation defined in (34). Then

$$Q_{TV}(K, A) := \sum_{k=1}^p \psi_k \sum_{i=1}^n \frac{\varepsilon^2_{TV} + \sum_{\ell \in N_i} \left( (K_{ik} - K_{ik})(A_{ik} - A_{ik}) + (K_{ik} - A_{ik})^2 \right)}{|\nabla_{ik} A|}$$

defines a separable and convex surrogate functional.

The separability of the surrogate is not obvious. The proof (see Appendix C) delivers the following notation, which we also need for an description of the algorithms in the next section. First of all we need the definition of the so-called adjoint neighborhood pixels $N_{\ell_i}$ given by

$$\ell \in N_{\ell_i} \Leftrightarrow i \in N_{\ell}. \quad (38)$$

One then introduces matrices $P(A) \in \mathbb{R}^{n \times p}$ and $Z(A) \in \mathbb{R}^{n \times p}$ via

$$P(A)_{ik} := \frac{1}{|\nabla_{ik} A|} \sum_{\ell \in N_i} 1 + \sum_{\ell \in N_i} \frac{1}{|\nabla_{\ell k} A|}. \quad (39)$$
\[ Z(A)_{ik} := \frac{1}{P(A)_{ik}} \left( \frac{1}{|\nabla_{ik} A|} \sum_{\ell \in N_i} \frac{A_{ik} + A_{\ell k}}{2} + \sum_{\ell \in \bar{N}_i} \frac{A_{ik} + A_{\ell k}}{2|\nabla_{ik} A|} \right). \] (40)

Using these notations, it can be shown that the surrogate can be written as
\[ Q_{TV}(K, A) = \sum_{k=1}^{p} \psi_k \sum_{i=1}^{n} \left[ P(A)_{ik} (K_{ik} - Z(A)_{ik})^2 \right] + C(A), \] (41)
such that we obtain the desired separability. Here, \( C(A) \) denotes some function depending on \( A \). The description of \( Q_{TV} \) with the help of \( P(A)_{ik} \) and \( Z(A)_{ik} \) will also allow us to compute the partial derivatives in a way more comfortable way (see also Appendix A.2).

4.7 Surrogates for Supervised NMF

As a motivation for this section, we consider classification tasks. We view the data matrix \( Y \) as a collection of \( n \) data vectors, which are stored in the rows of \( Y \). Moreover, we do have an expert annotation \( u_i \) for \( i = 1, \ldots, n \), which assigns a label to each data vector. For a classification problem with two classes we have \( u_i \in \{0,1\} \).

As already stated, the rows of the matrix \( X \) of an NMF decomposition can be regarded as a basis for approximating the rows of \( Y \). Hence, one assumes that the correlations between a row \( Y_{i, \cdot} \) of \( Y \) and all row vectors of \( X \), i.e. computing \( Y_{i, \cdot} X^\top \), contains the relevant information of \( Y_{i, \cdot} \). The vector of correlations yields a so-called feature vector of length \( p \). A classical linear regression model, which uses these feature vectors, then asks to compute weights \( \beta_k \) for \( k = 1, \ldots, p \), such that \( Y_{i, \cdot} X^\top \beta \approx u_i \) (for more details on linear discriminant analysis methods, we refer to Chapter 4 in [4]).

In matrix notation and using least squares, this is equivalent to computing \( \beta \) as a minimizer of
\[ \| u - Y X^\top \beta \|^2. \]

We now use \( X \) and \( \beta \) to define
\[ x^* := X^\top \beta. \]

In tumor typing classifications, where the data matrix \( Y \) is obtained by MALDI measurements, the vector \( x^* \) can be interpreted as a characteristic mass spectrum of some specific tumor type and can be directly used for classification tasks in the arising field of digital pathology (see also Section 6 and [26]).

The classification of a new data vector \( y \) is then simply obtained by computing the scalar product \( w = x^* \cdot y \) and assigning either the class label 0 or 1 by comparing \( w \) with a pre-assigned threshold \( s \). This threshold is typically obtained in the training phase of the classification procedure by computing \( Y X^\top \beta \) for some given training data \( Y \) and choosing \( s \), such that a performance measure of the classifier is optimized.

The approach we have described is based on first computing an NMF, i.e. \( K \) and
X, and then computing the weights $\beta$ of the classifier. Hence, the computation of the NMF is done independently of the class labels $u$, which is also referred to as an unsupervised NMF approach. We might expect, that computing the NMF by minimizing a functional which includes the class labels, i.e.

$$F(K, X, \beta) := D_\beta(Y, KX) + \frac{\rho}{2} \| u - YX^\top \beta \|^2,$$

will lead to an improved classifier. In the application field of MALDI imaging, this supervised approach yields an extraction of features from the given training data, which allow a better distinction between spectra obtained from different tissue types such as tumorous and non-tumorous regions (see also [26]). Surrogates for the first term have been determined in the previous section for the case of the Frobenius norm and the Kullback-Leibler divergence. Hence, we need to determine surrogates of the new penalty term for minimization with respect to $X$ and $\beta$:

$$F(X) := \frac{1}{2} \| u - YX^\top \beta \|^2,$$

$$G(\beta) := \frac{1}{2} \| u - YX^\top \beta \|^2.$$  

(42)

(43)

Surrogates can be obtained by extending the Jensen principle to the matrix valued case. Here, we consider a convex subset $\Omega \subset \mathbb{R}^{N \times M}_{>0}$ and define

$$\tilde{F} : \Omega \to \mathbb{R}$$

$$V \mapsto f(c^\top V d)$$

(44)

with a convex and continuously differentiable function $f$ and auxiliary variables $c \in \mathbb{R}^N_{>0}$ and $d \in \mathbb{R}^M_{>0}$. We now use the following generalized Jensen’s inequality

$$f \left( \sum_{j=1}^N \sum_{k=1}^M \lambda_{jk} \alpha_{jk} \right) \leq \sum_{j=1}^N \sum_{k=1}^M \lambda_{jk} f(\alpha_{jk}).$$

(45)

Setting

$$\lambda_{jk} = \frac{Y_{ij}A_{kj}\beta_k}{Y_{ij}A_{ij}^\top \beta}$$

$$\alpha_{jk} = \frac{Y_{ij}X_{kj}\beta_k}{\lambda_{jk}}.$$ 

(46)

(47)

for some $i \in \{1, \ldots, n\}$ yields by inserting $\lambda_{jk}$ and $\alpha_{jk}$ into (45)

$$F(X) \leq \frac{1}{2} \sum_{i=1}^n \frac{1}{(YA^\top \beta)_i} \sum_{j=1}^N \sum_{k=1}^M Y_{ij}A_{kj}\beta_k \left( u_i - \frac{X_{kj}(YA^\top \beta)_i}{A_{kj}} \right)^2 := Q_F(X, A).$$

The computation of a surrogate for minimization with respect to $\beta$ proceeds analogously. We summarize the results in the following theorem.
Theorem 6 (Surrogate Functionals for Linear Regression) Let \( F(X) := \frac{1}{2} \| u - YX^\top \beta \|^2 \) und \( G(\beta) := \frac{1}{2} \| u - YX^\top \beta \|^2 \) denote a cost functional with respect to \( X \) and \( \beta \). Then

\[
Q_F(X, A) := \frac{1}{2} \sum_{i=1}^{n} \frac{1}{(YA)^i} \sum_{j=1}^{m} \sum_{k=1}^{p} Y_{ij} A_{kj} \beta_k \left( u_j - \frac{X_{kj}}{A_{kj}} (Y A^\top \beta)_i \right)^2 \tag{48}
\]

\[
Q_G(\beta, a) := \frac{1}{2} \sum_{i=1}^{n} \frac{1}{(YX^\top a_i)} \sum_{j=1}^{p} (YX^\top)_{ik} a_k \left( u_i - \frac{\beta_k}{a_k} (YX^\top a)_i \right)^2 \tag{49}
\]

define separable and convex surrogate functionals.

A big advantage of linear regression models is their simplicity and manageability. However, they are far from the optimal approach to approximate the binary output data \( u \) with a continuous input. Logistic regression models offer a way more natural method for binary classification tasks. Together with the supervised NMF as a feature extraction method, this overall workflow leads in \cite{26} to excellent classification results and outperformed classical approaches.

However, the proposed model is based on a gradient descent approach, such that the non-negativity of the iterates can only be guaranteed by a projection step. Appropriate surrogate functionals for this workflow are still ongoing research and could lead to even better outcomes (see also \cite{35,36}).

5 Surrogate Based NMF Algorithms

In the previous section we have defined surrogate functionals for various NMF cost functions. Besides the necessary surrogate properties we also expect that the minimization of these surrogates is straightforward and can be computed efficiently. In our case we demand additionally, that the minimization schemes based on solving the first order optimality conditions leads to a separable algorithm and that it only requires multiplicative updates, which automatically preserve the non-negativity of its iterates. Let us start with denoting the most general functional with Kullback-Leibler divergence, the Frobenius case follows similarly.

For constructing non-negative matrix factorizations, we incorporate \( \ell_2 \)-, \( TV \)-, orthogonality-\(, \sigma_k \)-, sparsity-\(, \omega \)-, \( \tau \)-, and also the penalty terms coming from the supervised NMF. Of course, in most applications one only uses a subset of these constraints for stabilization and for enhancing certain properties. These algorithms can readily be obtained from the general case by putting the respective regularization parameters to zero. The corresponding update rules are classical results and can be found in numerous works \cite{11,24,25}.

Definition 9 (NMF Problem) For \( Y \in \mathbb{R}_{\geq 0}^{n \times m}, K, V \in \mathbb{R}_{\geq 0}^{n \times p}, X, W \in \mathbb{R}_{\geq 0}^{p \times m}, \beta \in \mathbb{R}^p, \) and a set of regularization parameters \( \lambda, \mu, \nu, \omega, \tau, \sigma_{K,1}, \sigma_{K,2}, \sigma_{X,1}, \sigma_{X,2}, \rho \geq 0 \), we define the NMF minimization problem by

\[
\min_{K, X, V, W, \beta \geq 0} \left\{ \text{KL}(Y, KX) + \lambda \| X \|_1 + \frac{\mu}{2} \| K \|_F^2 + \frac{\nu}{2} \| X \|_F^2 + \omega \| K \|_1 + \frac{\tau}{2} \| V \|_F^2 + \frac{\sigma_{K,1}}{2} \| I - V^\top K \|_F^2 + \frac{\sigma_{K,2}}{2} \| V - K \|_F^2 + \frac{\sigma_{X,1}}{2} \| I - X W^\top \|_F^2 + \frac{\sigma_{X,2}}{2} \| W - X \|_F^2 + \frac{\rho}{2} \| u - YX^\top \beta \|_F^2 \right\}
\]
The choice of the various regularization parameters occurring in Definition 9 is often based on heuristic approaches. We will not focus on that issue in this work and refer instead to [19] and the references therein, where two methods are investigated for the general case of multi-parameter Tikhonov regularization.

The algorithms studied in this section will start with positive initializations for $K, X, V, W$ and $\beta$. These matrices are updated alternatingly, i.e. all matrices except one matrix are kept fixed and only the selected matrix is updated by solving the respective first order optimality condition.

We will focus in this section on the derivation of the update rules of $K$ (see also Appendix A.2). The iteration schemes for the other matrices follow analogously. For that, we only have to consider those terms in the general functional which depend on $K$, i.e. we aim at determining a minimizer for

$$F(K) := KL(Y, KX) + \frac{\mu}{2} \|K\|_F^2 + \omega \|K\|_1 + \frac{\tau}{2} TV(K) + \frac{\sigma_{K,1}}{2} \|I - V^T K\|_F^2 + \frac{\sigma_{K,2}}{2} \|V - K\|_F^2.$$ 

Instead of minimizing this functional we exchange it with the previously constructed surrogate functionals, which leads to

$$Q_F(K, A) := Q_{KL}(K, A) + \frac{\mu}{2} \|K\|_F^2 + \omega \|K\|_1 + \frac{\tau}{2} Q_{TV}(K, A) + Q_{Orth}(K, A)$$

with the surrogates $Q_{KL}$ for the Kullback-Leibler divergence in Theorem 3, $Q_{TV}$ for the TV penalty term in Theorem 5 and $Q_{Orth}$ for the orthogonality penalty terms in Theorem 4.

The computation of the partial derivatives leads to a system of equations

$$\frac{\partial Q_F}{\partial K_{\xi \zeta}}(K, A) = 0$$

for $\xi \in \{1, \ldots, n\}$ and $\zeta \in \{1, \ldots, p\}$. This leads to a system of quadratic equations

$$K_{\xi \zeta}^2 \left( \mu + \tau \psi_{\xi \zeta} P_\zeta(A) + \frac{\sigma_{K,1}}{A_{\xi \zeta}} (V^T A)_{\xi \zeta} + \sigma_{K,2} \right) + K_{\xi \zeta} \sum_{j=1}^m X_{\zeta j} + \omega - \tau \psi_{\xi \zeta} P(A)_{\xi \zeta} Z(A)_{\xi \zeta} - (\sigma_{K,1} + \sigma_{K,2}) V_{\xi \zeta}$$

$$= A_{\xi \zeta} \sum_{j=1}^m \frac{Y_{\xi j}}{(AX)_{\xi j}} X_{\zeta j}.$$ 

Solving for $K_{\xi \zeta}$ and denoting the Hadamard product by $\odot$ as well as the matrix division for each entry separately by a fraction line, yields the following update rule for $K$. (Note that the notation for $P(A)$ and $Z(A)$ was introduced in the
section on TV-regularization above.)

\[
K^{[d+1]} = \left( \frac{K^{[d]}}{\nu I_{n \times p} + \tau \Psi \circ P(K^{[d]}) + \sigma_{K,1} \frac{VV^\top K^{[d]}}{K^{[d]}} + \sigma_{K,2} I_{n \times p}} \right) \circ \left( \frac{Y}{X^\top} \right) \\
= \begin{bmatrix}
\begin{pmatrix}
K^{[d]}
\mu I_{n \times p} + \tau \Psi \circ P(K^{[d]}) + \sigma_{K,1} \frac{VV^\top K^{[d]}}{K^{[d]}} + \sigma_{K,2} I_{n \times p}
\end{pmatrix}
\end{bmatrix}
\end{bmatrix}
\]

\[
= \left( \frac{1}{4} \left( 1_{n \times m} X^\top + \omega 1_{n \times p} - \tau \Psi \circ P(K^{[d]}) \circ Z(K^{[d]}) - (\sigma_{K,1} + \sigma_{K,2}) V \\
\mu I_{n \times p} + \tau \Psi \circ P(K^{[d]}) + \sigma_{K,1} \frac{VV^\top K^{[d]}}{K^{[d]}} + \sigma_{K,2} I_{n \times p}
\right) \right)^{1/2}
\]

In the above update rule, \( 1_{n \times p} \) denotes an \( n \times p \) matrix with ones in every entry and \( \Psi \in \mathbb{R}^{n \times p} \) is defined as

\[
\Psi := \begin{pmatrix}
\psi_1 & \psi_2 & \cdots & \psi_p \\
\psi_1 & \psi_2 & \cdots & \psi_p \\
\vdots & \vdots & \ddots & \vdots \\
\psi_1 & \psi_2 & \cdots & \psi_p
\end{pmatrix}
\]

Details on the derivation can be found in Appendix A.2

The partial derivatives with respect to \( X \) are computed similarly. Defining

\[
A^{[d]} := \begin{pmatrix}
X^{[d]}
\nu 1_{p \times m} + \sigma_{X,1} \frac{X^{[d]} W^\top W}{X^{[d]}} + \sigma_{X,2} 1_{p \times m} + \rho \beta \beta^\top \frac{X^{[d]} Y^\top Y}{X^{[d]}}
\end{pmatrix} \circ \begin{pmatrix}
K^\top \ Y \\
K \ X^{[0]}
\end{pmatrix}
\]

\[
\Gamma^{[d]} := \begin{pmatrix}
K^\top 1_{n \times m} + \lambda I_{p \times m} - (\sigma_{X,1} + \sigma_{X,2}) W - \rho \beta u^\top Y \\
\nu 1_{p \times m} + \sigma_{X,1} \frac{X^{[d]} W^\top W}{X^{[d]}} + \sigma_{X,2} 1_{p \times m} + \rho \beta \beta^\top \frac{X^{[d]} Y^\top Y}{X^{[d]}}
\end{pmatrix}
\]

leads to the update

\[
X^{[d+1]} \left( \frac{\sqrt{A^{[d]}} + \frac{1}{4} \Gamma^{[d]} \circ \Gamma^{[d]} - \frac{1}{2} \Gamma^{[d]} }{4}
\right).
\]

The updates for \( V, W \) are straightforward and we obtain the following theorem.

**Theorem 7 (Alternating Algorithm for NMF Problem in Definition 9)**

The initializations \( K^{[0]}, V^{[0]} \in \mathbb{R}_{>0}^{n \times p}, X^{[0]}, W^{[0]} \in \mathbb{R}_{>0}^{p \times m}, \beta^{[0]} \in \mathbb{R}_{>0}^p \) and the iterative updates

\[
V^{[d+1]} = \left( \frac{\sigma_{K,1} + \sigma_{K,2} K^{[d]} }{\sigma_{K,1} K^{[d]} K^{[d \top]} V^{[d]} + \sigma_{K,2} 1_{n \times p}}
\right)
\]

(51)
The classical update rule for $X$ to zero. In the case of $\ell_2$, this leads to the cost functional

$$F(K, X, V, W, \beta) := \text{KL}(Y, KX) + \lambda \|X\|_1 + \frac{\mu}{2} \|K\|_F^2 + \frac{\nu}{2} \|X\|_F^2 + \omega \|K\|_1$$

It is easy to see that the classical, regularized NMF algorithms described in [11, 24, 25] can be regained by putting the corresponding regularization parameters to zero. In the case of $\ell_1$- and $\ell_2$-regularized NMF, this leads to the cost function

$$F(K, X) = \text{KL}(Y, KX) + \lambda \|X\|_1 + \frac{\mu}{2} \|K\|_F^2 + \frac{\nu}{2} \|X\|_F^2 + \omega \|K\|_1.$$ 

The classical update rule for $X$ is obtained by setting

$$\tilde{A}^{[d]} := X^{[d]} \circ \left( K^{[d+1]^\top} \frac{Y}{K^{[d+1]} X^{[d]}} \right),$$

$$\tilde{\Gamma}^{[d]} := K^{[d+1]^\top} I_{n \times m} + \lambda I_{p \times m},$$

which - in connection with the update rule for $X$ of the previous theorem - leads to

$$X^{[d+1]} = \sqrt{\frac{1}{\beta} \tilde{A}^{[d]} + \frac{1}{4\beta^2} \tilde{\Gamma}^{[d]} \circ \tilde{\Gamma}^{[d]} - \frac{1}{2\beta^2} \tilde{\Gamma}^{[d]}}$$

$$= \frac{2 \tilde{A}^{[d]}}{\tilde{\Gamma}^{[d]} + \sqrt{4\beta^2 \tilde{A}^{[d]} + \tilde{\Gamma}^{[d]} \circ \tilde{\Gamma}^{[d]}}}$$

which is the update rule described in [11].

By the same approach and with the surrogate functionals derived in Section [3] we obtain the update rules for the Frobenius discrepancy term, i.e. we consider the functional

$$F(K, X, V, W, \beta) := \frac{1}{2} \|Y - KX\|_F^2 + \lambda \|X\|_1 + \frac{\mu}{2} \|K\|_F^2 + \frac{\nu}{2} \|X\|_F^2 + \omega \|K\|_1$$

$$+ \frac{\sigma K_{1,1}}{2} \|I - V \top K\|_F^2 + \frac{\sigma K_{2,2}}{2} \|V - K\|_F^2$$

$$+ \frac{\sigma X_{1,1}}{2} \|I - X W \top\|_F^2 + \frac{\sigma X_{2,2}}{2} \|W - X\|_F^2 + \frac{\rho}{2} \|u - YX \top \beta\|^2$$
A monotone decrease of this functional is obtained by the following iteration in combination with the update rules for $V$, $W$, $\beta$ as in Theorem 7 (see also Appendix A.1 for more details on the derivation of these algorithms.)

$$ K^{[d+1]} = \frac{YX^{[d]} + \tau \Psi \circ P(K^{[d]}) \circ Z(K^{[d]}) + (\sigma_{K,1} + \sigma_{K,2}) V^{[d+1]}}{\tau \Psi \circ P(K^{[d]}) + \sigma_{K,2} I_{n \times p} + \mu K^{[d]} + \omega I_{n \times p} + \sigma_{K,1} V^{[d+1]} V^{[d+1] \top} Y^{\top} Y + \sigma_{X,1} X^{[d]} W^{[d+1] \top} W^{[d+1]}} $$

$$ X^{[d+1]} = \frac{K^{[d+1]} \top Y + (\sigma_{X,1} + \sigma_{X,2}) W^{[d+1]} \top \rho \beta^{[d]} u \top Y}{\sigma_{X,2} I_{p \times m} + \rho \beta^{[d]} \beta^{[d]} \top X^{[d]} Y \top Y + \sigma_{X,1} X^{[d]} W^{[d+1] \top} W^{[d+1]}} $$

6 MALDI Imaging

As a test case we analyse MALDI imaging data (matrix assisted laser desorption/ionization) of a rat brain. MALDI imaging is a comparatively novel modality, which unravels the molecular landscape of tissue slices and allows a subsequent proteomic or metabolic analysis [1,6,22]. Clustering this data reveals for example different metabolic regions of the tissue, which can be used for supporting pathological diagnosis of tumors.

The data used in this paper was obtained by a MALDI imaging experiment, see Figure 2 for a schematic experimental setup.

In our numerical experiments, we used a classical rat brain dataset which has been used in several data processing papers before [2,3,22]. It constitutes a standard test set for hyperspectral data analysis.

The tissue slice was scanned at 20185 positions. At each position a full mass spectrum with 2974 m/z (mass over charge) values was collected. I.e. instead of three color channels, as it is usual in image processing, this data has 2974 channels, each channel containing the spatial distribution of molecules having the same m/z value.

The following numerical examples were obtained with the multiplicative algorithms described in the previous section. We just illustrate the effect of the different penalty terms for some selected functionals. One can either display the columns of $K$ as the pseudo channels of the NMF decomposition or the rows of $X$ as pseudo spectra characterizing the different metabolic processes present in the tissue slice, see the Figures 3-6.

Both ways of visualization do have their respective value. Looking at the pseudo spectra in connection with orthogonality constraints leads to a clustering of the spectra and to a subdivision of the tissue slice in regions with potentially different metabolic activities, see [22]. Considering instead the different pseudo spectra, which were constructed in order to have a bases which allows a low dimensional approximaton of the data set, is the basis for subsequent proteomic analysis. E.g. one may target pseudospectra where the related pseudo channels are concentrated in regions, which were annotated by pathological experts. Mass values which are dominant in those spectra may stem from proteins/peptides relating to biomarkers as indicators for certain diseases. Hence, classification schemes based on NMF decompositions have been widely investigated [28,90,54].
Fig. 2: Structure of MALDI imaging data: A mass spectrum is obtained at different positions of a tissue slice. The full data set is a data cube, which can be visualized with different perspectives. Fixing a position of the tissue slice gives the mass spectrum at this position. Fixing a particular molecular weight reveals the distribution of molecules across the tissue slice with this weight.

Fig. 3: NMF of the rat brain dataset for $p = 6$. Orthogonality constraints on the channels with $\sigma_{K,1} = 1$ and $\sigma_{K,2} = 1$. 
Fig. 4: NMF of the rat brain dataset for $p = 6$. Orthogonality constraints on the channels with $\sigma_{K,1} = 200$ and $\sigma_{K,2} = 200$.

7 Conclusion

In this paper, we investigated methods based on surrogate minimization approaches for the solution of NMF problems. The interest in NMF methods is related to its importance for several machine learning tasks. Application for large data sets require that the resulting algorithms are very efficient and that iteration schemes only need simple matrix-vector multiplications.

The state of the art for constructing appropriate surrogates are based on case-by-case studies for the different, considered NMF models. In this paper, we embedded the different approaches in a general framework, which allowed us to analyze several extensions to the NMF cost functional, including $\ell_1$- and $\ell_2$-regularization, orthogonality constraints, total variation penalties as well as extensions, which led to supervised NMF concepts.

Secondly, we analyzed surrogates in the context of the related iteration schemes, which are based on first order optimality conditions. The requirement of separability as well as the need of having multiplicative updates, which preserve non-negativity without additional projections, were analyzed. This resulted in a general description of algorithms for alternating minimization of constrained NMF functionals. The potential of these methods is confirmed by numerical tests using hyperspectral data from a MALDI imaging experiment.

Several further directions of research would be of interest. First of all, besides the most widely used penalty terms discussed in this paper further penalty terms, e.g. higher order TV-terms, could be considered. Secondly, construction principles for
more general discrepancy terms could be analyzed (see also [16]).

Potentially more importantly, this paper contains only very first results for combining NMF constructions directly with subsequent classification tasks. The question of an appropriate surrogate functional for the supervised NMF model with logistic regression used in [26] remains unanswered and also the comparison with algorithmic alternatives such as ADMM methods needs to be explored.

Acknowledgements The authors want to thank Christine DeMol: her excellent presentations at several conferences and our joint discussions were the starting point for this paper. The authors would also like to thank the German Science Foundation. The work on this paper was funded by project no. GRK 2224/1 (Research training group π parameter identification: analysis, algorithms, applications).

The results in this paper are based on the Master thesis of the first author.

References

1. Aebersold, R., Goodlett, D.: Mass spectrometry in proteomics. Chem. Rev. 101(2), 269–295 (2001)
2. Alexandrov, T., Bartels, A.: Testing for presence of known and unknown molecules in imaging mass spectrometry. Bioinformatics 29(18), 2335–2342 (2013)
3. Alexandrov, T., Becker, M., Deininger, S., Ernst, G., Wehder, L., Grasmair, M., von Eggeling, F., Thiele, H., Maass, P.: Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. Journal of Proteome Research 9(12), 6535–6546 (2010)
Fig. 6: NMF of the rat brain dataset for $p = 6$. Orthogonality constraints on the channels with $\sigma_{K,1} = 10$ and $\sigma_{K,2} = 10$ and sparsity penalty term on $X$ with $\lambda = 0.06$. The sparsity penalty term has in connection with the orthogonality constraint a comparatively strong influence on the NMF computation: The sparsity in the spectra increases significantly and thus their biological interpretability, whereas the anatomic structure in the pseudo channels diminishes.

4. Bishop, C.: Pattern Recognition and Machine Learning. Springer-Verlag New York (2006)
5. Böhning, D., Lindsay, B.G.: Monotonicity of quadratic approximation algorithms. Annals of the Institute of Statistical Mathematics 40(4), 641–663 (1988)
6. Caprioli, R., Terry, B., Gile, J.: Molecular imaging of biological samples: localization of peptides and proteins using maldi-tof ms. Anal. Chem. 69(23), 4751–4760 (1997)
7. Chambolle, A., Caselles, V., Novaga, M., Cremers, D., Pock, T.: An introduction to total variation for image analysis. Theoretical Foundations and Numerical Methods for Sparse Recovery 9, 263–340 (2010)
8. Chan, T., Esedoglu, S., Park, F., Yip, A.: Recent developments in total variation image restoration. In: In Mathematical Models of Computer Vision. Springer Verlag (2005)
9. Condat, L.: Discrete total variation: New definition and minimization. Research report, GIPS-lab (2016)
10. Cvetkovski, Z.: Inequalities - Theorems, Techniques and Selected Problems. Springer-Verlag Berlin Heidelberg (2012)
11. De Mol, C.: Regularized multiplicative algorithms for nonnegative matrix factorization. Methodological Aspects of Hyperspectral Imaging Workshop (2013)
12. De Pierro, A.R.: On the relation between the isra and the em algorithm for positron emission tomography. IEEE Transactions on Medical Imaging 12(2), 328–333 (1993)
13. Defrise, M., Vanhove, C., Liu, X.: An algorithm for total variation regularization in high-dimensional linear problems. Inverse Problems 27(6) (2011)
14. Fessler, J.A.: Statistical image reconstruction methods for transmission tomography. In: M. Sonka, J. Fitzpatrick (eds.) Handbook of Medical Imaging, Volume 2. Medical Image Processing and Analysis, pp. 1–70. SPIE (2000)
15. Févotte, C., Bertin, N., Durrieu, J.L.: Nonnegative matrix factorization with the itakura-saito-divergence: With application to music analysis. Neural Computation 21(3), 793–830 (2009)
16. Févotte, C., Idier, J.: Algorithms for nonnegative matrix factorization with the beta-divergence. Neural Computation 23(9), 2421–2456 (2011)
17. Hennquin, R., David, B., Badeau, R.: Beta-divergence as a subclass of bregman divergence. IEEE Signal Processing Letters 18(2), 83–86 (2011)
18. Hunter, D.R., Lange, K.: A tutorial on mm algorithms. The American Statistician 58(1), 30–37 (2004)
19. Ito, K., Jin, B., Takeuchi, T.: Multi-parameter tikhonov regularization – an augmented approach. Chinese Annals of Mathematics, Series B 35(3), 383–398 (2014)
20. Jin, B., Lorenz, D.A., Schiffler, S.: Elastic-net regularization: error estimates and active set methods. Inverse Problems 25(11) (2009)
21. Jin, B., Maass, P.: Sparsity regularization for parameter identification problems. Inverse Problems 28(12) (2012)
22. Kobarg, J.H., Maass, P., Oetjen, J., Hirsch, E., Sagiv, C., Golabbae, M., Vanderheynst, P.: Numerical experiments with maldi imaging data. Advances in Computational Mathematics 40(3), 667–682 (2014)
23. Lange, K.: Optimization, Springer Texts in Statistics, vol. 95, 2 edn. Springer-Verlag New York (2013)
24. Lecharlier, L., Mol, C.D.: Regularized blind deconvolution with poisson data. Journal of Physics: Conference Series 464(1) (2013)
25. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
26. Leuschner, J., Fersnel, P., Schmidt, M., Lachmund, D., Boskamp, T., Maass, P.: Supervised non-negative matrix factorization methods with maldi-imaging applications. Bioinformatics (in review) (2018)
27. Li, T., Ding, C.: Non-negative matrix factorization for clustering: A survey. In: Data Clustering: Algorithms and Applications, pp. 149–176 (2013)
28. Louis, A.K.: Inverse und schlecht gestellte Probleme. Vieweg+Teubner Verlag (1989)
29. Oliveira, J.P., Bioucas-Dias, J.M., Figueiredo, M.A.T.: Review: Adaptive total variation image deblurring: A majorization-minimization approach. Signal Process. 89(9), 1683–1693 (2009)
30. Phon-Amnuaisuk, S.: Applying non-negative matrix factorization to classify superimposed handwritten digits. Procedia Computer Science 24, 261–267 (2013)
31. Rudin, L.L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D 60(1-4), 259–268 (1992)
32. Sun, D.L., Févotte, C.: Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6201–6205 (2014)
33. Tan, V.Y.F., Févotte, C.: Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. arXiv preprint arXiv:1111.6085v3 (2012)
34. Tang, J., Ceng, X., Peng, B.: New methods of data clustering and classification based on nmf. In: 2011 International Conference on Business Computing and Global Informationization, pp. 432–435 (2011)
35. Zhang, Z., Kwok, J.T., Yeung, D.Y.: Surrogate maximization/minimization algorithms for adaboost and the logistic regression model. In: Proceedings of the Twenty-first International Conference on Machine Learning, ICML ’04 (2004)
36. Zhang, Z., Kwok, J.T., Yeung, D.Y.: Surrogate maximization/minimization algorithms and extensions. Machine Learning 69(1), 1–33 (2007)
Appendix A  Details on the Derivation of the Algorithms in Section 5

In this section, we give a more detailed derivation of the algorithms presented in Section 5. We start with the less complex case of the Frobenius norm as discrepancy term and then turn to the Kullback-Leibler divergence. To cover both aspects, we derive the update rules of $X$ for the Frobenius discrepancy term and of $K$ in the case of the KLD. We will also take a closer look at the effect of $\kappa$ in equation (18) with respect to the LQBP construction principle.

A.1 Frobenius Norm

We consider the general cost function described in Section 5 for the case of the Frobenius norm. To compute the update rules for $X$, it is enough to examine the function

$$F(X) := \frac{1}{2} \| Y - KX \|_F^2 + \lambda \| X \|_1 + \frac{\nu}{2} \| X \|_F^2,$$

where all terms independent from $X$ are omitted. Based on Remark 1 and following the discussion of Section 4.4, the construction of a surrogate for $F$ can be done separately for $F_1$ and the remaining penalty terms.

The construction of a surrogate for $F_1$ with the LQBP principle as it has been done similarly in Subsection 4.4 is essential. If we would use instead a surrogate for the discrepancy term $\frac{1}{2} \| Y - KX \|_F^2$ from Subsection 4.1 or 4.2 and take the $\ell_1$-penalty term $\lambda \| X \|_1$ as surrogate itself, it is easy to see that this would not lead to multiplicative update rules. It is the $\ell_1$-penalty term which causes the difficulty. Computing the first order optimality condition for the corresponding surrogate $\tilde{Q}_F(X, A) := \lambda \| X \|_1 + \tilde{Q}_F(X, A)$ with respect to $X$ would lead to

$$0 = \frac{\partial \tilde{Q}_F}{\partial X_{\xi \zeta}}(X, A) = \lambda + \frac{\partial \tilde{Q}_F}{\partial X_{\xi \zeta}}(X, A),$$

where the second term on the right hand side does not depend on $\lambda$. Hence, we get a sign in front of $\lambda$ by solving the equation for $X_{\xi \zeta}$ and we will not obtain multiplicative updates for $X$.

A correct surrogate is obtained by using the LQBP principle to $F_1$ and leads to

$$Q_F(X, A) := Q_{F_1}(X, A) + \frac{\nu}{2} \| X \|_F^2 + Q_{Orth}(X, A) + Q_{LR}(X, A)$$

with

$$Q_{F_1}(X, A) = \sum_{j=1}^m f_{Y_{\ast, j}}(A_{\ast, j}) + \nabla f_{Y_{\ast, j}}(A_{\ast, j})^T(X_{\ast, j} - A_{\ast, j})$$

$$+ \frac{1}{2}(X_{\ast, j} - A_{\ast, j})^T A_{f_{Y_{\ast, j}}}(A_{\ast, j})(X_{\ast, j} - A_{\ast, j}),$$
where $f_{Y_{*,j}} : \mathbb{R}_{\geq 0}^p \to \mathbb{R}$ is defined as

$$f_{Y_{*,j}}(x) := \frac{1}{2}\|Y_{*,j} - Kx\|^2 + \lambda\|x\|_1$$

and with the diagonal matrix

$$A_{f_{Y_{*,j}}}(A_{*,j})_{kk} = \frac{(\nabla^2 f_{Y_{*,j}}(A_{*,j}) A_{*,j})_k + \kappa_k}{A_{kk}} = \frac{(K^T K A_{*,j})_k + \kappa_k}{A_{kk}}.$$

The functionals $Q_{Orth}$ resp. $Q_{LR}$ are the surrogates obtained from Theorem 4 resp. Theorem 6. It will turn out that an appropriate choice of $\kappa_k$ will ensure a multiplicative NMF algorithm.

The computation of the first order optimality condition for $Q_F$ leads to

$$0 = \frac{\partial Q_F(X, A)}{\partial X_{\xi \zeta}} = (K^T K A)_{\xi \xi} - (K^T Y)_{\xi \zeta} + \lambda + \frac{(K^T K A)_{\xi \zeta} + \kappa_{\xi}(X_{\xi \zeta} - A_{\xi \zeta})}{A_{\xi \xi}}$$

$$+ \sigma_{X,1} \sum_{k=1}^p W_k \left( \frac{X_{\xi \zeta}(AW^T)_{\xi k} - \delta_{\xi k}}{A_{\xi \xi}} \right) + \sigma_{X,2}(X_{\xi \zeta} - W_{\xi \zeta})$$

$$+ \rho \beta \sum_{i=1}^n Y_{i \xi} \left( \frac{X_{\xi \zeta}(YA^T\beta)_i - u_i}{A_{\xi \xi}} \right) + \nu X_{\xi \zeta}.$$ 

One can see immediately, that the choice of $\kappa_{\xi} := \lambda$ for all $\xi \in \{1, \ldots, p\}$ is appropriate to get rid of the problematic term $\lambda$. Hence, we obtain

$$0 = -(K^T Y)_{\xi \zeta} + \frac{X_{\xi \zeta}}{A_{\xi \xi}} ((K^T K A)_{\xi \xi} + \lambda)$$

$$+ \sigma_{X,1} \sum_{k=1}^p W_k \left( \frac{X_{\xi \zeta}(AW^T)_{\xi k} - \delta_{\xi k}}{A_{\xi \xi}} \right) + \sigma_{X,2}(X_{\xi \zeta} - W_{\xi \zeta})$$

$$+ \rho \beta \sum_{i=1}^n Y_{i \xi} \left( \frac{X_{\xi \zeta}(YA^T\beta)_i - u_i}{A_{\xi \xi}} \right) + \nu X_{\xi \zeta}.$$

Reordering the terms leads to

$$(K^T Y)_{\xi \zeta} + (\sigma_{X,1} + \sigma_{X,2})W_{\xi \zeta} + \rho \beta (Y^T u)_\zeta$$

$$= \frac{X_{\xi \zeta}}{A_{\xi \xi}} ((K^T K A)_{\xi \xi} + \nu A_{\xi \xi} + \lambda + \rho \beta_{\xi}(Y^T Y A^T \beta)_\zeta + \sigma_{X,1}(AW^T W)_{\xi \zeta} + \sigma_{X,2} A_{\xi \xi}).$$

Solving for $X_{\xi \zeta}$ and extending the equation to the whole matrix $X$ yields finally

$$X = A \odot \frac{K^T Y + (\sigma_{X,1} + \sigma_{X,2}) W + \rho \beta u^T Y}{K^T K A + (\sigma_{X,2} + \nu) A + \lambda I_{p \times m} + \rho \beta \beta^T AY^T Y + \sigma_{X,1} AW^T W}.$$ 

By exploiting the surrogate minimization principle as described in Lemma 1 we get finally the update rule for $X$ presented in Section 5.
A.2 Kullback-Leibler Divergence

We take Equation (50) as our starting point. The computation of the first order optimality condition gives

$$\frac{\partial Q}{\partial K_{\xi\zeta}} = \sum_{j=1}^{m} \left[ X_{\xi j} - \frac{Y_{\xi j}}{(AX)_{\xi j}} A_{\xi\xi} X_{\xi j} \right] + \mu K_{\xi\zeta} + \omega + \sigma_{K,2} (K_{\xi\zeta} - V_{\xi\zeta})$$

$$+ \tau \psi \xi P(A)_{\xi\zeta} (K_{\xi\zeta} - Z(A)_{\xi\zeta}) + \sigma_{K,1} \sum_{k=1}^{p} V_{\xi k} \left( \frac{K_{\xi\zeta} (V^T A)_{k\zeta} - \delta_{k\zeta}}{A_{\xi\xi}} \right)$$

$$= 0.$$ 

Multiplying on both sides with $K_{\xi\zeta}$ and sorting the terms already gives the system of quadratic equations mentioned in Section 5, namely

$$K_{\xi\zeta}^2 \left( \mu + \tau \psi \xi P(A)_{\xi\zeta} + \frac{\sigma_{K,1}}{A_{\xi\xi}} (VV^T A)_{\xi\zeta} + \sigma_{K,2} \right)$$

$$+ K_{\xi\zeta} \left( \sum_{j=1}^{m} X_{\xi j} + \omega - \tau \psi \xi P(A)_{\xi\zeta} Z(A)_{\xi\zeta} - (\sigma_{K,1} + \sigma_{K,2}) V_{\xi\zeta} \right)$$

$$= A_{\xi\xi} \sum_{j=1}^{m} \frac{Y_{\xi j}}{(AX)_{\xi j}} X_{\xi j}. $$

Taking into account that

$$\sum_{j=1}^{m} \frac{Y_{\xi j}}{(AX)_{\xi j}} X_{\xi j} = \left( \frac{Y}{AX} X^T \right)_{\xi\zeta} \quad \text{and} \quad \sum_{j=1}^{m} X_{\xi j} = (1_{n \times m} X^T)_{\xi\zeta},$$

we obtain the explicit solution of $K_{\xi\zeta}$ by completing the square and get

$$K_{\xi\zeta} = \left[ \frac{A_{\xi\xi}}{\mu + \tau \psi \xi P(A)_{\xi\zeta} + \frac{\sigma_{K,1}}{A_{\xi\xi}} (VV^T A)_{\xi\zeta} + \sigma_{K,2}} \left( \frac{Y}{AX} X^T \right)_{\xi\zeta} \right]^{1/2}$$

$$+ \frac{1}{4} \left( \frac{(1_{n \times m} X^T)_{\xi\zeta} + \omega - \tau \psi \xi P(A)_{\xi\zeta} Z(A)_{\xi\zeta} - V_{\xi\zeta} (\sigma_{K,1} + \sigma_{K,2})}{\mu + \tau \psi \xi P(A)_{\xi\zeta} + \frac{\sigma_{K,1}}{A_{\xi\xi}} (VV^T A)_{\xi\zeta} + \sigma_{K,2}} \right)$$

$$- \frac{1}{2} \left( \frac{(1_{n \times m} X^T)_{\xi\zeta} + \omega - \tau \psi \xi P(A)_{\xi\zeta} Z(A)_{\xi\zeta} - V_{\xi\zeta} (\sigma_{K,1} + \sigma_{K,2})}{\mu + \tau \psi \xi P(A)_{\xi\zeta} + \frac{\sigma_{K,1}}{A_{\xi\xi}} (VV^T A)_{\xi\zeta} + \sigma_{K,2}} \right)^{1/2}. $$
This equation holds for arbitrary $\xi \in \{1, \ldots, n\}$ and $\zeta \in \{1, \ldots, k\}$. We therefore can extend this relation to the whole matrix $K$ and obtain

$$
K = \left[ \begin{array}{c}
\mu I_{n \times p} + \tau \Psi \circ P(A) + \sigma_{K,1} \frac{VV^T A}{A} + \sigma_{K,2} I_{n \times p}
\end{array} \right] \circ \left( \begin{array}{c}
Y A^T X^T
\end{array} \right)
$$

$$
+ \frac{1}{4} \left[ \begin{array}{c}
I_{n \times m} X^T + \omega I_{n \times p} - \tau \Psi \circ P(A) \circ Z(A) - (\sigma_{K,1} + \sigma_{K,2}) V
\end{array} \right]
\mu I_{n \times p} + \tau \Psi \circ P(A) + \sigma_{K,1} \frac{VV^T A}{A} + \sigma_{K,2} I_{n \times p}
$$

$$
- \frac{1}{2} \left[ \begin{array}{c}
I_{n \times m} X^T + \omega I_{n \times p} - \tau \Psi \circ P(A) \circ Z(A) - (\sigma_{K,1} + \sigma_{K,2}) V
\end{array} \right]
\mu I_{n \times p} + \tau \Psi \circ P(A) + \sigma_{K,1} \frac{VV^T A}{A} + \sigma_{K,2} I_{n \times p}
$$

which is exactly the described update rule in Section 5.

**Appendix B  Kullback-Leibler Divergence Discrepancy and LQBP**

In this Section, we will use the LQBP construction principle to derive a multiplicative algorithm for the cost function

$$
F(X) := \text{KL}(Y, KX) = \sum_{j=1}^{m} \text{KL}(Y_{*,j}, KX_{*,j}) := \sum_{j=1}^{m} f_{Y_{*,j}}(X_{*,j})
$$

Similar to the approach in Appendix A.2, we define according to the LQBP principle the surrogate

$$
Q_F(X, A) = \sum_{j=1}^{m} f_{Y_{*,j}}(A_{*,j}) + \nabla f_{Y_{*,j}}(A_{*,j})^T (X_{*,j} - A_{*,j})
+ \frac{1}{2} (X_{*,j} - A_{*,j})^T A_{f_{Y_{*,j}}}(A_{*,j})(X_{*,j} - A_{*,j})
$$

with the diagonal matrix

$$
A_{f_{Y_{*,j}}}(A_{*,j})_{kk} = \frac{\nabla^2 f_{Y_{*,j}}(A_{*,j}) A_{*,j})}{A_{*,j}} + \kappa_k.
$$

It follows for the partial derivatives of $f$

$$
\frac{\partial f_{Y_{*,j}}}{\partial X_{\beta \zeta}}(X_{*,j}) = -\sum_{i=1}^{n} Y_{i \zeta} K_{i \beta} + K_{i \beta},
$$

$$
\frac{\partial^2 f_{Y_{*,j}}}{\partial X_{\alpha \zeta} \partial X_{\beta \zeta}}(X_{*,j}) = \sum_{i=1}^{n} Y_{i \zeta} K_{i \alpha} K_{i \beta} (KX)^2_{i \zeta}.
$$
The first order optimality condition of the surrogate functional leads then to
\[
0 = \frac{\partial Q_F}{\partial X_{\kappa \xi}}(X, A) = - \sum_{i=1}^{n} Y_i \xi K_{i \xi} \left( (KA)_{i \xi} \right) + \sum_{i=1}^{n} K_{i \xi}
\]
\[
+ \frac{\sum_{i=1}^{n} Y_i \xi K_{i \xi} \left( (KA)_{i \xi} \right) + \kappa \xi}{A_{\kappa \xi}} (X_{\kappa \xi} - A_{\kappa \xi}).
\]
Setting \( \kappa \xi := \sum_{i=1}^{n} K_{i \xi} \) and solving for \( X_{\kappa \xi} \) leads finally to the multiplicative update rule
\[
X_{[d+1]} = \frac{2X_{[d]}^T Y}{K_{[d+1]}^T X_{[d]}^T} + K_{[d+1]}^T I_{n \times m} \frac{Y}{K_{[d+1]}^T X_{[d]}^T},
\]
which differs from the classical update rule for the KLD described in [11,24,25].

Appendix C Surrogate of the TV Penalty - Separability

In this section, we will prove the separability of the surrogate functional
\[
Q_{TV}(K, A) = p \sum_{k=1}^{p} \psi_k \sum_{i=1}^{n} \frac{1}{|\nabla_i K|} \left( \varepsilon_{TV}^2 + \sum_{\ell \in N_i} K_{ik}^2 + K_{k\ell}^2 - K_{ik}(A_{ik} + A_{k\ell}) - K_{k\ell}(A_{ik} + A_{k\ell}) + A_{ik}^2 + A_{k\ell}^2 \right),
\]
(57)
described in Theorem 5. Furthermore, we choose an arbitrary \( s \in \{1, \ldots, n\} \) and \( t \in \{1, \ldots, p\} \). The aim is now to find all terms in (57) with \( K_{st}^2 \) and \( K_{st} \).

To find all quadratic terms \( K_{st}^2 \) in (57), we see that we have to fix the index \( k \), such that \( k = t \). The remaining indices in (57), which have to be analyzed, are \( i \) and \( \ell \).

For the case \( i = s \), we find that the preceding coefficient is
\[
\psi_t \sum_{r \in N_s} \frac{1}{|\nabla_i K|}.
\]
The case \( \ell = s \) can only occur for those indices \( i \), which satisfy \( s \in N_i \). The definition of the adjoint neighbourhood pixels gives
\[
\forall i: s \in N_i \iff \forall i: i \in \bar{N}_s.
\]
Therefore, the corresponding preceding coefficient is here
\[
\psi_t \sum_{r \in \bar{N}_s} \frac{1}{|\nabla_{rt} K|}.
\]
Altogether, we obtain for the quadratic terms \( K_{st}^2 \) the coefficient
\[
\tilde{P}_{st}(A) := \psi_t \left( \frac{1}{|\nabla_{st} A|} \sum_{r \in N_s} \frac{1}{|\nabla_r A|} + \sum_{r \in \bar{N}_s} \frac{1}{|\nabla_{rt} A|} \right).
\]
such that $\tilde{P}_{st}(A) \cdot K_{st}^2$ takes all quadratic terms of the matrix entries of $K$ in the surrogate functional into account.

The same can be done with the linear terms $K_{st}$, which leads to the coefficient

$$\tilde{Z}_{st}(A) := -\psi_2 \left( \frac{1}{|\nabla_{st} A|} \sum_{r \in N_s} [A_{st} + A_{rt}] + \sum_{r \in \bar{N}_s} \frac{A_{st} + A_{rt}}{|\nabla_{rt} A|} \right).$$

Therefore, the surrogate $Q_{TV}$ can be written as

$$Q_{TV}(K, A) = \sum_{t=1}^{p} \sum_{s=1}^{n} \left[ \tilde{P}_{st}(A) \cdot K_{st}^2 + \tilde{Z}_{st}(A) \cdot K_{st} \right] + \tilde{C}(A)$$

for some function $\tilde{C}$, which only depends on $A$. This shows the separability of the surrogate.