Minimax estimation in linear models with unknown finite alphabet design

Merle Behr\textsuperscript{*‡} and Axel Munk\textsuperscript{*†}

Abstract. We provide minimax theory for joint estimation of $F$ and $\omega$ in linear models $Y = F\omega + Z$ where the parameter matrix $\omega$ and the design matrix $F$ are unknown but the latter takes values in a known finite set. We show that this allows to separate $F$ and $\omega$ uniquely under weak identifiability conditions, a task which is not doable, in general. These assumptions are justified in a variety of applications, ranging from signal processing to cancer genetics. We then obtain in the noiseless case, that is, $Z = 0$, stable recovery of $F$ and $\omega$ in a neighborhood of $Y$. Based on this, we show for Gaussian error matrix $Z$ that the LSE attains minimax rates for both, prediction error of $F\omega$ and estimation error of $F$ and $\omega$, separately. Due to the finite alphabet, estimation of $F$ amounts to a classification problem, where we show that the classification error $P(\hat{F} \neq F)$ decreases exponentially in the dimension of one component of $Y$.

Keywords: Finite alphabet, Combinatorial linear model, Least squares estimator, Minimax estimation, Exact recovery, Blind source separation, Measurement uncertainty.

MSC 2010 subject classifications: Primary 62F12, 62H30, Secondary 62F30, 62J05.

1 Introduction

Linear models (LM) are arguably one of the most prominent tools in statistical modeling. In such a (multivariate) LM one observes a matrix $Y \in \mathbb{R}^{n \times M}$,

$$Y = F\omega + Z,$$

which is linked to the parameter matrix of interest $\omega \in \mathbb{R}^{m \times M}$ via a design matrix $F \in \mathbb{R}^{n \times m}$ and an additive noise matrix $Z \in \mathbb{R}^{n \times M}$, which in this work is assumed to be i.i.d. Gaussian $Z_{ij} \sim N(0, \sigma^2)$ for $i = 1, \ldots, n, j = 1, \ldots, M$, with (unknown) variance $\sigma^2 > 0$. Usually, $F$ is assumed to be known and analysis on $\omega$ is performed conditioned on $F$, as, e.g., in classical ANOVA or in regression analysis where $F$ is determined by the design of the experiment. In contrast, in the following, we want to consider the situation where the matrix $F$ is unknown and has to be estimated from the data $Y$ jointly with the parameter matrix $\omega$. In general, separation of

\textsuperscript{*}Department of Statistics, University of California Berkeley, 367 Evans Hall, Berkeley, CA 94720, \textsuperscript{*}Institute for Mathematical Stochastics, University of Göttingen, Goldschmidtstraße 7, 37077 Göttingen, \textsuperscript{†}Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Göttingen, Germany, Email: behr@berkeley.edu, munk@math.uni-goettingen.de ; \textsuperscript{‡}corresponding author
\( F \) and \( \omega \) from \( F\omega \) is not possible, of course, and therefore existing approaches (see, e.g., \cite{25}) focus on estimation of \( F\omega \). However, if we assume that \( F \) can only attain values in a known, finite set \( \mathfrak{A} = \{a_1, \ldots, a_k\} \subset \mathbb{R} \), denoted as finite alphabet, we will show that the LM in \( (1) \) becomes identifiable, that is, \( F \) and \( \omega \) can be separated from \( F\omega \), under rather weak assumptions on \( \omega \) and \( F \). The aim of this paper is to provide estimates for joint recovery of \( F \) and \( \omega \) and to develop statistical theory for these quantities. For better understanding, it is convenient to rewrite the LM \( (1) \) with unknown design matrix \( F \) as a blind source separation problem (the terminology is borrowed from the signal processing literature)

\[
Y_l = \sum_{i=1}^{m} F_i \omega_m + Z_l, \quad l = 1, \ldots, M,
\]

with \( m \) source signals \( F_1, \ldots, F_m \in \mathfrak{A}^n \), each only taking values in the finite alphabet \( \mathfrak{A} \), which are obtained in \( M \) mixtures with unknown mixing weights \( \omega_1, \ldots, \omega_M \in \mathbb{R}^m \). Therefore, we will denote model \( (1) \) with \( F \in \mathfrak{A}^{n \times m} \) unknown as the Multivariate finite Alphabet Blind Separation (MABS) model.

**Notation** Throughout the following bold letters, e.g., \( F, \omega \), denote the underlying truth generating the observations \( Y \) in \( (1) \). Further, throughout the following \( \|A\| \) denotes the Frobenius norm for a matrix \( A \) and \( \|x\| \) denotes the Euclidean norm for vectors \( x \). For a matrix \( A \) we denote by \( A_{i} \cdot \) its \( i \)th row and by \( A \cdot i \) its \( i \)th column. For a vector \( x \) we denote by \( x_i \) its \( i \)th entry.

**Applications** MABS occurs in many different fields. For instance, in digital communications \cite{26,30,34,36,28}, where \( m \) digital signals (e.g., binary signal with \( \mathfrak{A} = \{0,1\} \)) are modulated (e.g., with pulse amplitude modulation), transmitted through wireless channels (each having different channel response), and received by \( M \) antennas. In signal processing this is known as MIMO (multiple input multiple output) and (ignoring time shifts, i.e., considering instantaneous mixtures) can be described by MABS when the channel response is unknown, see e.g., \cite{30,20} for details.

Another example where MABS is relevant arises in cancer genetics \cite{35,7,19,14}. Specific mutations in cancer tumors are copy number variations (CNV’s) where some parts of the genome are either duplicated or deleted. CN’s of a single tumor only take integer values, i.e., \( \mathfrak{A} = \{0,1,2,\ldots,k\} \) (with good biological knowledge of a maximum copy number \( k \), see e.g., \cite{19}). However, tumors are known to be heterogeneous, i.e., they consists of a few different types of tumor cells, so called clones, see e.g., \cite{29,13}. In whole genome sequencing (WGS) data the CN’s of the single clones overlap according to the relative (unknown) proportion of the clone in the tumor. Important for the analysis of this data is that often \( M \) different probes of the tumor cells are available, taken at different time points or at different locations. Each of these contain the same clones but at different relative proportions. This can be modeled with MABS, where \( m \) is the number of clones, \( M \) is the number of probes, \( \omega_{ij} \) corresponds to the unknown proportion of clone \( i \) in probe \( j \) of the tumor, and \( F_i \) corresponds to the CN’s of
clone $i$, see e.g., \cite{5,13,29} for details. For $M = 1$, see \cite{3}. Analog, one can model point mutations in tumors with MABS with alphabet $\mathcal{A} = \{0, 1\}$, where $F_{ji} = 1$ if and only if point mutation $j$ is present in clone $i$ \cite{15}.

Two simplifications Motivated from the application in cancer genetics, where the mixing weights correspond to physical mixing proportions of DNA strands, we will in the following assume that the mixing weights $\omega_{ij}$ are positive and sum up to one for each $j$. This assumption simplifies the corresponding identifiability conditions to decompose $F$ and $\omega$ uniquely. However, we stress that all results can be extended for general mixing weights which only requires a slight modification of the corresponding identifiability assumptions, see Remark 2.4. The minimax rates which are derived in the following do not depend on this assumption. More precisely, for a given number of sources $m$ and a given number of mixtures $M$, the set of possible mixing weights $\omega$ is defined as

$$\Omega_{m,M} := \left\{ \omega \in \mathbb{R}_{+}^{m \times M} : 0 < \|\omega_1\| < \ldots < \|\omega_m\|, \sum_{i=1}^{m} \omega_{ij} = 1 \forall j = 1, \ldots, M \right\}. \quad (3)$$

Note that a fixed ordering of the row-sums is necessary as otherwise for any permutation matrix $P$ one finds that $F\omega = FP^{-1}P\omega$ with $\omega$ and $P^{-1}\omega$ both valid mixing weights. Moreover, throughout the following, we may assume w.l.o.g. that the fixed given alphabet $\mathcal{A}$ is ordered and that $a_1 = 0$ and $a_2 = 1$, that is

$$\mathcal{A} = \{0, 1, a_3, \ldots, a_k\} \quad \text{with} \quad 1 < a_3 < \ldots < a_k. \quad (4)$$

Otherwise, one may instead consider the observations $(Y_{ij} - a_1)/(a_2 - a_1)$ with alphabet $\mathcal{A} = \{0, 1, \frac{a_3-a_1}{a_2-a_1}, \ldots, \frac{a_k-a_1}{a_2-a_1}\}$ in \cite{1}.

Identifiability A minimal requirement underlying any recovery algorithm of $F$ and $\omega$ from (a possibly noisy version of) $G := F\omega$ in \cite{1} to be valid is identifiability, that is, a unique decomposition of the mixture $G$ into finite alphabet sources $F$ and weights $\omega$.

For illustration, consider a binary alphabet $\mathcal{A} = \{0, 1\}$ with two sources $m = 2$ and a single mixture $M = 1$. The question is as follows: When is it possible to uniquely recover the underlying weights $\omega \in \mathbb{R}_+^2$ and sources $F \in \mathbb{R}^{n \times 2}$ from the mixture $G = F\omega \in \mathbb{R}^n$? In this example the answer is simple: As the entries in $F$ can only attain the values 0 and 1 the smallest possible value for $G_j$, $j = 1, \ldots, n$, is 0, which corresponds to both sources taking the smallest alphabet value $F_{j1} = F_{j2} = 0$. Analog, when $F_{j1} = 0$ and $F_{j2} = 1$, $G_j$ takes the second smallest possible value, denoted as $\omega_1$ (recall that $0 < \omega_1 \leq \omega_2$ and $\omega_1 + \omega_2 = 1$ by \cite{3}). Similar, the third smallest value for $G_j$ then equals $\omega_2$ with $F_{j1} = 0, F_{j2} = 1$ and the largest value equals 1 with $F_{j1} = F_{j2} = 1$. Thus, one can (almost) always uniquely identify $\omega$ and $F$ from $G = F\omega$ simply by looking at the ordering structure of the values $G_1, \ldots, G_n$. There are just two situations where this fails:
1. If $\omega_1 = \omega_2$, one cannot identify from $G$ whether $F_{j1} = 0, F_{j2} = 1$ or $F_{j1} = 1, F_{j2} = 0$.

2. If $F_{j1} = F_{j2}$, one cannot identify from $G$ the values $\omega_1, \omega_2$.

Consequently, in order to guarantee identifiability (in this simple example), we have to exclude these two situations. That is, we need to exclude from the parameter space the single weight vector $\omega = (0.5, 0.5)$ (the only one with $\omega_1 = \omega_2$) and sources $F = (F_1, F_2)$ with equal components $F_1 = F_2$ (or equivalently $\omega_1, \omega_2 \not\in \{G_1, \ldots, G_n\}$). Clearly, this is not very restrictive in most situations, as it simply excludes that only one source is visible.

Now we turn to the general case, of arbitrary $\mathfrak{A}, m,$ and $M$. It is shown in [4] that identifiability has a complete combinatorial characterization via the given alphabet and that the above assumptions can be extended to a universal (for any $\mathfrak{A}, m, M$) simple sufficient condition, called separability, which guarantees identifiability. For sake of completeness and to ease reading, we recapitulate these conditions in the following in our context.

First, we discuss conditions on $\omega$. For fixed $\omega$ each row of $G = F\omega$ can take any of at most $k^m$ (recall that the alphabet $\mathfrak{A}$ has size $k$) values of the form $e\omega = \sum_{i=1}^{m} e_i \omega_i$ with $e = (e_1, \ldots, e_m) \in \mathfrak{A}^m$ (elements in $\mathfrak{A}^m$ are considered as row vectors). Clearly, if for any two $e \neq e' \in \mathfrak{A}^m$ it holds that $e\omega = e'\omega$, then $F$ is not identifiable from $G$, in general, as for any row $i \in \{1, \ldots, n\}$ with $G_i = e\omega$ it cannot be distinguished whether $F_i = e$ or $F_i = e'$. Hence, we require that the alphabet separation boundary (ASB) [4], that is, the minimal distance between any of these values, is positive, i.e.,

$$\text{ASB}(\omega) = \text{ASB}(\omega, \mathfrak{A}) := \min_{e \neq e' \in \mathfrak{A}^m} \frac{1}{\sqrt{M}} \|e\omega - e'\omega\| > 0. \quad (5)$$

Recall that $e\omega \in \mathbb{R}^M$ and $\|\cdot\|$ denotes the Euclidean norm, hence, $1/\sqrt{M}$ is the appropriate scaling factor. Further, we will see that when $e\omega$ is corrupted by noise as in the MABS model [1], the ASB quantifies perturbation stability of $G$ when any two of these values are close, that is, $\|e\omega - e'\omega\| < \delta \sqrt{M}$ for small $\delta > 0$.

Recall that for $\Omega_{m,M}$ in [3], to ensure identifiability one has to assume $\|\omega_1\| < \ldots < \|\omega_m\|$, as otherwise the ordering of the rows of $\omega$ (and columns of $F$, respectively) are not well defined via the mixture $F\omega$. Hence, in a noisy setting, as in the MABS model, the minimal distance between the row norms of $\omega$ will be crucial for separation and hence recovery of $F$ and $\omega$. Therefore, for $\omega \in \Omega_{m,M}$ we introduce the weights separation boundary

$$\text{WSB}(\omega) = \text{WSB}(\omega, \mathfrak{A}) := \frac{1 + ma_k}{2\sqrt{M}} \min_{i = 2, \ldots, m} (\|\omega_i\| - \|\omega_{i-1}\|). \quad (6)$$

Again, $\omega_i \in \mathbb{R}^M$, which results in the scaling factor $1/\sqrt{M}$. The additional factor $(1 - ma_k)/2$ just ensures that the ASB in (5) and the WSB in (6) have comparable scaling properties in terms of $\mathfrak{A}$ and $m$ (see proof of Theorem 2.2). Figure 1 illustrates for $m = M = 2$ in [3] the set

$$\Omega_{2,2} = \left\{ \left( \begin{array}{cc} a & b \\ 1-a & 1-b \end{array} \right) : a, b \in [0,1], a + b < 1 \right\}. \quad (7)$$
For $\mathcal{A} = \{0, 1\}$, regions where $\min(ASB(\omega), WSB(\omega)) \geq \delta$ are displayed blue in Figure 1.1 for different values of $\delta$. In particular, this illustrates that the set of regularized weights $\{\omega \in \Omega_{2, 2} : \min(ASB(\omega), WSB(\omega)) \geq \delta\}$ is reasonably large when $\delta$ is sufficiently small. For illustration, recall the simple example from above with $\mathcal{A} = \{0, 1\}, m = 2, M = 1$, there the condition $ASB(\omega), WSB(\omega) > 0$ is equivalent to $\omega_1 \neq \omega_2$.

Second, we discuss conditions on $F$. In order to identify $\omega$ from $G$ it is necessary that the sources $F_1, \ldots, F_m$ differ sufficiently much. For instance, if $F_1 = \ldots = F_m$ then $G = F_1$ irrespective of $\omega$. Here, we employ the separability condition from [4], which provides a sufficient variability of $F$ and guarantees identifiability. More precisely, separability guarantees that for each $i = 1, \ldots, m$ there exists some $j = 1, \ldots, n$ where $F_{ij}$ takes the second smallest alphabet value and all other sources $F_{ki}, k \neq j$, take the smallest alphabet value. As the alphabet is of the form (4), this is equivalent to $\omega_1, \ldots, \omega_m, F_1, \ldots, F_m \in \{G_1, \ldots, G_n\}$.

In summary, we denote a pair $(\omega, F)$ in MABS as $\delta$-separable if $\delta = 0.007, 0.07, 0.14, 0.28$ (from left to right).

For $\mathcal{A} = \{0, 1\}$, regions where $\min(ASB(\omega), WSB(\omega)) \geq \delta$ are displayed blue in Figure 1.1 for different values of $\delta$. In particular, this illustrates that the set of regularized weights $\{\omega \in \Omega_{m, M} : \min(ASB(\omega), WSB(\omega)) \geq \delta\}$ is reasonably large when $\delta$ is sufficiently small. For illustration, recall the simple example from above with $\mathcal{A} = \{0, 1\}, m = 2, M = 1$, there the condition $ASB(\omega), WSB(\omega) > 0$ is equivalent to $\omega_1 \neq \omega_2$.

Second, we discuss conditions on $F$. In order to identify $\omega$ from $G$ it is necessary that the sources $F_1, \ldots, F_m$ differ sufficiently much. For instance, if $F_1 = \ldots = F_m$ then $G = F_1$ irrespective of $\omega$. Here, we employ the separability condition from [4], which provides a sufficient variability of $F$ and guarantees identifiability. More precisely, separability guarantees that for each $i = 1, \ldots, m$ there exists some $j = 1, \ldots, n$ where $F_{ij}$ takes the second smallest alphabet value and all other sources $F_{ki}, k \neq j$, take the smallest alphabet value. As the alphabet is of the form (4), this is equivalent to $\omega_1, \ldots, \omega_m, F_1, \ldots, F_m \in \{G_1, \ldots, G_n\}$.
\((\omega, F), (\omega, F)\) both \(\delta\)-separable
\[
\max_{j=1,\ldots,n} \| (F\omega)_j - (F\omega)_j\| < c(\delta)
\]
\[
\Rightarrow F = F\text{ and } \max_{i=1,\ldots,m} \| \omega_i - \omega_i\| < c(\delta),
\]
where \(c(\delta) \to 0\) as \(\delta \to 0\). Stable recovery in \((9)\) will provide the basis to extend minimax rates for prediction error to minimax rates for estimation error. Estimation of \(\Pi\) and \(\omega\) separately is of primary importance in many applications and the finite alphabet provides the basis for this.

**Main results** In Section 3 and 4 we derive estimators for the unknown quantities \(F\) and \(\omega\), which are asymptotically minimax optimal (up to constants) for prediction and estimation, where we assume the number of sources \(m\) and the alphabet \(\mathcal{A}\) to be fixed and known. The aim is to study the influence of all quantities on recovery as \(n, M \to \infty\), where the total sample size is \(nM\). Our minimax rates apply to all regimes where
\[
n/M \to \infty \quad \text{and} \quad \log(n)/M \to 0,
\]
that is, where the number of observations grows faster than the number of mixtures and the number of mixtures grows at least logarithmic with the number of observations. This describes a very realistic regime of sample sizes in many applications, where it is usually much easier to collect a large number of source samples \(n\) (e.g., genome locations in cancer genetics), than a large number of mixture samples \(M\) (e.g., number of tissue probes of a cancer patient at different locations or time points). See Remark 3.6 for more details.

It is intuitive that increasing \(M\) makes inference on \(F\) in \((1)\) easier (we observe more mixtures of the same sources) while inference on \(\omega\) becomes more difficult (the dimension of \(\omega\) and thus the number of parameters to be estimated increases with \(M\)). In contrast, increasing \(n\) makes inference on \(F\) more difficult (the dimension of \(F\) and thus the number of parameters to be estimated increases with \(n\)) while inference on \(\omega\) becomes easier (we observe more samples from the same mixture matrix \(\omega\)). This is quantified in Theorem 3.1 (lower bound) and Theorem 3.5 (upper bound) which provide under weak identifiability conditions the minimax rate for the prediction error as
\[
\frac{\sigma^2 m}{n} + \frac{\sigma e^{-c\frac{\sigma^2}{M^2}}}{\sqrt{M}} \lesssim \inf_{\hat{\theta}} \sup_{F,\omega} \mathbb{E}_{F,\omega} \left( \left\| \hat{\theta} - F\omega \right\|^2 / nM \right) \lesssim \frac{\sigma^2 m}{n} + \frac{\sigma e^{-c'\frac{\sigma^2}{M^2}}}{\sqrt{M}},
\]
whenever \((10)\) holds. Further, we show that the least squares estimator (LSE) achieves this rate. Here \(\lesssim\) and \(\gtrsim\) denote inequalities up to a universal constant which does not depend on any model parameter and \(c = c(m, \mathcal{A}), c' = c'(m, \mathcal{A}) > 0\) are positive constants. A major consequence of \((11)\) is that when \(M \gg \log(n)\) it does not play much of a role for the prediction error in model \((1)\) whether the design matrix \(F\) is known or unknown. In particular, for \(M \gg \log(n)\) the precise form of the alphabet \(\mathcal{A}\) does not influence the prediction rate. The alphabet only
enters in the second term, that is, it determines the constant of the $O(\ln(n))$ mixtures needed to remove the influence of the unknown design $F$ in (1). This dependence on $\mathfrak{A}$ does not coincide for our lower and upper bound and it seems natural that the precise constant is not a simple feature of $\mathfrak{A}$.

The exact recovery result in (9) puts the basis to relate the prediction error $\|\hat{\theta} - F\omega\|$ in (11) to the estimation error via the metric

$$d((F, \omega), (\hat{F}, \hat{\omega})) = \sqrt{M} 1_{F \neq \hat{F}} + \max_{i=1, \ldots, m} \|\omega_i - \hat{\omega}_i\|$$

(12)

(see Theorem 4.2), where the scaling factor $\sqrt{M}$ naturally arises from the dimensionality of $\|\omega_i - \hat{\omega}_i\|$. In Theorem 4.3 we show that

$$c_1 \frac{\sigma^2}{M} + \frac{\sigma e^{-c \frac{M}{\sigma^2}}}{\sqrt{M}} \lesssim \inf_{F, \omega, \hat{F}, \hat{\omega}} \sup_{(F, \omega), (\hat{F}, \hat{\omega})} \left( \frac{d((F, \omega), (\hat{F}, \hat{\omega}))^2}{M} \right)$$

(13)

whenever (10) holds, with $c_1 = c_1(m, \mathfrak{A}), c_1' = c_1'(m, \mathfrak{A}) > 0$ positive constants and this rate is achieved by the LSE. A major consequence of (13) is that also for the estimation error if $M \gg \ln(n)$ the unknown $F$ in (1) does not play much of a role. Moreover, for the LSE we provide an explicit exponential bound on the classification error (see Corollary 4.1)

$$P_{F, \omega}(\hat{F} \neq F) \lesssim \frac{\sigma e^{-c \frac{M}{\sigma^2}}}{\sqrt{M}}.$$  

This shows that as $M$ increases the unknown design is estimated exactly with probability increasing exponentially in $M$. Thus, for $M$ large enough, $\hat{F} = F$ finally, and thus, the LSE for joint estimation of $\omega, F$ coincides with the ordinary LSE for $\omega$ (and given $F$).

**Related work**

Identifiability for finite alphabet sources has been considered e.g., in [30, 9]. Here, we extend the identifiability result of [4] to a stable and exact recovery result for an arbitrary mixture dimension $M \in \mathbb{N}$ (see Theorem 2.2). The univariate case $M = 1$ was considered in [3], where the temporal structure of a change-point regression setting allows recovery of sources and mixing weights. Here, we treat mixtures with arbitrary dimension $M \in \mathbb{N}$ and the validity of our methodology now results from observing $M$ (possibly different) mixtures of the same sources. In particular, we cannot rely on any “temporal” structure as in the univariate case.

Mostly related to our work is [25], who also considered model (1), but with $F$ being an arbitrary design matrix (without finite alphabet constraint) which is unknown only up to a permutation matrix. They derive minimax prediction rates, that is, for estimation of $G = F\omega$, and show that the LSE for $G$ obtains these rates (up to log-factors). They also consider the case where $F$ is unknown up to
a selection matrix (i.e. not every row of the design necessarily appears in the data $Y$ and some rows might be selected several times). One can rewrite the MABS model in an analog way, to obtain a model as in (1) where the design matrix equals $F = \Pi A$, with $\Pi$ an unknown selection matrix and $A$ being the matrix where the rows constitute of all different combinations of alphabet values (see equations (15) - (17) at beginning of Section 2 for details). [25] consider general $A$ and derive minimax prediction rates of the form

$$\inf_{\hat{\theta}} \sup_{\Pi A \omega} E_{\Pi A \omega} \left( \frac{1}{nM} \| \hat{\theta} - \Pi A \omega \|^2 \right) \approx \frac{\sigma^2 m}{n} + \frac{\sigma^2 (\ln(n))}{M}, \quad (14)$$

where the log-term only appears in their upper bound. In our situation, where we assume a specific finite alphabet for the design matrix, thus a specific matrix $A$, the second term in the minimax rate becomes exponential in $M$ instead of parametric. The rate (14) is obtained in [25] by treating the whole matrix $\Pi A$ as unknown. In this paper, we exploit a specific structure of $A$ and thus obtain a faster rate. Note that, just as in our setting (see (11)), [25] obtain with (14) that whenever $\ln(n) \gg M$ the unknown permutation $\Pi$ does not play much of a role for the prediction error. The major difference, however, is that under the finite alphabet we can now provide identifiability conditions on $F = \Pi A$ and $\omega$ in (1) and thus, in contrast to [25], we do obtain estimators for $\omega$ and $\Pi$ and bounds for the estimation error. By regularizing the model in an appropriate way, we obtain the minimax estimation rate (for $n, M \to \infty$) for $F$ and $\omega$ up to constants and show that it is achieved by the LSE.

Note that model (1) with either of both, $F$ or $\omega$, known corresponds to standard models in statistics: If $F$ is known, (1) is a classical multivariate LM. If $\omega$ is known, the finite alphabet assumption turns model (1) into a clustering problem with known centers $\{a \omega : a \in A^m\}$. In contrast, in MABS both, $F$ and $\omega$, are unknown. Hence, MABS can be seen as a hybrid model of clustering and parameter estimation, but has received rather few attention so far. Only some specific instances of MABS, e.g., specific alphabets, have been considered previously, see, for example, [30, 23, 9, 33]. However, all these works only focus on algorithms (often in the noiseless case), but do not provide any theory in a statistical context.

A similar problem as in [25], which we discussed above, was considered in [24]. However, they assume $M = 1$ (not general $M \in \mathbb{N}$ as in this paper) and unknown permutation matrices $\Pi$, which is more restrictive than general selection matrices, as considered here. Moreover, they assumed a random design $A$ with Gaussian entries, in contrast to MABS where $A$ is a fixed finite alphabet matrix. This makes the analysis severely different, as the finite alphabet assumption allows to separate $\Pi$ and $\omega$ from their mixture $\Pi A \omega$. In their setting, they give a sharp condition on the signal to noise ratio $\|\omega\|/\sigma$ and the number of observations $n$ under which it is possible to exactly recover the permutation $\Pi$ with large probability. They show that the LSE recovers $\Pi$ with large probability whenever this is possible. [31] study a similar model as [24] but mainly focus on the noiseless case. They also consider a random design for $A$ (in contrast to MABS). They focus on recovery of $\omega$ (not on $\Pi$) and show that whenever $n > 2m$ with probability one $\omega$ can be recovered from the (noiseless) observations $Y$. [22] consider a similar model
as in the context of object recognition, where \( m = 3 \) and \( M = 2 \). There \( m = 3 \) corresponds to the dimension of an object, \( M = 2 \) to the dimension of a photo of this object, and the unknown mixture matrix \( \omega \) to an unknown camera perspective. They also focus on recovery of the unknown permutation \( \Pi \). Their results basically require that sufficiently many of the \( n \) permutations are known in advance, which is a rather strong assumption.

A structural similarity to MABS appears in nonnegative matrix factorization (NMF), where one assumes \( \text{with } F \) and \( \omega \) both non-negative \([17, 11, 1]\). Here, however, we do have the additional assumption of a finite alphabet for the sources \( F \), which leads to a model structure more related to a classification problem. Hence, estimation rates are expected to be in a completely different regime, although, we stress that we are not aware of any minimax results for prediction error in NMF, as derived here for MABS. See, however, recent results by \([16, 6]\) who study NMF in the context of topic models and provide minimax rates for L1-error of estimating the word-topic matrix, which corresponds to our matrix \( F \). Note that in our setting, due to the finite alphabet, \( F \) is estimated exactly eventually and thus, we cannot directly compare their results to ours. Further, it should be stressed that from a computational perspective both models are very different, as NMF algorithms usually do not allow to incorporate a finite alphabet assumption. An exception is binary matrix factorization \([13]\), where both, \( F \) and \( \omega \) are assumed to have binary entries. There, however, also the data matrix \( Y \) will be binary, which makes both, theory as well as computations very different.

The separability conditions \( \delta^{-IC1} \) and \( IC2 \), which we introduce here (see Section 2), are closely related to standard conditions for NMF \([11, 1]\), from where the notation separable originates. The proofs for separability are, however, different. Whereas in MABS they are build on combinatorics, in NMF they are build on geometric considerations. More details on the relation between separability in MABS, as it is employed here, and separability in NMF can be found, for example, in \([4, 3]\).

Also related is statistical seriation \([12]\). There \( F = \Pi \) in \([1]\) is itself an \( n \times n \) permutation matrix and the \( n \times M \) matrix \( \omega \) is assumed to have unimodal columns. Just as in this paper for MABS, \([12]\) obtain that the LSE is (almost) minimax optimal for the statistical seriation problem. Similar as in \([11]\) and \([14]\), the respective minimax prediction rate is the sum of two terms: The first term corresponds to the rate which would be achieved when the permutation matrix was known. The second term corresponds to the price one has to pay for the combinatorial uncertainty in form of the unknown permutation matrix. Assuming that the variation of \( \omega \) is not too small, it is of the same form as in \([14]\), that is, it vanishes with rate \( 1/M \), instead of exponentially as in this work due to the finite alphabet.

Finally, note that the MABS model \([1]\) can be seen as a particular type of dictionary learning (see e.g., \([21, 27]\) for a review), where the dictionary constitutes of all vectors with elements in the finite alphabet \( \mathcal{A} \). We are not aware of any other work which provides statistical theory for finite alphabet dictionaries.
Organization of the paper  In Section 2 we introduce the MABS model and corresponding identifiability conditions. From this we derive stable recovery under suitable regularization, see Theorem 2.2. In Section 3 we derive lower bounds (Theorem 3.1) and upper bounds (Theorem 3.5) for the minimax prediction rate. In Section 4 we give an upper bound for the classification error (Corollary 4.1) and, based on this, lower and upper bounds for the minimax estimation rate (Theorem 4.3). We conclude in Section 5. All proofs are postponed to the appendix.

2 Model assumptions and identifiability

It is illustrative to rewrite the MABS model to highlight its combinatorial structure. To this end, we rewrite the unknown finite alphabet design matrix \( F \in \mathbb{A}^{n \times m} \) as a product of an unknown selection matrix \( \Pi \) and the known design matrix \( A \) with rows consisting of all different alphabet combinations in \( \mathbb{A}^{m} \). Then the MABS model (1) is equivalent to

\[
Y = \Pi A \omega + Z,
\]

with an unknown selection matrix

\[
\Pi \in \{0, 1\}^{n \times k^m}, \quad \sum_{j=1}^{n} \Pi_{ij} = 1 \quad \forall i = 1, \ldots, n,
\]

and known finite alphabet design matrix

\[
A := \begin{pmatrix}
0 & 0 & 0 & \ldots & 0 & 0 & 0 \\
0 & 0 & 0 & \ldots & 0 & 0 & 1 \\
0 & 0 & 0 & \ldots & 0 & a_3 \\
\vdots \\
0 & 0 & 0 & \ldots & 0 & a_k \\
0 & 0 & 0 & \ldots & 0 & 1 & 0 \\
\vdots \\
a_k & a_k & a_k & \ldots & a_k & a_k & a_k
\end{pmatrix} \in \{0, a_3, \ldots, a_k\}^{k^m \times m},
\]

where the rows of \( A \) constitute all different vectors in \( \mathbb{A}^{m} \) (recall (4)). Further, the unknown mixing matrix \( \omega \in \Omega_{m,M} \) is as before in (1) and we assume i.i.d. normal noise \( Z_{ij} \sim \mathcal{N}(0, \sigma^2) \), \( i = 1, \ldots, n \), \( j = 1, \ldots, M \), with unknown variance \( \sigma^2 \). As discussed in the introduction, we employ \( \delta \)-separability on \( \Pi \) and \( \omega \) to guarantee identifiability and stable recovery. That is, for \( \omega \in \Omega_{m,M} \) and \( \delta > 0 \), we assume the identifiability condition (IC)

\[
\min (ASB(\omega), WSB(\omega)) \geq \delta.
\]

(\( \delta \)-IC 1)

The separability condition for \( F \) in (8) now translates to \( \Pi \) in (16) as follows. \( \Pi \) is separable if and only if \( F = \Pi A \) is separable, namely,

\[
\exists j_1, \ldots, j_m \in \{1, \ldots, n\} \text{ such that } ((\Pi A)_{rs})_{r=j_1, \ldots, j_m, s=1, \ldots, m} = I_{m \times m},
\]

(\( \text{IC 2} \))

that is, if \( \Pi \) selects at least once each of the unit vectors \( e^1, \ldots, e^m \).
Theorem 2.1 ([4], Theorem 4.1).

Consider the MABS model (15). Then, for any \( \delta > 0 \) under (\( \delta \)-IC 1) and (IC 2), \( (\omega, \Pi) \) is identifiable:

If \( (\omega, \Pi), (\omega', \Pi) \) both satisfy (\( \delta \)-IC 1) and (IC 2) for some \( \delta > 0 \), then \( \Pi A \omega = \Pi A \omega' \) implies \( \omega = \omega' \) and \( \Pi = \Pi' \).

The following theorem shows (for a proof see Appendix A.1) how the parameter \( \delta \) in (\( \delta \)-IC 1) regularizes the space of possible mixtures \( \Pi A \omega \) in such a way that recovery is still valid within perturbations of \( G = \Pi A \omega \).

Theorem 2.2 (Stable and exact recovery).

Consider the MABS model (15). Fix \( \delta > 0 \) and let \( 0 < \epsilon < \sqrt{M\delta}/(1 + ma_k) \). Assume that (\( \delta \)-IC 1) and (IC 2) hold for \( (\omega, \Pi), (\omega', \Pi) \). If

\[
\max_{j=1,...,n} \| (\Pi A \omega)_j - (\Pi A \omega')_j \| < \epsilon,
\]

then

1. \( \max_{i=1,...,m} \| \omega_i - \omega_i' \| < \epsilon \) (stable recovery) and
2. \( \Pi = \Pi' \) (exact recovery).

In words, whenever two \( \delta \)-separable mixtures are close enough, the corresponding weights are also close and the corresponding sources are even the same.

Remark 2.3 (Converse exact recovery).

Note that the converse direction of Theorem 2.2 also holds up to a constant factor. More precisely, for any \( \epsilon > 0 \), if for some \( (\omega, \Pi), (\omega', \Pi) \) it holds that \( \max_{i=1,...,m} \| \omega_i - \omega_i' \| < \epsilon \) and \( \Pi = \Pi' \), then \( \max_{j=1,...,n} \| (\Pi A \omega)_j - (\Pi A \omega')_j \| < ma_k \epsilon \). This follows directly from the triangle inequality. Note that the identifiability conditions (\( \delta \)-IC 1) and (IC 2) are not needed for this direction.

Remark 2.4 (Separability for general mixing weights).

For arbitrary mixing matrices, not necessarily positive and summing up to one, \( \omega \in \{ \omega \in \mathbb{R}^{m \times M} : 0 < \| \omega_1 \| < \ldots < \| \omega_m \| \} \), separability (IC 2) can be defined analogously, where the uncertainty about sign and scaling of the weights leads to additional vectors which must be selected by \( \Pi \) to guarantee identifiability, see [4, Theorem 7.1]. Theorem 2.1 and 2.2 can be adapted accordingly.

Analog to \( \delta \) which regularizes the identifiability condition on \( \omega \), we further introduce a second parameter which regularizes separability of \( \Pi \) in (IC 2). To this end, \( \Pi A \) is denoted as \( \Lambda \)-separable, if the \( i \)-th unit vector \( e^i \) appears in \( \Pi A \) at least \( M \Lambda \) times for each \( i = 1, \ldots, m \), that is

\[
\# \{ j \in \{1, \ldots, n\} : (\Pi A)_j = e^i \} \geq M \Lambda \quad \text{for all } i = 1, \ldots, m.
\]

(Lambda-IC 2)

Note that the above condition is equivalent to \( \omega_i \) appearing at least \( M \Lambda \) times in the rows of \( G \). The scaling factor \( M \) naturally corresponds the dimension on \( \omega_i \).

Summing up, for fixed alphabet \( A \), number of sources \( m \), and \( A \) as in (17), we consider the parameter space of \( (\delta, \Lambda) \)-regularized response matrices

\[
N^{\delta, \Lambda} := \{ \Pi A \omega : \omega \in \Omega_{m,M}, \Pi \text{ as in (16) with (\( \delta \)-IC 1), (\Lambda-IC 2)}\}.
\]

Note that \( N^{\delta, \Lambda} \) depends on \( n \) and \( M \), which we suppress in the following whenever the dependence is not important.
3 Minimax rates for the prediction error

In the following, we provide lower and upper bounds for the minimax prediction rate. We consider the situation where both \( n, M \to \infty \) (note that the total sample size is \( nM \)). These bounds match (up to constants) whenever (10) holds, which is a very realistic sample size regime in many practical settings (see Remark 3.6).

The following theorem gives a lower bound on the minimax prediction error in the MABS model. To this end, for the fixed alphabet \( \mathcal{A} \), define the smallest pairwise difference

\[
\Delta A_{\min} := \min \{|a - a' : a \neq a' \in \mathcal{A}|\}.
\]

(19)

**Theorem 3.1** (Lower bound).

Consider the MABS model (15) with \((\delta, \Lambda)\)-regularized parameter space \( N_{\delta, \Lambda} \) as in (18). Further, assume that \( \sigma/\sqrt{8M} < \delta \leq (\Delta A_{\min})^2(90a_km)^{-1} \) and \( 1/M \leq \Lambda \leq \lfloor n/m \rfloor / M \). Let \( \min(n, M) \geq 2m \). Then

\[
\inf \sup_{\hat{\theta}} \mathbb{E}_{\Pi A_\omega} \left\| \hat{\theta} - \Pi A_\omega \right\|^2 \geq 0.4M \left( \frac{1}{(m - 1)\sigma^2} + \frac{288m^5a_k^2}{(\Delta A_{\min})^2n} \right)^{-1} + \frac{1}{2} \sigma \delta \sqrt{M} e^{-\frac{\mu_M^2}{\delta}}.
\]

The next theorem gives an upper bound on the minimax prediction error. It almost coincides with the lower bound from Theorem 3.1 (see Corollary 3.7). This upper bound is achieved by the LSE

\[
\hat{\theta} \in \arg\min_{\tilde{\theta} \in N_{\delta, \Lambda}} \| Y - \tilde{\theta} \|^2.
\]

(20)

Before we give the upper bound, let us make a few remarks about the LSE over the class \( N_{\delta, \Lambda} \) in (20).

**Remark 3.2** (Existence and uniqueness of LSE).

The LSE \( \hat{\theta} \) exists, that is the minimum in (20) is attained. To see this, note that

\[
N_{\delta, \Lambda} = \bigcup_{\Pi \Lambda\text{-separable}} \{\Pi A_\omega : \omega \in \Omega_{m, M}, \min(ASB(\omega), WSB(\omega)) \geq \delta\}
\]

(21)

is a finite union of closed, bounded in \([0, a_k]^{n \times M}\), hence, compact sets. The LSE \( \hat{\theta} \) in (20) is not always unique, but the upper bound in Theorem 3.5 holds for any minimizer. A counterexample is the following. Let \( n = m = 2, M = 1, \Lambda = 1, \mathcal{A} = \{0, 1, 2\}, \) and \( \delta = 0.01 \). By separability, \( \Pi A \) is restricted to the identity matrix. Thus

\[
\arg\min_{\tilde{\theta} \in N_{\delta, \Lambda}} \| Y - \tilde{\theta} \|^2 = \arg\min_{\omega \in \Omega_{2, 1}} \| Y - \omega \|^2 = \arg\min_{\omega \in \Omega_{2, 1}} \omega_1^2 - (1 + Y_1 + Y_2)\omega_1.
\]
with $\omega_1 = 1 - \omega_2$. Simple calculations give that $\Omega_{\delta,1} = \{ (\omega_1, 1 - \omega_1)^{\top} : \omega_1 \in [0.1, 3/10] \cup [11/30, 0.45] \}$. If the observations $Y_1, Y_2$ are such that $(1+Y_1+Y_2)/2 = 1/3$, it is easy to check that

$$\arg\min_{\omega \in \Omega_{\delta,1}} \omega_1^2 - (1 + Y_1 + Y_2)\omega_1 = \{ (3/10, 7/10)^{\top}, (11/30, 19/30)^{\top} \}.$$ 

**Remark 3.3** (Computation of LSE).

We are not aware of an efficient implementation of the LSE and we speculate that this is an NP-hard problem in general. Note that in [21] for $\Pi \neq \Pi'$, both $\Lambda$-separable, Theorem 2.2 implies that the corresponding two sets in the union in (21) are disjoint. Thus, computation of the LSE amounts to minimization over exponentially many (in $n$) disjoint, compact sets (see also Figure 1.1). [23] have shown that exact computation of the LSE is NP-hard in general, for the MABS model [15] with $M = 1$, but for arbitrary design $A$ (not the specific form in (17)) and restricted to permutation matrices $\Pi$ (not the bigger class of selection matrices). Although, their results do not directly apply to our setting, it is near at hand that exact computation of the LSE for the MABS model [15] is also not feasible. A natural computationally efficient approximation of the LSE in (20) is via an iterative Lloyd’s algorithm: On the one hand, given $\Pi$ computation of the LSE corresponds to a convex optimization problem which can be solved efficiently, see e.g. [32]. On the other hand, given the mixture matrix $\omega$ computation of the LSE corresponds to a simple LS clustering with known centers $A\omega \in \mathbb{R}^{K \times M}$. We found such an iterative approximation schemes to work well in practice, see [2].

In future work, it will be interesting to analyze its theoretical properties in more detail and derive whether its estimation rates match with the optimal ones derived in this paper.

**Remark 3.4** (Dependence on $\Lambda$ and $\delta$).

The LSE in (20) depends on the regularization parameters $\delta$ and $\Lambda$. One may ask whether this is indeed necessary or whether the full LSE (without restriction on the space $N^{\delta,\Lambda}$) would achieve the same optimal rates. At least for estimation, which is the main focus of this paper, this cannot hold true, in general. To see this, note that for given $\delta, \Lambda$ one can easily construct examples of pairs $(\omega, \Pi) \in N^{\delta,\Lambda}$ and $(\omega, \Pi') \notin N^{\delta,\Lambda}$ such that $\Pi A \omega = \Pi' A \omega$ (note that Theorem 2.2 requires both, $(\omega, \Pi)$ and $(\omega, \Pi')$, to fulfill the identifiability condition (8-IC 1) and (IC 2)). Hence, without prior knowledge on $\delta$ and $\Lambda$ it cannot be possible to consistently estimate $\Pi$ and $\omega$.

However, in most applications prior knowledge on $\delta$ and $\Lambda$ is usually available. $\Lambda$ corresponds to the number of times that specific alphabet combinations appear in the data. Often there is a good prior knowledge of the relative occurrence of alphabet values in a signal (e.g., the relative occurencies of 0’s and 1’s in a standard digital signal or the typical mutation patterns in cancer). From this, lower bounds on $\Lambda$ can easily be derived. $\delta$ corresponds to the minimal distance between mixture values. Although, prior knowledge for this quantity might be more difficult to obtain in practice, one can always work with the lower bound of Lemma A.2, which holds almost surely for $M$ sufficiently large (recall that this paper consider the asymptotic regime $n, M \to \infty$).
Theorem 3.5 (Upper bound).
Consider the MABS model \((15)\) with \((\delta, \Lambda)\)-regularized parameter space \(\mathcal{N}^{\delta, \Lambda}\) as in \((18)\) and let \(\hat{\theta}\) be the LSE in \((20)\). Then
\[
\sup_{\Pi, A_{\omega} \in A^{\delta}} \mathbb{E}_{\Pi, A_{\omega}} \left( \| \hat{\theta} - \Pi A_{\omega} \|^2 \right) 
\leq 4\sigma^2 mM + 12\sigma n^2 k^m m^{7/2} a_k^2 M^{3/2} e^{-\frac{\Lambda^2 M}{8(m+1)(1+ma_k)^2}\sigma^2}.
\]

Let throughout the following \(\gtrsim\) and \(\lesssim\) denote inequalities up to a universal constant which does not depend on any model parameter. Further, for the sample sizes \(n, M\) assume that
\[
M \geq \frac{144\sigma^2 m^3 a_k^3}{\Lambda^2 \Delta A_{\text{min}}^2} \left( 2 \ln(n) + 2 \ln(M) + \ln \left( a_k^3 m^{7/3} k^m \right) \right),
\]
\[
n \geq \min \left( m^6 a_k^2 \sigma^2 \left( \Delta A_{\text{min}} \right)^2, 2 \Lambda M m \right).
\]
(22)

Note that (22) holds eventually in the asymptotic regime \((10)\).

Remark 3.6 (Sample size regime in applications).
The asymptotic regime \((10)\), which is considered here, appears to be realistic in many applications. For example in digital communications \(n\) relates the length of the source signals, \(M\) relates to the number of receiver antennas, and \(m\) relates to the number of source signals which are send through a channel simultaneously. In some situations, the number of receiver antennas \(M\) might be fixed and given and one is interested in the minimal length of signal \(n\) that has to be processes at a time. In other situations, one might be interested in the minimal number of receiver antennas \(M\) which need to be employed in order to recover a signal of fixed and given length \(n\) exactly. Our results show that the number of receiver antennas \(M\) which are necessary to exactly recovery signals of length \(n\) grows logarithmic with \(n\). Similar, in cancer genetics, \(n\) corresponds to the number of genetic locations where CNV’s or SNP’s are measured. Again, our results show that when at least \(M \sim \log(n)\) samples of a patient at different locations or different time points are available, then \(n\) mutations of the \(m\) single tumor clones can be recovered exactly.

Note that in both situations it is much easier to collect a large number of source samples \((e.g.,\, genome-locations\, or\, length\, of\, transmitted\, digital\, signal)\) than a large number of mixture samples \((e.g.,\, tissue\, samples\, of\, a\, patient\, at\, different\, time\, points/locations\, or\, number\, of\, receiver\, antennas\, in\, a\, MIMO\, channel)\). Thus, the asymptotic regime \((10)\) appears realistic.

Corollary 3.7 (Minimax prediction rate).
Assume the setting of Theorem 3.5 and that \(\sigma/\sqrt{8M} < \delta < (\Delta A_{\text{min}})^2 (90a_k m)^{-1}\). Then, whenever (22) holds,
\[
\inf_{\hat{\theta}} \sup_{\Pi, A_{\omega} \in A^{\delta, \Lambda}} \mathbb{E}_{\Pi, A_{\omega}} \left( \frac{1}{nM} \| \hat{\theta} - \Pi A_{\omega} \|^2 \right) \gtrsim \frac{\sigma^2 m}{n} + \frac{\sigma \delta}{\sqrt{M}} e^{-\frac{\Lambda^2 M}{8(m+1)(1+ma_k)^2}\sigma^2},
\]
\[
\inf_{\hat{\theta}} \sup_{\Pi, A_{\omega} \in A^{\delta, \Lambda}} \mathbb{E}_{\Pi, A_{\omega}} \left( \frac{1}{nM} \| \hat{\theta} - \Pi A_{\omega} \|^2 \right) \lesssim \frac{\sigma^2 m}{n} + \frac{\sigma}{\sqrt{\Lambda \delta \sqrt{M}}} e^{-\frac{\Lambda^2 M}{8(m+1)(1+ma_k)^2}\sigma^2},
\]
with \( c := \frac{\Lambda \delta^2}{(16m^3(1 + ak)^2)} \), and the LSE achieves the second inequality.

Corollary 3.7 sheds light on the specific tradeoff between \( n \) and \( M \) regarding the prediction error of \( \Pi_A \omega \). The dependence on \( M \) vanishes exponentially fast (in the asymptotic regime (10)). Hence, for sufficiently large \( M \) the prediction rate is dominated by its first term, which is parametric in \( n \). Thus, as long as (10) holds, the unknown selection matrix \( \Pi \) in the linear model (15) does not play much of a role. Put it differently, if, in a multivariate linear model (1) the dimension \( M \) (the number of mixtures of the BSS problem (2)) is at least of order \( O(\ln(n)) \), then, under weak identifiability conditions, for the estimation accuracy of the signal \( F\omega \) it is irrelevant whether the design \( F \) is completely known or it is only known up to a finite set of possible values for its entries. This observation was already made for the more general setting of arbitrary design \( A \) with unknown selection matrices \( \Pi \) in [25]. However, here the additional assumption of the given finite alphabet constraint leads to a much faster decay (recall our discussion in the introduction). More importantly, the finite alphabet assumption provides identifiability and hence, allows separation of \( \omega \) and \( \Pi \), such that the results on the prediction error can be extend to the estimation error, as we will show in the following. This is not possible for the more general setting as in [25].

4 Estimation error

From the proof of Theorem 3.5 it is easy to derive the following upper bound on the maximal classification error \( P(\hat{\Pi} \neq \Pi) \).

**Corollary 4.1 (Upper bound on classification error).**

Assume the setting of Corollary 3.7 and let \( \hat{\theta} = \hat{\Pi}A\omega \) be the LSE in (20). Then, whenever (22) holds

\[
\sup_{\Pi A \omega, \Pi A' \omega \in N^\delta} P_{\Pi A \omega} (\hat{\Pi} \neq \Pi) \lesssim \frac{\sigma}{\sqrt{A \delta \sqrt{M}}} e^{-16m^3(1 + ak)^2} \frac{M \sigma}{\sqrt{M}}. \tag{23}
\]

In order to derive lower bounds for the maximal estimation error, one can combine Corollary 3.7 with Theorem 2.2. To this end, recall the metric \( d \) in (12) with \( F \) replaced by \( \Pi \), which combines the classification \( \Pi \neq \Pi \) and estimation \( \|\omega_i - \hat{\omega}_i\| \) error. The metrics \( d((\Pi, \omega), (\Pi', \omega)) \) and \( \|\Pi A \omega - \Pi A \omega\| \) are locally equivalent on \( N^{\delta, 1} \) as the following theorem shows.

**Theorem 4.2.**

Let \( \Pi A \omega, \Pi' A \omega' \in N^{\delta, 1} \) as in (18), then

1. \[
d((\Pi, \omega), (\Pi', \omega')) \geq \|\Pi A \omega - \Pi' A \omega'\|/(\sqrt{\delta} ma_k),
\]

2. if \( \|\Pi A \omega - \Pi' A \omega'\| \leq \delta \sqrt{M}/(1 + ma_k) \), then

\[
d((\Pi, \omega), (\Pi', \omega')) \leq \|\Pi A \omega - \Pi' A \omega'\|.
\]

15
The following theorem shows that the LSE is not only asymptotically minimax rate optimal for the prediction error as in Corollary 3.7, but also minimax rate optimal for the estimation error in terms of the metric \( d \) in (12).

**Theorem 4.3 (Minimax estimation rate).**

Assume the setting of Corollary 3.7. Then, whenever (22) holds

\[
\inf_{\Pi, \omega} \sup_{\Pi, \omega} E_{\Pi, \omega} \left( \frac{1}{M} d \left( (\Pi, \omega), (\hat{\Pi}, \hat{\omega}) \right)^2 \right) \gtrsim \frac{\sigma^2}{\Lambda M} + \frac{\sigma \delta}{m^2 a^2 \sqrt{M}} e^{-\frac{\sigma^2}{4m^3}} \frac{M}{\Lambda},
\]

\[
\inf_{\Pi, \omega} \sup_{\Pi, \omega} E_{\Pi, \omega} \left( \frac{1}{M} d \left( (\Pi, \omega), (\hat{\Pi}, \hat{\omega}) \right)^2 \right) \lesssim \frac{\sigma^2}{\Lambda M} + \frac{\sigma}{\sqrt{\Lambda \delta}} \sqrt{M} e^{-c M},
\]

with \( c := \Lambda \delta^2 / (16m^3(1 + a^2)^2) \), and the LSE in (20) achieves the second inequality.

Again, Theorem 4.3 shows that when \( M \) is sufficiently large, increasing \( M \) further does not influence the estimation rate in terms of \( d(\cdot, \cdot)^2 \). Moreover, the minimax estimation rate of \( d(\cdot, \cdot)^2 \) does not depend on \( n \), although the dimension of \( \Pi \) and \( \hat{\Pi} \), respectively, increase with \( n \). Thus, if \( \ln(n) \ll M \) the unknown selection \( \Pi \) in the linear model (15) does not play much of a role for the estimation rate, as well.

## 5 Conclusion and discussion

In this paper we introduced the Multivariate finite Alphabet Blind Separation (MABS) model with Gaussian noise, where we imposed weak regularity conditions to ensure identifiability of the model parameters. Depending on these quantities, we derived lower and upper bounds (attained by the LSE) of the maximal prediction error which coincide up to constants. In particular, our results reveal that, due to the finite alphabet structure, minimax rates are significantly improved, in the sense that the classification error vanishes exponentially as the number of mixtures grows, instead of parametric estimation rates of convergence which are optimal without the finite alphabet assumption (recall the discussion in Section 1 regarding the results of [23]). Most importantly, we could derive bounds for the estimation error from those for the prediction error. This only becomes feasible due to the finite alphabet structure, which provides identifiability and stable recovery under reasonable conditions. Again, we showed that the LSE attains the optimal rates for the estimation error. In particular, our results demonstrate that the unknown design does not influence the minimax rates when the number of mixtures \( M \) is at least of order \( \ln(n) \), where \( n \) is the number of observations. This is in strict contrast to a computational view on the MABS model. Whereas for known design computation of the LSE amounts to a convex optimization problem, for unknown finite alphabet design as in (1) it amounts to minimization over a disjoint union of exponentially many (in \( n \)) sets. A natural approximation scheme for the LSE, which performs well in practice, is via an iterative Lloyd’s algorithm (recall Remark 3.3). In future work, it will be of great interest to analyze the theoretical properties of such an algorithm in more details and compare its theoretical and practical behavior with the minimax benchmarks obtained in this paper.
In summary, a major consequence of this paper is that finite alphabet structures can significantly improve prediction accuracy and provide identifiability, thus enabling us to estimate \( \Pi \) and \( \omega \), separately. In a broader context, finite alphabet structures may be considered as a new type of sparsity – promising to be explored further.

A Appendix

A.1 Proof of the exact recovery result Theorem 2.2

Proof of Theorem 2.2 The separability condition implies that there exists \( e_i, \tilde{e}_i \in A^m \) for \( i = 1, \ldots, m \) such that

\[
\| \omega_i - \tilde{e}_i \omega \| < \epsilon \quad \text{and} \quad \| \omega_i - e_i \omega \| < \epsilon. \tag{24}
\]

We start with proving the first assertion by induction for \( i = 1, \ldots, m \). If either \( e_1 \) or \( \tilde{e}_1 \) equals the unit vector \((1, 0, \ldots, 0) \in \mathbb{R}^m \), (24) yields

\[
\| \omega_1 - \omega_1 \| < \epsilon. \tag{25}
\]

If \( e_1 \) or \( \tilde{e}_1 \) equals the zero vector \((0, \ldots, 0) \in \mathbb{R}^m \), then \( \text{ASB}(\omega), \text{ASB}(\omega) \geq \delta \) and (24) contradict. So assume that \( e_1 \) and \( \tilde{e}_1 \) both neither equal the first unit vector nor the zero vector and, in particular,

\[
e_1 \leq 1 \Rightarrow \sum_{i=2}^{m} e_i \geq 1 \tag{26}
\]

and analog for \( \tilde{e}_1 \). W.l.o.g. assume that \( \| \omega_1 \| \geq \| \omega_2 \| \). Then

\[
\| e_1 \omega \|^2 = \sum_{j=1}^{M} \left( \sum_{i=1}^{m} e_i \omega_{ij} \right)^2 \geq \min \left( a_2^2 \| \omega_1 \|^2, \| \omega_2 \|^2 \right), \tag{27}
\]

where the inequality follows from separating into the following cases. If \( e_1 \geq a_2 > 1 \), then

\[
\left( \sum_{i=1}^{m} e_i \omega_{ij} \right)^2 \geq (e_1)^2 \omega_{ij}^2 \geq a_2^2 \omega_{ij}^2.
\]

If \( e_1 \leq 1 \), then by (26) \( \exists r > 1 \) such that \( e_r \geq 1 \), and

\[
\left( \sum_{i=1}^{m} e_i \omega_{ij} \right)^2 \geq (e_r)^2 \omega_{ij}^2 \geq \omega_{ij}^2. \tag{28}
\]

In particular, (27) gives

\[
\| e_1 \omega \| - \| \omega_1 \| \geq \min \left( a_2 \| \omega_1 \|, \| \omega_2 \| - \| \omega_1 \| \right) \\
\geq \min \left( \| (a_2-1,0,\ldots,0)\omega \|, \| \omega_2 \| - \| \omega_1 \| \right) \geq \frac{2 \delta \sqrt{MT}}{(1 + ma_k)} > \epsilon \tag{29}
\]
and (24) gives
\[ \|e^1 \omega\| - \|\omega_1\| \leq \|e^1 \omega\| - \|\omega_1\| \leq \|e^1 \omega - \omega_1\| < \epsilon. \] (30)

(29) and (30) contradict, which shows (25).

Now assume that
\[ \|\omega_i \cdot - \omega_i \cdot\| < \epsilon \quad \text{for} \quad i = 1, \ldots, r - 1 \] (31)
and w.l.o.g. assume that
\[ \|\omega_r \cdot\| \geq \|\omega_r \cdot\|. \] (32)
First, assume that \( \sum_{i=r+1}^{m} e^r_i \geq 1 \). Then it follows from (24) that
\[ \|\omega_r \cdot\| = \|\omega_r \cdot - e^r \omega + e^r \omega\| \geq \|e^r \omega\| - \|\omega_r \cdot - e^r \omega\| \geq \|e^r \omega\| - \epsilon \]
\[ \geq \|\omega_{r+1} \cdot\| - \epsilon \geq \|\omega_r \cdot\| + 2\delta \sqrt{M}/(1 + ma_k) - \epsilon > \|\omega_r \cdot\|, \] (33)
where for the third inequality we used an analog argument as in (28). (33) contradicts (32). Thus, it follows that
\[ e^r_{r+1} = \ldots = e^r_m = 0. \] (34)
Further, if \( e^r_r = 0 \), then
\[ \|\omega_r \cdot - e^r \omega\| \leq \|\omega_r \cdot - e^r \omega\| + \|e^r \omega - e^r \omega\| \]
\[ \leq \|\omega_r \cdot - e^r \omega\| + (r - 1)a_k \epsilon \leq (1 + (r - 1)a_k) \epsilon < \delta \sqrt{M}, \] (35)
where the second inequality follows from (31) and third inequality from (24). (35) and \( ASB(\omega) \geq \delta \) contradict. Thus, it follows that
\[ e^r_r \geq 1. \] (36)

Note that (34), (36) and (24) imply that
\[ \epsilon > \|\omega_r \cdot - e^r \omega\| = \|\omega_r \cdot - \left( (e^r_r - 1)\omega_r \cdot + \sum_{i=1}^{r-1} e^r_i \omega_i \right)\| \]
\[ \geq \|\omega_r \cdot - (\omega_r \cdot - x)\|, \] (37)
with \( x_j = \)
\[ \begin{cases} 
(e^r_r - 1)\omega_r \cdot + \sum_{i=1}^{r-1} e^r_i \omega_i \bigg)_{j} & \text{if } \left( (e^r_r - 1)\omega_r \cdot + \sum_{i=1}^{r-1} e^r_i \omega_i \bigg) \leq \omega_{rj}, \\
\omega_{rj} & \text{otherwise.}
\end{cases} \]

As \( x, \omega_r \), and \( \omega_r - x \) have non-negative entries, it also follows that
\[ \|\omega_r \cdot\| \geq \|\omega_r \cdot - x\|. \] (38)

18
Thus, by (32) and (24) it follows that which contradicts (24). Hence, it follows that where for the second inequality we used (24) and (31). (39) and

Further, if \( \tilde{e}_r \) = 0, then \( \tilde{e}_r \omega = \sum_{i=1}^{r-1} e_i \omega_i \) and

where for the second inequality we used (24) and (31). (39) and ASB(\( \omega \)) \( \geq \delta \) contradict. Thus it follows that \( \tilde{e}_r \geq 1 \). However, this implies that

which contradicts [24]. Hence, it follows that

Thus, by [32] and [24] it follows that

where for the first inequality we used the fact that for two vectors \( a, b \) with \( \|a\| \geq \|b\| \) and a constant \( c \geq 1 \) it follows that \( \|ca - b\| \geq \|a - b\| \). Thus, the first assertion follows by induction.

To show the second assertion, assume the contrary, i.e., that there exists \( e \neq e' \in \mathcal{K}^m \) such that

As ASB(\( \omega \)) \( \geq \delta \), it follows that \( \|\omega - e'\omega\| \geq \delta \sqrt{M} \) and by the first assertion of the theorem, it follows that \( \|e\omega - e'\omega\| \leq ma\epsilon \). Therefore [42] implies \( \epsilon > \delta \sqrt{M} - ma\epsilon \), which is a contradiction.

### A.2 Additional lemmas on the ASB

For the following considerations we define the space of \( \delta \)-separable mixing weights as

\[
\Omega_{m,M}^\delta := \{ \omega \in \Omega_{m,M} : \min (\text{ASB}(\omega), WSB(\omega)) \geq \delta \}.
\]  

#### Lemma A.1

If \( \Omega_{m,M}^\delta \) in (43) is non-empty for some \( m, M \in \mathbb{N} \), then \( \delta \leq \frac{(1+ma\epsilon)}{\sqrt{2m(m+1)}} \).
Proof. If $\Omega^\delta_{m,M}$ in (43) is non-empty, then there exists an $\omega \in \Omega^\delta_{m,M}$ with

$$\delta \leq ASB(\omega) \leq \frac{1}{\sqrt{M}} \sqrt{\sum_{j=1}^{M} \omega_{ij}^2}$$

and

$$\sqrt{M} \delta \leq \sqrt{M}WSB(\omega) \leq (1 + ma_k)/2 (\|\omega_1\| - \|\omega_{i-1}\|) \leq (1 + ma_k)/2 \sqrt{\|\omega_i\|^2 - \|\omega_{i-1}\|^2}$$

for all $i = 1, \ldots, m$, with $\omega_{mj} = 1 - \omega_1j - \ldots - \omega_{m-1}j$. In particular, there exists $\omega \in \mathbb{R}^{m \times M}$ with

$$\frac{4\delta^2 M}{(1 + ma_k)^2} \leq \min \left( \sum_{j=1}^{M} \omega_{1j}, \min_{i=2, \ldots, m-1} \sum_{j=1}^{M} (\omega_{ij}^2 - \omega_{i-1j}^2) \right),$$

$$\sum_{j=1}^{M} \left( (1 - \omega_1j - \ldots - \omega_{m-1}j)^2 - \omega_{m-1j}^2 \right).$$

Moreover,

$$(1 - \omega_1j - \ldots - \omega_{m-1}j)^2 = 1 + (\omega_1j + \ldots + \omega_{m-1}j)^2 - 2\omega_1j - \ldots - 2\omega_{m-1}j$$

$$= 1 - \sum_{i=1}^{m-1} \omega_{ij} (2 - \sum_{s=1}^{m-1} \omega_{sj}) \leq 1 - \omega_{1j}^2 - \ldots - \omega_{m-1j}^2.$$

And thus,

$$\frac{\delta^2 4M}{(1 + ma_k)^2} \leq \min \left( \sum_{j=1}^{M} \omega_{1j}^2, \min_{i=2, \ldots, m-1} \sum_{j=1}^{M} (\omega_{ij}^2 - \omega_{i-1j}^2), \sum_{j=1}^{M} (1 - \omega_1j - \ldots - 2\omega_{m-1j}) \right)$$

$$\leq \max_{x \in \mathbb{R}^{m-1}} \min (x_1, x_2, \ldots, x_{m-1}, (M - mx_1 - \ldots - 2x_{m-1}) = \frac{2M}{m(m + 1)},$$

where for the second inequality we used $x_1 := \sum_{j=1}^{M} \omega_{1j}^2$ and $x_i := \sum_{j=1}^{M} \omega_{ij}^2 - \sum_{j=1}^{M} \omega_{i-1j}^2$ for $i = 2, \ldots, m$. \hfill $\Box$

For the following considerations, we extend the definition in (19) for the fixed alphabet $\mathcal{A}$ and number of sources $m$ by

$$\Delta^2 \mathcal{A}^m := \{ e_1 - e_2 : e_1 \neq e_2 \in \mathcal{A}^m \}, \quad \Delta^2 \mathcal{A}^m_{\min} := \min_{x \in \Delta^2 \mathcal{A}^1} |x|,$$

$$\Delta^2 \mathcal{A}^m := \{ e_1 - e_2 : e_1 \neq e_2 \in \Delta^2 \mathcal{A}^m \}, \quad \Delta^2 \mathcal{A}^m_{\min} := \min_{x \in \Delta^2 \mathcal{A}^1} |x|. \quad (44)$$

20
Define the constants
\[ c = c(m, \mathfrak A) := \frac{\sqrt{2} \Delta_{\text{min}}}{\sqrt{3} k^2 m^2 (m-1)}, \quad C = C(m, \mathfrak A) := \frac{\sqrt{2} (1 + ma_k)}{\sqrt{m(m-1)}}. \] 

Lemma A.2.
If \( \omega \) is uniformly distributed on \( \Omega_{m,M} \) in (3), then for \( c, C \) as in (45) it holds almost surely that
\[ c < \liminf_{M \to \infty} \text{ASB}(\omega) \leq \limsup_{M \to \infty} \text{ASB}(\omega) < C. \]

Proof of Lemma A.2. It follows from Lemma A.1 that \( \text{ASB}(\omega) \) is surely bounded from above by \( (1 + ma_k)/\sqrt{2(m(m-1))} \), which shows the inequality on the right hand side. Further, if \( M = 1 \) and \( \omega \) is drawn uniformly, then it can be shown that \( P(\text{ASB}(\omega) > \delta) \geq 1 - d \delta \) with \( d = k^2 m^2 (m-1)/(\sqrt{2} \Delta_{\text{min}}) \). For arbitrary \( M \in \mathbb{N} \), if \( \omega \) is drawn uniformly its ASB is bounded by the sum of the corresponding ASB's of the single components, i.e., \( \text{MASB}(\omega) \geq \sum_{j=1}^{M} \text{ASB}(\omega_j)^2 \), where \( \text{ASB}(\omega_j), j = 1, \ldots, M \), are independent and identically distributed with
\[ E(\text{ASB}(\omega_j)^2) \geq \int_0^\infty (1 - d \sqrt{x})_+ dx = \frac{1}{3d^2}. \]
Hence, for \( c < \frac{1}{\sqrt{3} d} \) it follows from the strong law of large numbers that almost surely
\[ \liminf_{M \to \infty} \text{ASB}(\omega)^2 \geq \liminf_{M \to \infty} \frac{1}{M} \sum_{j=1}^{M} \text{ASB}(\omega_j)^2 = E(\text{ASB}(\omega_j)^2) > c^2, \]
which shows the inequality on the left hand side. \( \square \)

Lemma A.3.
\( \Omega_{m,m}^{\Delta_{\text{min}}/\sqrt{m}} \) as in (44) is non-empty for any \( m \in \mathbb{N} \), with \( \Delta_{\text{min}} \) as in (44).

Proof. For \( 1/(2\sqrt{m}) > \delta > 0 \) define
\[ \omega^\delta := I_{m \times m} - \frac{2\delta \sqrt{m}}{1 + ma_k} \begin{pmatrix} m-1 & 0 & \ldots & 0 & 0 \\ 0 & m-2 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \\ -(m-1) & -(m-2) & \ldots & -1 & 0 \end{pmatrix}, \] 
where \( I_{m \times m} \) denotes the \( m \times m \) identity matrix. As \( \frac{2\delta \sqrt{m}}{1 + ma_k} (m-1) < 1 \), it holds for \( i = m-1, \ldots, 1 \) that
\[ \left| 1 - \frac{2\delta \sqrt{m}}{1 + ma_k} (i-1) \right| \geq \left| 1 - \frac{2\delta \sqrt{m}}{1 + ma_k} \right| \geq \frac{2\delta \sqrt{m}}{1 + ma_k}. \]
and thus $WSB(\omega^\delta) \geq \delta$. Consequently, if $ASB(\omega^\delta) \geq \delta$ it follows that $\Omega_{m,m}^\delta$ is non-empty. We have that

$$\sqrt{m}ASB(\omega^\delta) \geq \sqrt{m}ASB(I_{m \times m}) - \frac{2\delta \sqrt{m}}{1 + ma_k} \max_{e \neq 0 \in \Delta \mathbb{A}^m} \sqrt{(m-1)(e_1 - e_m)^2 + \ldots + (e_{m-1} - e_m)^2}

= \Delta \mathbb{A}_{\min} - \frac{2\delta \sqrt{m}}{1 + ma_k} \max_{e \neq 0 \in \Delta \mathbb{A}^m} \sqrt{(m-1)(e_1 - e_m)^2 + \ldots + (e_{m-1} - e_m)^2}

\geq \Delta \mathbb{A}_{\min} - \frac{2\delta \sqrt{m} 2a_k}{1 + ma_k} \sum_{i=1}^{m-1} (m-i) = \Delta \mathbb{A}_{\min} - \frac{\delta \sqrt{m} 4a_k}{1 + ma_k} \sqrt{\frac{m(m-1)}{2}},$$

which implies that if

$$\sqrt{m}\delta \leq \Delta \mathbb{A}_{\min} - \frac{\delta \sqrt{8a_k m(m-1)}}{1 + ma_k} \quad (47)$$

then $\Omega_{m,m}^\delta$ is non-empty. As

$$\left(1 + \frac{\sqrt{8a_k m(m+1)}}{1 + ma_k}\right)^{-1} \geq \left(1 + \sqrt{\frac{8(m+1)}{m}}\right)^{-1} \geq 0.2,$$

[47] holds for $\delta = 0.2\Delta \mathbb{A}_{\min}/\sqrt{m}$. \hfill \Box

**Lemma A.4.**

For any $M \geq m$ it holds that $\Omega_{m,M}^{0.2\Delta \mathbb{A}_{\min} \sqrt{\lceil M/m \rceil}/\sqrt{M}}$ in (43) is non-empty.

**Proof.** By Lemma A.3, there exists $\omega' \in \Omega_{m,m}^{0.2\Delta \mathbb{A}_{\min}/\sqrt{m}}$. In particular, in holds that $\min(ASB(\omega'), WSB(\omega')) \geq 0.2\Delta \mathbb{A}_{\min}/\sqrt{m}$. Define

$$\omega = (\omega', \ldots, \omega', e^m, \ldots, e^m) \in \Omega_{m,M}.$$ Then $ASB(\omega), WSB(\omega) \geq 0.2\Delta \mathbb{A}_{\min} \sqrt{\lceil M/m \rceil}/\sqrt{M}$. \hfill \Box

**Lemma A.5.**

If $\Omega_{m,M}^\delta$ is non-empty for some $\delta > 0$, then there exists $\omega \in \Omega_{m,M}^\delta$ with $ASB(\omega) = \delta$.

**Proof.** Fix some $\omega \in \Omega_{m,M}^\delta$ and for $0 \leq \epsilon \leq 1$ define $\omega^\epsilon \in \Omega_{m,M}$ as

$$\omega^\epsilon_i = \begin{cases} \omega_{ij} & \text{if } i \notin \{1, m\}, \\ \epsilon \omega_{ij} & \text{if } i = 1, \\ \omega_{mj} + (1 - \epsilon)\omega_{1j} & \text{if } i = m. \end{cases}$$

Then $WSB(\omega^\epsilon) \geq WSB(\omega) \geq \delta$ for all $0 \leq \epsilon \leq 1$ and $\epsilon \mapsto ASB(\omega^\epsilon)$ is continuous with $ASB(\omega^0) = 0$ and $ASB(\omega^1) \geq \delta$. Thus, there exists an $\epsilon^* \in (0, 1]$ such that $\omega^{\epsilon^*}$ has the desired properties. \hfill \Box
Lemma A.6.  
If \( \Omega_{m,M}^{\delta} \) in (47) is non-empty, then there exists a quadratic matrix \( \omega \in \Omega_{m,M} \) such that \( \text{ASB}(\omega) = \frac{\delta \Delta \mathcal{A}_{\text{min}}}{(9 \sqrt{m} a_k)} \) and \( WSB(\omega) \geq \frac{\delta \Delta \mathcal{A}_{\text{min}}}{(9 \sqrt{m} a_k)} \).

Proof. If \( \Omega_{m,M}^{\delta} \) is non-empty, it follows from Lemma A.1 that \( \delta \leq \frac{1+m a_k}{\sqrt{2m(m+1)}} \), and hence, \( \delta \Delta \mathcal{A}_{\text{min}}/(9 a_k) \leq 0.2 \Delta \mathcal{A}_{\text{min}} \). Thus, by Lemma A.3 it follows that
\[
\Omega_{m,M}^{\delta \Delta \mathcal{A}_{\text{min}}}(9 \sqrt{m} a_k) \supseteq \Omega_{m,M}^{\delta \Delta \mathcal{A}_{\text{min}}}(9 \sqrt{m} a_k) \neq \emptyset.
\]
I.e., there exists \( \omega \in \Omega_{m,M} \) such that \( \text{ASB}(\omega), WSB(\omega) \geq \frac{\delta \Delta \mathcal{A}_{\text{min}}}{(9 \sqrt{m} a_k)} \).

Now the assertion follows from Lemma A.5. \( \square \)

A.3 Proof of the lower bound in Theorem 3.1

Proof of Theorem 3.1. The proof of Theorem 3.1 is divided into two steps, corresponding to the two different estimation errors of \( \bar{\Omega} \) and \( \hat{\Omega} \), respectively. We start with the first term on the r.h.s. of the assertion which corresponds to the estimation error of \( \hat{\Omega} \). The idea is to construct a hyperrectangle of maximal size which is a subset of \( \mathcal{X}^{\delta,\Lambda} \) and then apply results of [10] (for fixed selection matrix \( \Lambda \)).

In the following, \( \omega^* \) will denote the center of this hyperrectangle and the matrix \( E \) will denote the perturbation (of maximal size) around \( \omega^* \). To this end, let \( \omega^* \in \Omega_{m,M} \) be such that \( \text{ASB}(\omega^*) = 0.2 \Delta \mathcal{A}_{\text{min}} \sqrt{\frac{M}{m}} \), \( WSB(\omega^*) \geq 0.2 \Delta \mathcal{A}_{\text{min}} \sqrt{\frac{M}{m}} \), and \( \omega^*_j \geq 0.4 \Delta \mathcal{A}_{\text{min}}/(1 + m a_k) \) for \( j = 1, \ldots, M \) (existence follows from Lemma A.4 and A.5). For \( \epsilon \in (0,1)^{\epsilon+M} \) define
\[
\omega^\epsilon := \omega^* + \begin{pmatrix}
\epsilon_{11} & \cdots & \epsilon_{1M} \\
\vdots & & \vdots \\
\epsilon_{(m-1)1} & \cdots & \epsilon_{(m-1)M} \\
- \sum_{i=1}^{m-1} \epsilon_{i1} & \cdots & - \sum_{i=1}^{m-1} \epsilon_{iM}
\end{pmatrix} = \omega^* + \epsilon.
\]

Let \( \tau := \max_{ij} |\epsilon_{ij}| \). If
\[
\frac{0.4 \Delta \mathcal{A}_{\text{min}}}{1 + m a_k} \geq \tau (m-1),
\]
then all entries of \( \omega^\epsilon \) are non-negative and \( \omega^\epsilon \in \Omega_{m,M} \). For \( \omega^\epsilon \) to be an element of \( \Omega_{m,M}^\delta \) we further need that \( WSB(\omega^\epsilon), \text{ASB}(\omega^\epsilon) \geq \delta \). To this end, note that
\[
\| \omega^\epsilon_{i1} \| - \| \omega^\epsilon_{i-1} \| = \| \omega^\epsilon_{i1} + E_{i1} \| - \| \omega^\epsilon_{i-1} + E_{i-1} \|
\geq \| \omega^\epsilon_{i1} \| - \| \omega^\epsilon_{i-1} \| - \| E_{i1} \| - \| E_{i-1} \|
\geq \frac{2}{1 + m a_k} 0.2 \Delta \mathcal{A}_{\text{min}} \sqrt{\frac{M}{m}} - m \sqrt{M} \tau
\]
and thus
\[
\sqrt{M} WSB(\omega^\epsilon) \geq 0.2 \Delta \mathcal{A}_{\text{min}} \sqrt{\frac{M}{m}} - \frac{1 + m a_k}{2 m \sqrt{M} \tau}.
\]

23
Further, note that for \( e \in \Delta A_m \) and \( E \) as in \((48)\)
\[
\|eE\|_2^2 = \sum_{j=1}^{M} \left( \sum_{i=1}^{m-1} (e_i - e_m) e_{ij} \right)^2 \leq M ((m - 1)2a_k \tau)^2
\]
and thus,
\[
\sqrt{M}ASB(\omega^*') = \min_{e \in \Delta A_m} \|e(\omega^* + E)\| \geq \min_{e \in \Delta A_m} \|e\omega^*\| - \|eE\| = \sqrt{M}ASB(\omega^*) - \|eE\| \geq 0.2\Delta A_m \min \sqrt{\lfloor M/m \rfloor} - 2(m - 1)a_k \tau \sqrt{M}.
\]

Summing up, \((50)\) and \((51)\) yield that \( \omega^* \in \Omega^\delta_{m,M} \) if
\[
\sqrt{M} \delta \leq 0.2\Delta A_m \sqrt{\lfloor M/m \rfloor} - m^2 a_k \tau \sqrt{M}.
\]

Now let \( \Pi^* \) be the selection matrix such that
\[
\Pi^* A = \begin{pmatrix}
e1 & e2 & \ldots & e_m & e1 & e2 & \ldots
\end{pmatrix}^\top,
\]
where \( e_r \in \mathbb{R}^m \) is the \( r \)-th unit vector in \( \mathbb{R}^m \). Then, as \( \Lambda \leq \lfloor n/m \rfloor / M \),
\[
\Theta := \{ \Pi^* A \omega^* : \epsilon \in [0, \epsilon^*]^{m-1 \times M} \} \subset \mathcal{N}^{\delta, \Lambda}
\]
and for \( \Pi^* A \omega^* \in \Theta \) one observes in \((15)\)
\[
Y_1, \ldots, Y_{\lfloor n/m \rfloor} \overset{i.i.d.}{\sim} N(\omega^*, \sigma^2 I_{Mm \times Mm}).
\]

With this we mean that \( Y_i \) is the \( m \times M \) sub-matrix of \( Y \) in \((15)\) which consists of \( m \) successive row vectors of \( Y \). The expectation of \( Y_i \) is \( \omega^* \) and the entries of \( Y_i \) are independent normally distributed with variance \( \sigma^2 \). Define
\[
\tilde{\Theta} := \{ \omega^* : \epsilon \in [0, \epsilon^*]^{m-1 \times M} \}.
\]

\( \tilde{\Theta} \) is almost an hyperractangle. To make it a proper hyperractangle, we have to remove the last column of the matrices in \( \Theta \), namely
\[
\tilde{\Theta}' := \{ \omega^*_{ij} : 1 \leq i \leq m - 1 : \epsilon \in [0, \epsilon^*]^{m-1 \times M} \}.
\]

Note that
\[
\inf_{\tilde{\Theta}} \sup_{\theta} E_\theta \left( \|\hat{\theta} - \theta\|_2^2 \right) \geq \inf_{\tilde{\Theta}} \sup_{\theta} E_\theta \left( \|\hat{\theta} - \theta^\prime\|_2^2 \right).
\]
Then it follows from [10], (2.1), (3.4) and Proposition 3 that
\[
\inf_{\hat{\theta}} \sup_{\Pi A \omega \in \Theta} E_{\Pi A \omega} \left( \| \hat{\theta} - \Pi A \omega \|^2 \right) \\
\geq \inf_{\hat{\theta}} \sup_{\Pi A \omega \in \Theta} E_{\Pi A \omega} \left( \| \hat{\theta} - \Pi A \omega \|^2 \right) \\
= \lfloor n/m \rfloor \inf_{\omega \in \tilde{\Theta}} E_{\omega} \left( \| \hat{\theta} - \omega \|^2 \right) \\
\geq \lfloor n/m \rfloor \inf_{\omega \in \tilde{\Theta}} E_{\omega} \left( \| \hat{\theta} - \omega \|^2 \right) \\
\geq (1.25)^{-1} \lfloor n/m \rfloor M(m - 1) \left( \frac{\epsilon^* \sigma^2}{n/m} \right) \left( \frac{\epsilon^*}{n/m} + \frac{\sigma^2}{n \epsilon^2} \right)^{-1},
\]
where for the last inequality we used \( n \geq 2m \). Together with (53) this gives
\[
\inf_{\hat{\theta}} \sup_{\Pi A \omega \in \Theta} E_{\Pi A \omega} \left( \| \hat{\theta} - \Pi A \omega \|^2 \right) \geq 0.4 M \left( \frac{1}{\sigma^2(m - 1)} + \frac{288 m^5 a^2_k}{M(m - 1)} \right)^{-1} .
\]
Now we show the second part of the proof which corresponds to the estimation error of \( \Pi \). The idea is to fix a suitable mixing matrix \( \omega \in \Omega_{m,M}^a \) and thus, reduce the estimation problem to a classification problem on the finite set of possible selection matrices \( \Pi \). This can be considered as a testing problem which allows to apply the Neyman-Pearson lemma. As \( \delta \leq \frac{0.2}{9 a_k} (\Delta \Omega_{\min}^a)^2 [M/m]/M \) it follows that \( 9 \delta a_k / \Delta \Omega_{\min}^a \sqrt{M/[M/m]} \leq 0.2 \Delta \Omega_{\min} \sqrt{[M/m]/\sqrt{M}} \). Thus, by Lemma A.4, \( \Omega_{m,M}^a, \Omega_{m,M}^a = \Omega_{m,M}^a \) is non-empty and hence, by Lemma A.6 there exists a quadratic mixing matrix \( \omega^\delta \in \Omega_{m,m}^a \) such that
\[
\sqrt{m ASB} (\omega^\delta) = (9 \delta a_k / \Delta \Omega_{\min}^a \sqrt{M/[M/m]}) / (9 a_k / \Delta \Omega_{\min}^a) = \delta \sqrt{M/\sqrt{[M/m]}} \\
\text{and} \sqrt{m WSB} (\omega^\delta) \geq \delta \sqrt{M/\sqrt{[M/m]}}. \]
Hence,
\[
\Theta := \{ \Pi A(\omega^\delta, ..., \omega^\delta), e^m, ..., e^m) : \Pi \Lambda \text{-separable} \} \subset \mathcal{N}^{\delta, \Lambda} . \quad (56)
\]
Then for \( \Pi A \omega \in \Theta \) one observes in (15)
\[
Y_1, ..., Y_{[M/m]} \overset{i.i.d.}{\sim} \mathcal{N}(\Pi A \omega^\delta, \sigma^2 I_{nm \times nm}) . \quad (57)
\]
With this we mean that \( Y_i \) is the \( n \times m \) sub-matrix of \( Y \) in (15) which consists of \( m \) successive column vectors of \( Y \).
For the finite parameter space \( \tilde{\Theta} := \{ \Pi \omega^\delta : \Pi \text{ \Lambda-separable} \} \) it holds that
\[
\min_{\theta \neq \theta' \in \tilde{\Theta}} ||\theta - \theta'||^2 = \min_{\theta \neq \theta' \in \tilde{\Theta}} \sum_{j=1}^n ||\theta_j - \theta'_j||^2
\]
= \min_{\theta \neq \theta' \in \tilde{\Theta}} ||\theta_j - \theta'_j||^2 = mAB(\omega^\delta)^2 = \frac{\delta^2 M}{[M/m]}.

Lemma A.7 yields for any estimator \( \hat{\theta} \)
\[
\sup_{\Pi \omega \in N^{\delta, \Lambda}} E_{\Pi \omega} \left( ||\hat{\theta} - \Pi \omega||^2 \right) \geq \sup_{\theta \in \Theta} E_{\theta} \left( ||\hat{\theta} - \theta||^2 \right)
\]
= \sup \{M/m\} E_{\theta} \left( ||\hat{\theta} - \theta||^2 \right) \quad (58)
\[
\geq \delta^2 \sup_{\theta \in \Theta} P_{\theta} \left( \hat{\theta} \neq \theta \right).
\]
Now let \( \theta, \theta' \in \tilde{\Theta} \) be fixed such that
\[
||\theta - \theta'||^2 = \frac{\delta^2 M}{[M/m]}.
\]
Then the Neyman-Pearson lemma yields for \( \bar{Y} := \sum_{i=1}^{[M/m]} m Y_i / M \) with \( Y_i \) as in (57) and normally distributed \( Z \sim N(0, \sigma^2/[M/m]) \) that
\[
\sup_{\theta \in \tilde{\Theta}} P_{\theta} \left( \hat{\theta} \neq \theta \right) \geq \frac{1}{2} \left( P_{\theta} \left( \hat{\theta} \neq \theta \right) + P_{\theta'} \left( \hat{\theta} \neq \theta' \right) \right)
\]
= \frac{1}{2} \left( P_{\theta} \left( \hat{\theta} \neq \theta \right) + P_{\theta'} \left( \hat{\theta} = \theta \right) \right)
\]
\[
\geq \frac{1}{2} \inf_{u \in R} \left( P_{\theta} \left( ||\bar{Y} - \theta||^2 - ||\bar{Y} - \theta'||^2 > u \right) + P_{\theta'} \left( ||\bar{Y} - \theta||^2 - ||\bar{Y} - \theta'||^2 < u \right) \right)
\]
= \frac{1}{2} \inf_{u \in R} \left( P \left( 2Z^T(\theta' - \theta) > u + ||\theta - \theta'||^2 \right) + P \left( 2Z^T(\theta' - \theta) < u - ||\theta - \theta'||^2 \right) \right)
\]
= \frac{1}{2} \inf_{u \in R} \left( \frac{||\theta - \theta'||^2}{2} - \frac{||\theta - \theta'||^2}{2} \right)
\]
= \frac{1}{2} \inf_{u \in R} \left( \frac{||\theta - \theta'||^2}{2} \right)
\]
= 1 - \Psi \left( \frac{||\theta - \theta'||^2}{2} \right)
\]
= 1 - \Psi \left( \frac{\delta \sqrt{M}}{2 \sigma} \right)
\]
= 1 - \Psi \left( \frac{\delta \sqrt{M}}{2 \sigma} \right)
\]
= \frac{\sigma^2}{2 \delta \sqrt{M}} e^{-\frac{\delta^2 M}{8 \sigma^2}},
where \( \Psi \) denotes the cumulative distribution function of the standard normal and the last inequality follows from Mill’s ratio and \( 1 - 4\sigma^2/(\delta^2 M) \geq 1/2 \) as \( \sqrt{M} \delta \geq \sigma \sqrt{8} \). With (58) this gives
\[
\sup_{\Pi \omega \in N^{\delta, \Lambda}} E_{\Pi \omega} \left( ||\hat{\theta} - \Pi \omega||^2 \right) \geq \delta \sqrt{M} \frac{\sigma^2}{2 \sigma^2} e^{-\frac{\delta^2 M}{8 \sigma^2}}
\]
This finishes the proof.
A.4 Proof of the upper bound in Theorem 3.5

Proof of Theorem 3.5

By Theorem 2.2 we can write \( \hat{\theta} = \hat{\Pi}A\hat{\omega} \) in a unique way. We have that for any \( \Pi A\omega \in N^{\delta, \Lambda} \)

\[
E_{\Pi A\omega} \left( \| \hat{\theta} - \Pi A\omega \|^2 \right) = E_{\Pi A\omega} \left( \| \hat{\theta} - \Pi A\omega \|^2 I_{\{\hat{\Pi} = \Pi\}} \right) + E_{\Pi A\omega} \left( \| \hat{\theta} - \Pi A\omega \|^2 I_{\{\hat{\Pi} \neq \Pi\}} \right). \tag{59}
\]

We start with the second term. The idea is to bound it with the classification error, that is \( P_{\Pi A\omega} (\hat{\Pi} \neq \Pi) \), and then apply exact recovery as in Theorem 2.2.

As the entries of the \( n \times M \) matrices \( \hat{\theta} \) and \( \Pi A\omega \) are contained in the range of the alphabet \([0, a_k]\), it follows that

\[
E_{\Pi A\omega} \left( \| \hat{\theta} - \Pi A\omega \|^2 I_{\{\hat{\Pi} \neq \Pi\}} \right) \leq a_k^2 n M P_{\Pi A\omega} (\hat{\Pi} \neq \Pi) = a_k^2 n M \sum_{\Pi \neq \Pi} P_{\Pi A\omega} (\hat{\Pi} = \Pi).
\]

For a fixed \( \Lambda \)-separable \( \Pi \neq \Pi \) and any \( \omega \in \Omega^\delta \) it follows from Theorem 2.2 that \( \| \Pi A\omega - \Pi A\omega \| \geq M\sqrt{\Lambda \delta / (1 + ma_k)} =: c \). Further, as by separability rank(\( \Pi A\omega \)) = \( m \) and \( \Omega^\delta \subset \mathbb{R}^{m,M} \), there exists a rotation matrix \( R \) such that for \( \Theta := \{ \Pi A\omega - \Pi A\omega : \omega \in \Omega^\delta \} \subset \mathbb{R}^{nM} \) and \( \hat{\Theta} := R\Theta \) it holds for all \( \theta \in \hat{\Theta} \) that \( \theta_{mM+2} = \ldots = \theta_{nM} = 0 \). This gives

\[
P_{\Pi A\omega} (\hat{\Pi} = \Pi) \leq P_{\Pi A\omega} \left( \| Y - \Pi A\omega \|^2 > \min_{\omega \in \Omega^\delta_{m,M}} \| Y - \Pi A\omega \|^2 \right)

= P_{\Pi A\omega} \left( \| Z \|^2 > \min_{\theta \in \Theta} \| Z + \theta \|^2 \right)

= P \left( \| Z \|^2 > \min_{\theta \in \Theta} \| Z + \theta \|^2 \right)

= P \left( \max_{\theta \in \Theta} -2Z^T\frac{\theta}{\|\theta\|} = \frac{\|\theta\|}{\sigma} > 0 \right)

\leq P \left( \max_{i=1, \ldots, mM+1} \frac{Z_i}{\sigma} \max_{\theta \in \Theta} \frac{\sum_{j=1}^{mM+1} |\theta_j|}{\|\theta\|} > \frac{c}{2\sigma} \right)

\leq (Mm + 1) P \left( \frac{Z_i}{\sigma} \sqrt{mM + 1} > \frac{c}{2\sigma} \right)

= (Mm + 1) P \left( |\mathcal{N}(0,1)| > \frac{c}{2\sqrt{mM + 1}\sigma} \right)

\leq (Mm + 1) \frac{2\sqrt{mM + 1}\sigma}{c} e^{-\frac{c^2}{8(mM + 1)\sigma^2}},
\]

where we considered the noise matrix \( Z \in \mathbb{R}^{nM} \) in \([15]\) as a vector \( Z \in \mathbb{R}^{nM} \) (with entries \( Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \ldots, nM \)) and for the last inequality we used
Mill’s ratio. As the number of $\Lambda$-separable selection matrices $\Pi$ is bounded by $nmk^m$, it follows that
\[
P_{\Pi A\omega} \left( \hat{\Pi} \neq \Pi \right) \leq 2\sigma nk^m m(1 + ma_k) \left( \frac{1 + mM}{\sqrt{A\delta M}} \right)^{3/2} \left( \frac{\Lambda^2 M^2}{n(mM + 1)(1 + ma_k)^{3/2}} \right) e^{-\frac{\sqrt{\Lambda\delta} \varepsilon}{\sqrt{n(mM + 1)(1 + ma_k)^{3/2}}}}.
\]
and
\[
E_{\Pi A\omega} \left( \|\hat{\theta} - \Pi A\omega\|_2^2 I_{\{\hat{\Pi} \neq \Pi\}} \right) \leq 12\sigma^2 n^2 k^m m^2 \left( \frac{M^3}{2A\delta} \right) \left( \frac{\Lambda^2 M^2}{n(mM + 1)(1 + ma_k)^{3/2}} \right) e^{-\frac{\sqrt{\Lambda\delta} \varepsilon}{\sqrt{n(mM + 1)(1 + ma_k)^{3/2}}}}.
\]
This gives the second term of the r.h.s. of the assertion.

Now we consider the first summand on the r.h.s. of (59). The idea is to bound the minimax risk conditioned on $\hat{\Pi} = \Pi$ with the minimax risk of the LSE on the linear subvector space $\text{imag}(\Pi A)$. To this end, let $N^\delta(\Pi) \subset N^\delta(\Pi)$ denote the set of all $\Pi A\omega \in N^\delta(\Pi)$ with $\Pi = \Pi$. Further, let $\hat{\theta}' \in \text{argmin}_{\Pi A\omega \in N^\delta(\Pi)} \|Y - \Pi A\omega\|^2$ be the least-squares estimator restricted to $\hat{\Pi} = \Pi$. Then, clearly, $\hat{\theta} = \hat{\theta}'$ on $\{\hat{\Pi} = \Pi\}$ and thus
\[
E_{\Pi A\omega} \left( \|\hat{\theta} - \Pi A\omega\|_2^2 I_{\{\hat{\Pi} = \Pi\}} \right) = E_{\Pi A\omega} \left( \|\hat{\theta}' - \Pi A\omega\|_2^2 I_{\{\hat{\Pi} = \Pi\}} \right) \quad \leq E_{\Pi A\omega} \left( \|\hat{\theta}' - \Pi A\omega\|_2^2 \right).
\]
Thus, for all fixed $\Pi$ as in (16)
\[
\sup_{\Pi A\omega \in N^\delta(\Pi)} E_{\Pi A\omega} \left( \|\hat{\theta}' - \Pi A\omega\|_2^2 I_{\{\hat{\Pi} = \Pi\}} \right) \quad \leq \quad \sup_{\Pi A\omega \in N^\delta(\Pi)} E_{\Pi A\omega} \left( \|\hat{\theta}' - \Pi A\omega\|_2^2 \right).
\]
Clearly, $N^\delta(\Pi) \subset \text{imag}(\Pi A)^M$ with $\text{dim}(\text{imag}(\Pi A)^M) = mM$. Thus, for the LS estimator on $\text{imag}(\Pi A)^M$, $\hat{\theta}'' \in \text{argmin}_{\theta \in \text{imag}(\Pi A)^M} \|Y - \hat{\theta}\|^2$, it follows from Lemma A.8 that
\[
\sup_{\Pi A\omega \in N^\delta(\Pi)} E_{\Pi A\omega} \left( \|\hat{\theta}' - \Pi A\omega\|_2^2 \right) \quad \leq \quad 4 \quad \sup_{\theta \in \text{imag}(\Pi A)^M} E_{\theta} \left( \|\hat{\theta}'' - \theta\|_2^2 \right) = 4\sigma^2 mM,
\]
which finishes the proof.}

### A.5 Proofs for results on the estimation error in Section 4

**Proof of Theorem 4.2** The first assertion follows directly from the first part of Lemma A.9 with $\epsilon = \|\Pi A\omega - \Pi' A\omega\|/(\sqrt{ma_k})$. The second assertion follows from the second part of Lemma A.9 with $\epsilon \not\in \|\Pi A\omega - \Pi' A\omega\|$. 

28
Proof of Theorem 4.3. First, we show the lower bound. It follows directly from combining the first part of Theorem 4.2, Corollary 3.7, and (22) that

\[ \inf_{\hat{\Pi}, \hat{\omega}} \sup_{\Pi, \omega} E_{\Pi, \omega} \left( d \left( (\Pi, \omega), (\hat{\Pi}, \hat{\omega}) \right)^2 \right) \geq \sigma^2 M \frac{1}{m^2 a_k^2} + \sigma \sqrt{M} \frac{\delta}{m^2 a_k^2} e^{-\frac{\sigma^2}{M} \frac{m}{\sigma^2}}. \tag{60} \]

Moreover, let \( \Pi^* \) be such that

\[ \Pi^* A = \begin{pmatrix} I_{m \times m} \ldots, I_{m \times m} \\ 0_{m \times m} \ldots 0_{m \times m} \end{pmatrix}, \]

where \( I_{m \times m} \) is the \( m \times m \) identity matrix and \( 0_{m \times m} \) is the \( m \times m \) zero matrix.

Let \( \hat{\Theta}' \) be the hyperractangle as in (55) and \( \hat{\Theta} \) as in (54) its embedding in \( \Omega_{m,M} \). Then, as been shown in the proof of Theorem 3.1

\[ \{ \Pi^* A\omega : \omega \in \hat{\Theta} \} \subset N^a, \Lambda. \]

For \( \Pi A\omega = \Pi^* A\omega \) one observes in (15)

\[ Y_1, \ldots, Y_{M \Lambda} \overset{i.i.d.}{\sim} N(\omega, \sigma^2 I_{M m \times M m}). \]

With this we mean that \( Y_i \) is the \( m \times M \) sub-matrix of \( Y \) in (15) which consists of \( m \) successive row vectors of \( Y \).

Thus, one gets with \( \epsilon^* \) as in (53) that

\[
\begin{align*}
\inf_{\hat{\Pi}, \hat{\omega}} \sup_{\Pi, \omega} E_{\Pi, \omega} \left( \frac{1}{M} d \left( (\Pi, \omega), (\hat{\Pi}, \hat{\omega}) \right)^2 \right) \\
\geq \inf_{\hat{\omega} \in \hat{\Theta}} \sup_{\omega \in \Theta} E_{\Pi^*, \omega} \left( \frac{1}{M} d \left( (\Pi^*, \omega), (\Pi^*, \hat{\omega}) \right)^2 \right) \\
= \inf_{\hat{\omega} \in \hat{\Theta}} \sup_{\omega \in \Theta} E_{\Pi^*, \omega} \left( \frac{1}{M} \max_{i=1,\ldots,m} \| \hat{\omega}_i - \omega_i \|^2 \right) \\
\geq \inf_{\hat{\omega} \in \hat{\Theta}} \sup_{\omega \in \Theta} E_{\Pi^*, \omega} \left( \frac{1}{M m} \| \hat{\omega} - \omega \|^2 \right) \\
\geq \frac{1}{M m} M(m - 1) \frac{(\epsilon^*)^2 \sigma^2 / (\Lambda M)}{(\epsilon^*)^2 + \sigma^2 / (\Lambda M)} \\
\geq \frac{\sigma^2}{\Lambda M},
\end{align*}
\]

where the second last inequality follows from [10] (2.1), (3.4) and Proposition 3] and for the last inequality we used that by (22) \( M \geq \sigma 6 \sqrt{2 m^{5/2} a_k / (\Delta \lambda_{\min \Lambda})} \) and thus \( \sigma^2 / (\Lambda M) \leq (\epsilon^*)^2 \). Together with (60) this yields

\[ \inf_{\hat{\Pi}, \hat{\omega}} \sup_{\Pi, \omega} E_{\Pi, \omega} \left( \frac{1}{M} d \left( (\Pi, \omega), (\hat{\Pi}, \hat{\omega}) \right)^2 \right) \geq \frac{\sigma^2}{\Lambda M} + \sigma \frac{\delta}{m^2 a_k^2 \sqrt{M}} e^{-\frac{\sigma^2}{M} \frac{m}{\sigma^2}}. \tag{61} \]
Second, we show the upper bound. Let $\hat{\theta} = \hat{\Pi}A\hat{\omega}$ be the LSE in (20). By Corollary 4.1 it follows that

$$\sup_{\Pi, \omega} E_{\Pi, \omega} \left( d \left( (\Pi, \omega), (\hat{\Pi}, \hat{\omega}) \right)^2 | \hat{\Pi} = \Pi \right) \leq \sup_{\Pi, \omega} E_{\Pi, \omega} \left( d \left( (\Pi, \omega), (\hat{\Pi}, \hat{\omega}) \right)^2 I_{\hat{\Pi} = \Pi} \right) + \sup_{\Pi, \omega} E_{\Pi, \omega} \left( d \left( (\Pi, \omega), (\hat{\Pi}, \hat{\omega}) \right)^2 I_{\hat{\Pi} \neq \Pi} \right) \leq \sup_{\Pi, \omega} E_{\Pi, \omega} \left( \max_{i=1, ..., m} \| \omega_i - \hat{\omega}_i \|^2 I_{\hat{\Pi} = \Pi} \right) + \sigma \sqrt{\Lambda} e^{-\frac{\Lambda^2}{16m^2(1+a_k)^2}} \frac{M}{\sigma^2}.$$

As argued in the proof of Theorem 3.5, conditioned on $\hat{\Pi} = \Pi$, the MSE of the LSE $\hat{\theta}$ is bounded (up to a constant) by the MSE of the ordinary LSE (for given design matrix $F = \Pi A$), which is known to be minimax optimal among linear unbiased estimators. Further, by the separability condition (A-IC 2) it follows that each of the unit vectors $e_i \in \mathbb{R}^m$ for all $i = 1, ..., m$ appears at least $\Lambda M$ times in $\Pi A$. Thus, only considering those rows in $Y$ which correspond to the $\Lambda M$ unit vectors $e_i$ for all $i = 1, ..., m$, that is,

$$Y_1, ..., Y_{\Lambda M} \overset{i.i.d.}{\sim} N(\omega, \sigma^2 I_{Mm \times Mm})$$

(again, this means that $Y_i$ is the $m \times M$ sub-matrix of $Y$ in (15) corresponding to the $m$ unit vectors $e_1, ..., e_m$ in $\Pi A = F$), increases the MSE and hence,

$$\sup_{\Pi, \omega} E_{\Pi, \omega} \left( \max_{i=1, ..., m} \| \omega_i - \hat{\omega}_i \|^2 I_{\hat{\Pi} = \Pi} \right) \leq \sup_{\Pi, \omega} E_{\Pi, \omega} \left( \| \omega - \hat{\omega} \|^2 I_{\hat{\Pi} = \Pi} \right) \lesssim \frac{\sigma^2 m}{\Lambda}.$$

Summing up, we get

$$\sup_{\Pi, \omega} E_{\Pi, \omega} \left( | d \left( (\Pi, \omega), (\hat{\Pi}, \hat{\omega}) \right)|^2 \right) \lesssim \frac{\sigma^2 m}{\Lambda M} + \frac{\sigma}{\sqrt{\Lambda} \sqrt{M}} e^{-\frac{\Lambda^2}{16m^2(1+a_k)^2}} \frac{M}{\sigma^2},$$

which finishes the proof.

### A.6 Additional lemmas

**Lemma A.7.**

For a finite parameter space $\Theta \subset \mathbb{R}^n$ and any estimator $\hat{\theta}$

$$\min_{\theta' \neq \theta''} \| \theta' - \theta'' \|^2 \leq \frac{\sup_{\theta \in \Theta} E_{\theta} \left( \| \hat{\theta} - \theta \|^2 \right)}{\sup_{\theta \in \Theta} P_{\theta} \left( \hat{\theta} \neq \theta \right)} \leq \max_{\theta' \neq \theta''} \| \theta' - \theta'' \|^2.$$
Proof. It holds that
\[
\sup_{\theta \in \Theta} \mathbb{E}_\theta \left( \|\hat{\theta} - \theta\|^2 \right) = \sup_{\theta \in \Theta} \sum_{\theta \in \Theta \setminus \theta} \|\hat{\theta} - \theta\|^2 P_{\theta} \left( \hat{\theta} = \hat{\theta} \right)
\]
where the last inequality follows from the definition of \( \hat{\theta} \).

Further, let \( \hat{\theta}(\theta) = \frac{1}{\|\theta\|^2} \sum_{\theta \in \Theta \setminus \theta} \|\hat{\theta} - \theta\|^2 P_{\theta} \left( \hat{\theta} = \hat{\theta} \right) \).

Proof. Let \( \theta \in \Theta \) be fixed. If \( \|\hat{\theta} - \theta\|^2 = 0 \), (62) holds trivially. Further, if \( \theta \in \Theta \), it holds that \( \hat{\theta} = \hat{\theta} = Y \), and hence, (62) follows trivially, too. So assume that \( \|\hat{\theta} - \theta\|^2 > 0 \) and \( \theta \notin A \).

Choosing an appropriate coordinate system, we may w.l.o.g. assume that
\[
V = \{x \in \mathbb{R}^d : x_1 = \ldots = x_r = 0\},
\]
with \( \text{dim}(V) = d - r \). Let \( \text{pr} \) be the orthogonal projection onto \( V \), i.e.,
\[
\text{pr} : (x_1, \ldots, x_d)^\top \mapsto (x_{r+1}, \ldots, x_d)^\top.
\]

Then,
\[
\arg\min_{\theta \in A} \|Y - \tilde{\theta}\|^2 = \arg\min_{\tilde{\theta} \in \bar{A}} \sum_{i=1}^r Y_i^2 + \|\text{pr}(Y) - \text{pr}(\tilde{\theta})\|^2
\]
and analog \( \arg\min_{\theta \in V} \|Y - \tilde{\theta}\|^2 = \arg\min_{\tilde{\theta} \in \bar{V}} \|\text{pr}(Y) - \text{pr}(\tilde{\theta})\|^2 \).

Thus, we may w.l.o.g. assume that \( V = \mathbb{R}^d \), i.e., \( \theta_v = Y \). Then
\[
\frac{\|Y - \theta\|^2}{\|\hat{\theta} - \theta\|^2} \geq \left( \min_{x \notin A} \frac{\|x - \theta\|}{\|\hat{\theta}(x) - \theta\|} \right)^2 \geq \left( \min_{x \notin A} \frac{\|x - \theta\|}{\|\hat{\theta}(x) - x\| + \|x - \theta\|} \right)^2
\]
\[
= \left( 1 + \max_{x \notin A} \frac{\|x - \hat{\theta}(x)\|}{\|x - \theta\|} \right)^{-2} \geq \frac{1}{4},
\]
where the last inequality follows from the definition of \( \hat{\theta}_A \).
Lemma A.9. 
Let $\Pi A\omega, \Pi' A\omega' \in N^k$, then for all $\epsilon > 0$

1. 
$$
\|\Pi A\omega - \Pi' A\omega'\| \geq \sqrt{n} m a_k \epsilon \implies d(\Pi, \omega, \Pi', \omega') \geq \epsilon,
$$

2. If $\epsilon < \delta \sqrt{M} (1 + \max_k)$, then
$$
\|\Pi A\omega - \Pi' A\omega'\| < \epsilon \implies d(\Pi, \omega, \Pi', \omega') < \epsilon.
$$

Proof. From $\|\Pi A\omega - \Pi' A\omega'\| \geq \sqrt{n} m a_k \epsilon$ it follows that
$$
\max_{j=1, \ldots, n} \| (\Pi A\omega)_j - (\Pi' A\omega')_j \| \geq m a_k \epsilon
$$
and
$$
\epsilon \leq \max_{j=1, \ldots, n} \| (\Pi A\omega)_j - (\Pi' A\omega')_j \| / (m a_k) \leq \sqrt{M}.
$$
Hence, by Remark 2.3, $\max_{i=1, \ldots, m} \| \omega_i - \omega'_i \| \geq \epsilon$ or $\Pi \neq \Pi'$. Thus, it follows that $d(\Pi, \omega, \Pi', \omega') \geq \epsilon$, which shows the first assertion. If $\epsilon < \delta \sqrt{M} (1 + \max_k)$ and $\max_{j=1, \ldots, n} \| (\Pi A\omega)_j - (\Pi' A\omega')_j \| \leq \|\Pi A\omega - \Pi' A\omega'\| < \epsilon$ it follows from Theorem 2.2 that $\max_{i=1, \ldots, m} \| \omega_i - \omega'_i \| < \epsilon$ and $\Pi = \Pi'$ and thus it follows that $d(\Pi, \omega, \Pi', \omega') < \epsilon$, which shows the second assertion.

Acknowledgements

Merle Behr acknowledges support of RTG 2088 and CRC 803 Z2 (German Research Foundation). Axel Munk acknowledges support of CRC 803 Z2 (German Research Foundation). Helpful comments of Martin Wainwright, Boaz Nadler, and Burkhard Blobel are gratefully acknowledged.

References

[1] Arora, S., Ge, R., Kannan, R., and Moitra, A. (2012). Computing a nonnegative matrix factorization—provably. In Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing, pages 145–162. New York.

[2] Behr, M. (2018). Finite Alphabet Blind Separation. Ph.D. thesis, University of Göttingen.

[3] Behr, M., Holmes, C., and Munk, A. (2018). Multiscale blind source separation. The Annals of Statistics, 46(2):711–744.

[4] Behr, M. and Munk, A. (2017). Identifiability for blind source separation of multiple finite alphabet linear mixtures. IEEE Transactions on Information Theory, 63(9):5506–5517.
[5] Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Dono-
van, J., et al. (2010). The landscape of somatic copy-number alteration across
human cancers. *Nature*, 463(7283):899–905.

[6] Bing, X., Bunea, F., and Wegkamp, M. (2018). A fast algorithm with mini-
max optimal guarantees for topic models with an unknown number of topics.
*arXiv:1805.06837 [cs, stat]*.

[7] Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T.,
et al. (2012). Absolute quantification of somatic DNA alterations in human
cancer. *Nature Biotechnology*, 30(5):413–421.

[8] Comon, P. (1994). Independent component analysis, a new concept? *Signal
Processing*, 36:287–314.

[9] Diamantaras, K. I. (2006). A clustering approach for the blind separation
of multiple finite alphabet sequences from a single linear mixture. *Signal
Processing*, 86(4):877–891.

[10] Donoho, D., Liu, R. C., and MacGibbon, B. (1990). Minimax risk over hy-
perrectangles and implications. *The Annals of Statistics*, 18(3):1416–1437.

[11] Donoho, D. and Stodden, V. (2004). When does non-negative matrix fac-
torization give a correct decomposition into parts? In *Advances in Neural
Information Processing Systems (NIPS) 16*. MIT Press, Cambridge.

[12] Flammarion, N., Mao, C., and Rigollet, P. (2016). Optimal rates of statistical
seriation. *arXiv preprint arXiv:1607.02435*.

[13] Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature*,
481(7381):306–313.

[14] Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., et al. (2014).
TITAN: Inference of copy number architectures in clonal cell populations from
tumor whole-genome sequence data. *Genome Research*, 24(11):1881–1893.

[15] Jiang, Y., Qiu, Y., Minn, A. J., and Zhang, N. R. (2016). Assessing intratu-
mer heterogeneity and tracking longitudinal and spatial clonal evolutionary
history by next-generation sequencing. *Proceedings of the National Academy
of Sciences*, 113(37):E5528–E5537.

[16] Ke, Z. T. and Wang, M. (2017). A new SVD approach to optimal topic
estimation. *arXiv:1704.07016 [stat]*.

[17] Lee, D. and Seung, S. (1999). Learning the parts of objects by non-negative
matrix factorization. *Nature*, 401:788–791.

[18] Li, T. (2005). A general model for clustering binary data. In *Proceedings of the
Eleventh ACM SIGKDD International Conference on Knowledge Discovery
in Data Mining*, pages 188–197. Chicago.
[19] Liu, B., Morrison, C. D., Johnson, C. S., Trump, D. L., Qin, M., Conroy, J. C., et al. (2013). Computational methods for detecting copy number variations in cancer genome using next generation sequencing: Principles and challenges. *Oncotarget*, 4(11):1868.

[20] Love, D. J., Heath, R. W., Lau, V. K., Gesbert, D., Rao, B. D., and Andrews, M. (2008). An overview of limited feedback in wireless communication systems. *IEEE Journal on Selected Areas in Communications*, 26(8):1341–1365.

[21] Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2010). Online Learning for Matrix Factorization and Sparse Coding. *Journal of Machine Learning Research*, 11:19–60.

[22] Marques, M., Stošić, M., and Costeira, J. (2009). Subspace matching: Unique solution to point matching with geometric constraints. In *IEEE Twelfth International Conference on Computer Vision*, pages 1288–1294. Kyoto.

[23] Pajunen, P. (1997). Blind separation of binary sources with less sensors than sources. In *International Conference on Neural Networks*, pages 1994–1997. Houston.

[24] Pananjady, A., Wainwright, M. J., and Courtade, T. A. (2016). Linear regression with an unknown permutation: Statistical and computational limits. In *54th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, Monticello, IL.

[25] Pananjady, A., Wainwright, M. J., and Courtade, T. A. (2017). Denoising linear models with permuted data. In *IEEE International Symposium on Information Theory*, pages 446–450. Aachen, Germany.

[26] Proakis, J. G. (2007). *Digital Communications*. McGraw-Hill series in electrical and computer engineering. McGraw-Hill, Boston.

[27] Rubinstein, R., Bruckstein, A. M., and Elad, M. (2010). Dictionaries for Sparse Representation Modeling. *Proceedings of the IEEE*, 98(6):1045–1057.

[28] Sampath, H., Stoica, P., and Paulraj, A. (2001). Generalized linear precoder and decoder design for MIMO channels using the weighted MMSE criterion. *IEEE Transactions on Communications*, 49(12):2198–2206.

[29] Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403):395–399.

[30] Talwar, S., Viberg, M., and Paulraj, A. (1996). Blind separation of synchronous co-channel digital signals using an antenna array. I. Algorithms. *IEEE Transactions on Signal Processing*, 44(5):1184–1197.

[31] Unnikrishnan, J., Haghighatshoar, S., and Vetterli, M. (2015). Unlabeled sensing with random linear measurements. *arXiv preprint arXiv:1512.00115*. 34
[32] Van den Meersche, K., Soetaert, K., and Van Oevelen, D. (2009). Xsample(): An R function for sampling linear inverse problems. *Journal of Statistical Software*, 30:1–15.

[33] Van der Veen, A.-J. (1997). Analytical method for blind binary signal separation. *IEEE Transactions on Signal Processing*, 45(4):1078–1082.

[34] Verdu, S. (1998). *Multiuser Detection*. Cambridge University Press, Cambridge.

[35] Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):37–57.

[36] Zhang, Y. and Kassam, S. A. (2001). Blind separation and equalization using fractional sampling of digital communications signals. *Signal Processing*, 81(12):2591–2608.