Inverse Problems

PAPER

A Bayesian framework for molecular strain identification from mixed diagnostic samples

To cite this article: Lauri Mustonen *et al* 2018 *Inverse Problems* **34** 105009

View the article online for updates and enhancements.

You may also like

- Bayesian uncertainty analysis compared with the application of the GUM and its supplements
Clemens Elster
- A Bayesian interpretation of the L-curve
Jérôme Antoni, Jérôme Idier and Sébastien Bourguignon
- On the evaluation of uncertainties for state estimation with the Kalman filter
S Eichstädt, N Makarava and C Elster

This content was downloaded from IP address 18.118.32.213 on 25/04/2024 at 12:39

Inverse Problems **34** (2018) 105009 (22pp) https://doi.org/10.1088/1361-6420/aad7cd

A Bayesian framework for molecular strain identification from mixed diagnostic samples

Lauri Mustonen[1,5], **Xiangxi Gao**[1], **Asteroide Santana**[2], **Rebecca Mitchell**[1,3], **Ymir Vigfusson**[1,4] and **Lars Ruthotto**[1]

[1] Department of Mathematics and Computer Science, Emory University, 400 Dowman Drive, Atlanta, GA 30322, United States of America
[2] School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive, Atlanta, GA 30318, United States of America
[3] Nell Hodgson Woodruff School of Nursing, Emory University, 1520 Clifton Road, Atlanta, GA 30322, United States of America
[4] School of Computer Science, Reykjavik University, Menntavegi 1, Reykjavik, 101, Iceland

E-mail: lauri.mustonen@emory.edu

Received 7 March 2018, revised 7 July 2018
Accepted for publication 3 August 2018
Published 21 August 2018

Abstract

We provide a mathematical formulation and develop a computational framework for identifying multiple strains of microorganisms from mixed samples of DNA. Our method is applicable in public health domains where efficient identification of pathogens is paramount, e.g. for the monitoring of disease outbreaks. We formulate strain identification as an inverse problem that aims at simultaneously estimating a binary matrix (encoding presence or absence of mutations in each strain) and a real-valued vector (representing the mixture of strains) such that their product is approximately equal to the measured data vector. The problem at hand has a similar structure to blind deconvolution, except for the presence of binary constraints, which we enforce in our approach. Following a Bayesian approach, we derive a posterior density. We present two computational methods for solving the non-convex maximum *a posteriori* estimation problem. The first one is a local optimization method that is made efficient and scalable by decoupling the problem into smaller independent subproblems, whereas the second one yields a global minimizer by converting the problem into a convex mixed-integer quadratic programming problem. The decoupling approach also provides an efficient way to integrate over the posterior. This provides useful information about the ambiguity of the underdetermined problem and, thus,

[5] Author to whom any correspondence should be addressed.

1361-6420/18/105009+22$33.00 © 2018 IOP Publishing Ltd Printed in the UK

the uncertainty associated with numerical solutions. We evaluate the potential and limitations of our framework *in silico* using synthetic and experimental data with available ground truths.

Keywords: Bayesian inverse problems, mixed-integer programming, molecular biology

(Some figures may appear in colour only in the online journal)

1. Introduction

Many public health programs, such as epidemiological surveillance, rely crucially on taking and processing biological samples to gather information. Diagnostic samples can often contain multiple genetic variants, so-called *strains*, of the same microorganism (e.g. bacteria, viruses, or parasites) resulting from mutations and adaptation. For instance, blood samples of malaria patients may exhibit multiple concurrent strains of malaria contracted from bites of several parasite-infected mosquitoes or even a single bite of a mosquito carrying multiple genetic variants of the parasite [1]. Different strains of pathogens can exhibit different characteristics that directly impact human health, potentially affecting the severity of illness, contagiousness, and resistance to classes of drugs [2]. Thus, accurate identification of strains within diagnostic samples is critical. Epidemiological applications of strain identification include control effort evaluation, such as malaria reduction programs where different strains respond to different interventions, and tracing of outbreaks for pathogens that are similar to benign microorganisms, such as commensal (harmless) bacteria, or for which hosts may carry multiple pathogenic strains simultaneously, such as *E. coli* bacteria.

Unfortunately, strain identification of mixed diagnostic samples—samples with multiple strains of a pathogen—remains particularly vexing. The state-of-the-art approaches, detailed in the following section, have shortcomings that limit their applicability to these public health challenges. Culture-based approaches are resource and labor intensive to be deployed at scale; metagenomic approaches thus far lack sufficient discriminatory power within a species, and direct polymerase chain reaction (PCR) diagnoses do not provide a sufficient strain-level resolution for epidemiological outbreak investigations or national surveillance programs. A recent method of Zhu *et al* bypasses these problems, but critically depends on perfect knowledge of possible strain types—a dictionary—that is unrealistic to assume or generate for epidemiological surveillance in the field [3].

There are two parts to overcome the strain reconstruction challenge for mixed samples: defining an appropriate laboratory procedure for converting the sample into a data vector and designing algorithms for disambiguating the pathogen strains from that data vector. On the laboratory side, our approach uses fast, affordable and widely available biological diagnostic tools, specifically a combination of DNA barcoding [4] and whole genome multilocus sequence typing (wgMLST) [5–7], to produce short-read amplifications of independent genomic targets in the mixed sample. For each mixed sample, the deep sequencing platform in the procedure produces a measurement vector of locus-by-locus frequency information, denoting the percentage of mutations at every target location, or single-nucleotide polymorphism (SNP) site, in the amplified DNA sequences from the sample.

In this work, we tackle the algorithmic side of the strain reconstruction problem using this measurement vector as the only input. Specifically, we derive a mathematical formulation and computational framework for identifying distinct strains of microorganisms as well as their proportions from mixed diagnostic samples. From the measurement data vector, we infer the identity and frequency of the strains with a Bayesian inverse problem approach. We derive an expression for the posterior distribution and present numerical methods for computing *maximum a posteriori* (MAP) estimates and integrals over the posterior distribution (section 3). The problem is underdetermined, and thus we give particular emphasis to quantifying uncertainty in the reconstruction. We evaluate the potential and limitations of our methods *in silico* using numerical examples on both simulated and experimental data (malaria strains) with available ground truths (section 4). Our results suggest that strain sequences can be reconstructed from mixed samples with moderate to high fidelity for a range of input vectors, with important caveats that are influenced, e.g. by the input measurement errors and the true strain frequencies.

2. Background

Our approach to strain reconstruction depends on interpreting output from a particular biological pipeline that is applied to a sample. Before we formalize our algorithms that depend on these outputs, we provide some context about their possible application and a brief description of and references for possible experimental pipelines.

Traditionally, biological samples were cultured in the laboratory to isolate a single microorganism. The isolate was then cultured to obtain a single strain sample of the microorganism, which could be analyzed [6]. If a sample was expected to have multiple strains, then multiple subcultures were processed to obtain samples of each strain. However, not all microorganisms are amenable to artificial culture. Moreover, some subtypes often grow better than others, resulting in an unequal representation of the true subtypes in the sample [8]. The number of subtypes detected can also be highly influenced by the number of subsamples cultured [9]. This relationship can be seen, for example, in samples with low concentrations of minor strains, such as drug-resistant bacteria that respond poorer to the culturing process (lower fitness) than wild-type bacteria. It is also evident in mixed samples with a high diversity of strains, such as samples of the *P. falciparum* malaria parasite with five or more strains in a single sample.

Recently developed PCR-based DNA amplification techniques allow one to diagnose pathogens directly from original samples, alleviating the need of isolating individual strains in culture [2, 6, 7, 10, 11]. These PCR-based techniques are not only considerably faster than culture-based diagnosis, but also less expensive. Furthermore, Langley *et al* have shown that PCR-based, culture-independent, diagnostic tests can be more sensitive and are easier to perform than traditional, culture-based approaches [12].

Despite the benefits of the PCR techniques, direct PCR-based diagnosis lacks information on strains at the level necessary for some important epidemiological studies, such as outbreak investigations or national surveillance programs. These applications require more detailed microorganism resolution [2, 7] than provided by PCR-based diagnoses, which commonly focus on the clinically relevant species- or serotype-level identification.

To gain more information about strains, metagenomic approaches that evaluate all the DNA in a sample and screen for microorganisms of interest are being used to distinguish multiple genetic variants. These techniques, however, generally require ample depth of coverage (copies of genomic area of interest) across large sections of the genome, or long, linked reads to provide sufficient discrimination of strains for our target epidemiological applications [13].

To uniquely identify the targeted microorganisms within samples, our approach is instead to use a combination of DNA barcoding [4] and wgMLST diagnostic tools [4–7] that rely on coverage of specific, often unlinked, genomic targets to increase discriminatory power relative to metagenomic screening. An observation recently published by Zhu *et al* [3], which we made concurrently and independently in our project, is that the locus-by-locus frequency information provided by the deep sequencing programs of the pipeline can be used to distinguish between strains in mixed samples. We focus on the analysis of the data created by these targeted and unlinked short-read diagnostic techniques for the remainder of this paper.

3. Methods

Motivated by practical lab pipelines for generating measurement data from mixed diagnostic samples, we now detail the algorithmic part of the strain reconstruction problem: translating a vector of mutation frequencies at each location into strain genomic sequences and the relative proportions of each strain within the sample. We begin by specifying the problem, notation, and assumptions before deriving a Bayesian formulation of the ensuing inverse problem. Following recommended guidelines [14, 15], we model all quantities in the forward model as random variables, design prior distributions to incorporate prior knowledge, and use Bayes' formula to obtain a posterior distribution. We will explore the posterior in three ways.

(i) We address MAP estimation as a non-convex mixed-integer optimization problem and present an efficient local optimization method based on block coordinate descent to compute local modes of the posterior. The algorithm exploits the separability of the objective function and provides both deterministic running time and scalability with respect to the number of unknowns.

(ii) We reformulate the MAP estimation problem as a convex binary constrained problem and present a method based on existing algorithms for obtaining the largest posterior mode. Specifically, the method computes a *global* minimizer and, unlike the first method, yields a certificate of optimality. While we observed that the running time can vary significantly between inputs, the method enables one to validate and calibrate the often more efficient local optimization method.

(iii) We develop a computationally tractable method to quantify the uncertainty in the reconstructed strains and their frequencies. In particular, we propose an efficient integration technique for the posterior density that leverages the separability in the structure of the posterior.

3.1. Problem specification

Assume at first that the number of strains, denoted by n, is known, and consider measurements at m locations in a DNA sequence. At each location and for each strain, a mutation (relative to a reference strain) is either present, which can be encoded as 1, or absent, which corresponds to 0. This binary vector is called a molecular barcode in the biological literature [4]. Now let $\mathbf{d} \in \mathbb{R}^m$ denote the actual measurement data which, for this example, represents the percentage of mutations at m defined SNP sites in the DNA sequence. If $\mathbf{w} \in \mathbb{R}^n$ is a vector containing the relative frequency of each strain and $\mathbf{M} \in \{0, 1\}^{m\times n}$ is a binary matrix encoding the presence and absence of mutations, the forward problem can be written as

$$\mathbf{d} = \mathbf{M}\mathbf{w} + \mathbf{n}, \tag{1}$$

where $\mathbf{n} \in \mathbb{R}^m$ represents the inevitable measurement noise. The goal of the inverse problem is to infer both $\mathbf{M}$ and $\mathbf{w}$ from the measurement data $\mathbf{d}$. In other words, we aim to identify the strains and their respective proportions in a given sample.

Even in the absence of noise, (1) corresponds to an underdetermined and ill-posed inverse problem: There are $mn + n$ unknowns but only m knowns. For example, the problem is invariant to permutations of the columns of $\mathbf{M}$ and rows in $\mathbf{w}$. On the other hand, with noisy measurements, the equation $\mathbf{Mw} = \mathbf{d}$ does not, in general, have a solution that satisfies the prescribed binary constraint for $\mathbf{M}$.

3.1.1. Multiplicity of infection. In practical applications, the number of strains n—often referred to as the multiplicity of infection (MOI)—is usually unknown and often difficult to estimate [1]. We will discuss below how to include the estimation for n in the inverse problem.

3.1.2. Related problems. The inverse problem corresponding to (1) arises also in signal processing and wireless communications when one tries to reconstruct binary source signals from a linear mixture that is formed with unknown mixing weights [16, 17]. More generally, in *blind source separation* the aim is to recover the original (not necessarily binary) signals and a mixing matrix when multiple linear combinations are observed [18–20]. Having a matrix measurement also leads to a *non-negative matrix factorization* problem [21], which can be equipped with binary constraints as well [22]. Notice that although the number of unknowns increases if $\mathbf{w}$ is replaced with a matrix, the number of knowns increases as well and the problem actually becomes less underdetermined compared to our case. The inverse problem corresponding to (1) bears also similarities with blind deconvolution [23–25] in the sense that both the linear operator $\mathbf{M}$ (compare to blurring operator) and the vector (compare to image) are unknown. A common property for all abovementioned problems is that they are *bilinear*, which for (1) means being linear in $\mathbf{M}$ for a fixed $\mathbf{w}$ and linear in $\mathbf{w}$ for a fixed $\mathbf{M}$. On the other hand, linear inverse problems in which binary and continuous variables are present arise in, e.g. groundwater flow [26].

If the noise $\mathbf{n}$ in (1) is zero (or small enough) and if the matrix $\mathbf{M}$ contains sufficiently many distinct rows, then in most cases the vector $\mathbf{w}$ can be easily solved by sorting the values in the measurement $\mathbf{d}$ and assigning them to $\mathbf{w}$ while discarding values that are binary combinations of already assigned values [16, 17, 20]. However, in our case, there is no guarantee that the matrix contains enough distinct rows for this approach to work. Our measurements also contain noise, which is why we take the Bayesian approach for the inverse problem.

3.2. Generalization to multiple classes

We start by considering a generalization of (1) where each site in a strain can represent more than two classes of measured values (mutation versus no mutation in the binary case above), allowing modeling of SNPs that have multiple alternative options, insertions or deletions of genomic content, or multiple linked differences within a target location, to name a few examples. One immediate use of this generalization is to obtain more detailed strain reconstruction consisting of all four nucleotides found in DNA molecules: adenine, guanine, cytosine, and thymine. Even with multiple classes, we formulate the generalized problem by using a binary matrix and real-valued vectors, and the computational methods will thus remain similar to the binary case described above.

As before, let m denote the number of SNP sites (i.e. measurement locations) and n the number of strains, and let $p \geqslant 2$ now denote the number of classes each strain can be associated with at each measurement location. The measured data represents the frequencies of the classes at each location. Because the frequencies sum up to one, the measurement can be represented using $q := m(p-1)$ elements.

The measurement $\mathbf{d}$ is interpreted as a block vector having m blocks of size $(p-1)$. Within each block, the first value corresponds to the frequency of the second class, the second value to that of the third class, and so on. In this way, the pure binary case $p=2$ will be handled naturally, and in general, the frequency of the first class is just one minus the sum of the frequencies of other classes. The matrix $\mathbf{M}$ defining the strains now becomes a binary matrix with m blocks of size $(p-1)\times n$. One can think of replacing the zeros and ones of the pure binary case with vectors $\mathbf{0}\in\mathbb{R}^{p-1}$ and Euclidean $(p-1)$-dimensional unit vectors, respectively. As an example, if $p=4$, then a strain is characterized at one measurement location by either $(0,0,0)^\top$, $(1,0,0)^\top$, $(0,1,0)^\top$, or $(0,0,1)^\top$.

3.3. Deriving the posterior

The observations taken from the sample can be written as a bilinear forward problem

$$D = MW + N,$$

where the measurement, D, is a random variable of length q, M is a random variable of size $q\times n$ with binary entries whose columns represent the different strains, and W is a random variable of length n with real entries between 0 and 1 that model the proportions in which different strains are present in the sample. For each realization $\mathbf{w}$ of W we therefore know that $\sum_{j=1}^{n}\mathbf{w}_j = 1$.

In this work, we assume that the additive noise, N, is a multivariate Gaussian random variable with vanishing mean $\mathbf{0}\in\mathbb{R}^q$ and a known diagonal covariance matrix $\Gamma = \mathrm{diag}(\gamma_1^2,\gamma_2^2,\dots,\gamma_q^2)\in\mathbb{R}^{q\times q}$. In other words, $\mathbf{n}$ in (1) is a realization of $N\sim\mathcal{N}(\mathbf{0},\Gamma)$. This assumption is motivated by its simplicity but can also be justified when the data features a relatively high signal to noise ratio. Several other noise models are possible—note that data is obtained by a counting process—and will be investigated in future work. With Gaussian noise, the *likelihood* of the observation, $\mathbf{d}$, given some fixed realizations $\mathbf{M}$ and $\mathbf{w}$ is

$$\pi(\mathbf{d}\mid\mathbf{M},\mathbf{w}) = \left(\frac{1}{2\pi|\Gamma|}\right)^{q/2}\exp\left(-\frac{1}{2}\|\mathbf{M}\mathbf{w}-\mathbf{d}\|_\Gamma^2\right),$$

where for a vector $\mathbf{v}\in\mathbb{R}^q$ we define $\|\mathbf{v}\|_\Gamma^2 := \mathbf{v}^\top\Gamma^{-1}\mathbf{v}$ and denote the determinant of the noise covariance by $|\Gamma| = \prod_{i=1}^{q}\gamma_i^2$. Note that this simple noise model gives positive probability to negative observations, as well as observations where the sum of one block is greater than one, although in practice such observations should not exist. In the following, we will drop the normalization constant from the probability densities for readability.

We use prior distributions to incorporate *a priori* knowledge (i.e. knowledge uninformed by the data) on the distributions of M and W. In particular, the priors for M and W are assumed to be mutually independent so that the joint prior can be written as $\pi(\mathbf{M},\mathbf{w}) = \pi(\mathbf{M})\pi(\mathbf{w})$.

The feasible set for the binary matrix M is

$$\Omega_M := \left\{\mathbf{M}\in\{0,1\}^{q\times n}:\ \mathbf{M} = (\mathbf{M}_1,\dots,\mathbf{M}_m)^\top,\ \mathbf{M}_k\in\tilde{\Omega}_M, 1\leqslant k\leqslant m\right\},$$

where each block belongs to the set

$$\tilde{\Omega}_M := \left\{\mathbf{M}\in\{0,1\}^{(p-1)\times n}\ :\ \sum_{i=1}^{p-1}\mathbf{M}_{i,j}\leqslant 1,\ 1\leqslant j\leqslant n\right\}.$$

In other words, the column sums of each block are at most 1, since a strain cannot be associated with more than one class at each SNP site. It is easy to see that the cardinality of $\tilde{\Omega}_M$ is $|\tilde{\Omega}_M| = p^n$, and thus $|\Omega_M| = p^{mn}$. Henceforth, we will choose the prior distribution of M to be uniform, i.e. $\pi(\mathbf{M}) \propto \chi_{\Omega_M}(\mathbf{M})$, where χ_{Ω_M} denotes the characteristic function of the set Ω_M. However, our numerical methods can be readily generalized to support different distributions. For example, in sections 3.4 and 3.6 we could choose any distribution that is separable in the sense that

$$\pi(\mathbf{M}) \propto \chi_{\Omega_M}(\mathbf{M}) \exp\left(-\frac{1}{2}\sum_{k=1}^{m} r_k(\mathbf{M}_k)\right),$$

where r_k is a function depending only on the kth block of $\mathbf{M}$.

To reduce the ambiguity arising from different permutations of the columns in $\mathbf{M}$, we assume in this work that the entries in the vector of proportions W have non-increasing order. To be specific, we assume that W is supported in the set

$$\Omega_W := \left\{ \mathbf{w} \in \mathbb{R}^n \ : \ \sum_{j=1}^{n} \mathbf{w}_j = 1, \quad 1 \geqslant \mathbf{w}_1 \geqslant \mathbf{w}_2 \geqslant \ldots \geqslant \mathbf{w}_n \geqslant 0 \right\},$$

which is a subset of an $(n-1)$-dimensional affine hyperplane; see visualizations in figures 1 and 4 for the case $n = 3$. For simplicity, we assume that W is uniformly distributed in Ω_W, thus the prior density is $\pi(\mathbf{w}) \propto \chi_{\Omega_W}(\mathbf{w})$. Again, the setting can be readily generalized for other prior distributions. In particular, assuming that W is a truncated Gaussian random variable with mean $\overline{\mathbf{w}} \in \Omega_W$ and a positive-definite covariance matrix $\Gamma_W \in \mathbb{R}^{n\times n}$, i.e.

$$\pi(\mathbf{w}) \propto \chi_{\Omega_W}(\mathbf{w}) \exp\left(-\frac{1}{2}\|\mathbf{w} - \overline{\mathbf{w}}\|^2_{\Gamma_W}\right),$$

would not add any difficulties in the algorithms that follow.

Having discussed both the likelihood and prior terms, we apply Bayes' formula

$$\pi(\mathbf{M}, \mathbf{w} \mid \mathbf{d}) = \frac{\pi(\mathbf{d} \mid \mathbf{M}, \mathbf{w})\, \pi(\mathbf{M}, \mathbf{w})}{\pi(\mathbf{d})}$$

which for our choices for the priors leads to the posterior distribution

$$\pi(\mathbf{M}, \mathbf{w} \mid \mathbf{d}) \propto \exp\left(-\frac{1}{2}\|\mathbf{M}\mathbf{w} - \mathbf{d}\|^2_{\Gamma}\right)\ \chi_{\Omega_M}(\mathbf{M})\ \chi_{\Omega_W}(\mathbf{w}). \tag{2}$$

The posterior probability encodes both information provided by the data and by our prior knowledge about the biological applications at hand.

3.4. Block coordinate descent method for MAP estimation

Maximum *a posteriori* (MAP) estimation aims at finding the largest mode of the posterior distribution (2). Taking the negative logarithm of the posterior density and denoting

$$\varphi(\mathbf{M}, \mathbf{w}) := \|\mathbf{M}\mathbf{w} - \mathbf{d}\|^2_{\Gamma}$$

we obtain the constrained minimization problem

$$\min_{\mathbf{M},\mathbf{w}} \ \varphi(\mathbf{M}, \mathbf{w}) \quad \text{subject to} \quad \mathbf{M} \in \Omega_M,\ \mathbf{w} \in \Omega_W. \tag{3}$$

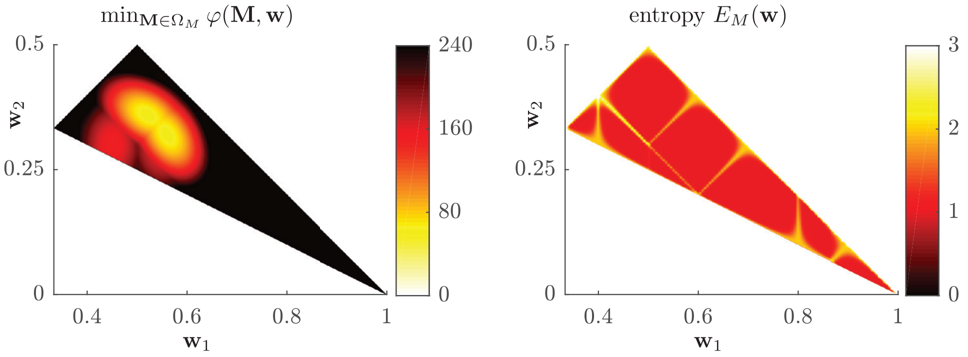

Figure 1. Left: minimum target function value as a function of the frequency vector $\mathbf{w}$ for the data $\mathbf{d}$ given in (15). Right: entropy of $\pi(\mathbf{M} \mid \mathbf{w}, \mathbf{d})$ for the same example, computed by (16).

Problem (3) is a *mixed-integer nonlinear programming* (MINLP) problem due to the binary constraints on $\mathbf{M}$ and the bilinear term $\mathbf{Mw}$.

Solving MINLP problems is known to be challenging as the computational complexity grows, in general, exponentially with the number of binary variables [27]. We will use the bilinearity of the forward problem and the separability of the posterior to obtain a block coordinate descent method whose computational cost is $\mathcal{O}(mp^n)$, i.e. the complexity grows linearly with the number of measurement locations and exponentially with the number of strains, which for the strain disambiguation application is rather small due to biological implausibility of a host simultaneously harboring dozens of competing pathogen strains. However, the method typically converges to a local minimum and may thus be needed to run several times to obtain a global minimizer; see also visualization of local minima in figure 1.

The block coordinate descent method decouples the problem (3) into two steps. The general idea is to alternate between updating the binary matrix $\mathbf{M}$ and the frequency vector $\mathbf{w}$ while keeping the respective other variable, or *block*, fixed. This is equivalent to maximizing the probabilities $\pi(\mathbf{M} \mid \mathbf{w}, \mathbf{d})$ and $\pi(\mathbf{w} \mid \mathbf{M}, \mathbf{d})$ repeatedly. At the ith iteration, starting from $(\mathbf{M}^{i-1}, \mathbf{w}^{i-1})$, we solve the two subproblems

$$\mathbf{M}^i = \arg\min_{\mathbf{M}} \varphi(\mathbf{M}, \mathbf{w}^{i-1}) \quad \text{subject to} \quad \mathbf{M} \in \Omega_M, \tag{4}$$

$$\mathbf{w}^i = \arg\min_{\mathbf{w}} \varphi(\mathbf{M}^i, \mathbf{w}) \quad \text{subject to} \quad \mathbf{w} \in \Omega_W. \tag{5}$$

For $p = 2$, this technique is presented in [19, 22, 28], and a similar alternating minimization approach has also been successfully employed in blind deconvolution [23].

In the first step, we find an exact solution to the binary-constrained optimization problem (4). Naïve solution of this problem would require full enumeration of all p^{mn} possible matrices and would be prohibitively expensive. However, we can decouple the problem along the blocks of the matrix, which yields m independent problems

$$\mathbf{M}_k^i = \arg\min_{\mathbf{M}_k} \left\|\mathbf{M}_k \mathbf{w}^{i-1} - \mathbf{d}_k\right\|_{\Gamma_k}^2 \quad \text{subject to} \quad \mathbf{M}_k \in \tilde{\Omega}_M, \tag{6}$$

where $\mathbf{d}_k \in \mathbb{R}^{p-1}$ denotes the kth block of the measurement vector for $k = 1, 2, \ldots, m$, and $\Gamma_k \in \mathbb{R}^{(p-1)\times(p-1)}$ is the corresponding block in the noise covariance matrix. Solving (6) can be parallelized, giving rise to additional computational savings. For the small problem sizes arising in the motivating public health applications, we use a full enumeration to solve each subproblem. Nevertheless, efficient software libraries such as Minotaur [29] and Gurobi [30] can be used for larger problem sizes. To summarize, solving (4) can be done in $\mathcal{O}(mp^n)$ flops, i.e. the complexity is linear with respect to the number of measurement locations and exponential in the number of strains.

The solution to (6) can be non-unique for two reasons. First, it may be possible to find two different blocks $\mathbf{M}_k, \mathbf{M}'_k \in \tilde{\Omega}_M$, both minimizing (6), such that $\mathbf{M}_k\mathbf{w}^{i-1} = \mathbf{M}'_k\mathbf{w}^{i-1}$, or equivalently $(\mathbf{M}_k - \mathbf{M}'_k)\mathbf{w}^{i-1} = \mathbf{0}$. This leads to the definition of *bi-independency* [31]: The values in $\mathbf{w}$ are bi-independent if $\mathbf{c}^\top\mathbf{w} \neq 0$ for all $\mathbf{c} \in \{0, -1, 1\}^n \setminus \{\mathbf{0}\}$. Clearly, for almost every $\mathbf{w} \in \Omega_W$ the values are bi-independent. Second, there may be two different minimizing blocks $\mathbf{M}_k, \mathbf{M}'_k \in \tilde{\Omega}_M$ such that $\mathbf{M}_k\mathbf{w}^{i-1} \neq \mathbf{M}'_k\mathbf{w}^{i-1}$. If $p = 2$, this means that $\mathbf{M}_k\mathbf{w}^{i-1} - \mathbf{d}_k = \mathbf{d}_k - \mathbf{M}'_k\mathbf{w}^{i-1}$, or equivalently

$$(\mathbf{M}_k + \mathbf{M}'_k)\mathbf{w}^{i-1} = 2\mathbf{d}_k. \tag{7}$$

For $\mathbf{d}_k = 1/2$ this holds for every $\mathbf{w}^{i-1} \in \Omega_W$ if we choose $\mathbf{M}'_k = \mathbf{1}^\top - \mathbf{M}_k$. Otherwise, for (7) to hold there must exist $\mathbf{c} \in \{0, 1/2, 1\}^n$ such that $\mathbf{c}^\top\mathbf{w}^{i-1} = \mathbf{d}_k$. If $p > 2$, the argument is not valid as such, but the general idea is still the same. We conclude that (6) has a unique solution for almost every $(\mathbf{w}^{i-1}, \mathbf{d}_k) \in \Omega_W \times \mathbb{R}^{p-1}$, and naturally (4) inherits a similar property as well.

In the second step of the block coordinate descent, we keep the binary matrix fixed and solve the convex quadratic programming problem (5) for the frequency vector. Due to the equality constraint for $\mathbf{w}$, the solution for (5) is unique if the rank of $\mathbf{M}^i$ is at least $n - 1$. We note that the gradient and Hessian of the objective function φ with respect to the continuous variable $\mathbf{w}$ are

$$\nabla_{\mathbf{w}}\varphi(\mathbf{M}, \mathbf{w}) = \mathbf{M}^\top\Gamma^{-1}(\mathbf{M}\mathbf{w} - \mathbf{d})$$

and

$$\nabla^2_{\mathbf{w}}\varphi(\mathbf{M}, \mathbf{w}) = \mathbf{M}^\top\Gamma^{-1}\mathbf{M},$$

respectively. The update $\delta\mathbf{w}$ is then obtained by approximately solving the convex quadratic program

$$\begin{aligned} &\min_{\delta\mathbf{w}} \quad \frac{1}{2}\delta\mathbf{w}^\top\nabla^2_{\mathbf{w}}\varphi(\mathbf{M}^i, \mathbf{w}^{i-1})\delta\mathbf{w} - \delta\mathbf{w}^\top\nabla_{\mathbf{w}}\varphi(\mathbf{M}^i, \mathbf{w}^{i-1}) \\ &\text{subject to} \quad 0 \leqslant \mathbf{w}^{i-1} + \delta\mathbf{w} \leqslant 1, \ \sum_{j=1}^{n}\delta\mathbf{w}_j = 0. \end{aligned} \tag{8}$$

To this end, we use a few steps of a standard active set method for quadratic programming; see, e.g. [32, ch. 16] for a detailed description.

The block coordinate descent approach for MAP estimation is listed in algorithm 1. Since there are only finitely many instances of the problem (5), and the value of the objective function cannot increase during the iteration, at some point the objective value must stagnate [28]. Thus, we repeat the steps (4) and (5) until there is no change in subsequent iterates $\mathbf{M}^i$ and $\mathbf{M}^{i-1}$. In practice, one may also want to monitor the change of the frequency vector $\|\mathbf{w}^i - \mathbf{w}^{i-1}\|$ and set a maximum number for the iterations to make sure that the algorithm also stops in the case of non-unique solutions for the subproblems.

There is no guarantee that the iteration converges to a global minimum, which is why the block coordinate method is run $n_T \in \mathbb{N}$ times with different random initial vectors $\mathbf{w}^0$, and the output with the highest probability is selected as the MAP estimate. This is usually a good strategy [19, 22, 28]; see also [25], where the dependency of the solution on the starting guess was established for the blind deconvolution problem without binary constraints.

In addition to the block-wise non-uniqueness stemming from the subproblems (4) and (5), the global minimum may be obtained with multiple elements of $\Omega := \Omega_M \times \Omega_W$ such that both blocks, i.e. the binary matrices and frequency vectors, differ. This can be interpreted as a generalization of the permutation invariance which is eliminated by our choice for the prior: If $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is a matrix such that $\mathbf{MQ} \in \Omega_M$ and $\mathbf{Q}^{-1}\mathbf{w} \in \Omega_W$, then clearly $(\mathbf{M}, \mathbf{w})$ and $(\mathbf{MQ}, \mathbf{Q}^{-1}\mathbf{w})$ correspond to the same value of the objective function φ. The existence of such nontrivial $\mathbf{Q}$ is discussed in detail in [20]. In short, the uniqueness (up to permutations) of the factorization $\mathbf{Mw}$ becomes rapidly more likely when the number of distinct rows in $\mathbf{M}$ increases.

Algorithm 1. Block coordinate descent for strain identification from mixed samples.

```
Input: Measurements $\mathbf{d} \in \mathbb{R}^q$, number of strains $n \in \mathbb{N}$, number of classes $p \geqslant 2$, number of trials $n_T \in \mathbb{N}$, tolerance $\varepsilon_w > 0$, maximum number of iterations $n_I \in \mathbb{N}$.
for $t = 1, 2, \ldots, n_T$ do
    Draw starting guess $\mathbf{w}^0$ uniformly from $\Omega_W$ and set, e.g. $\mathbf{M}^0 = -\mathbf{1}$.
    for $i = 1, \ldots, n_I$ do
        Get $\mathbf{M}^i$ block-wise by solving (6) for current $\mathbf{w}^{i-1}$.
        Get $\mathbf{w}^i$ by solving (8) for current $\mathbf{M}^i$.
        if $\mathbf{M}^i = \mathbf{M}^{i-1}$ and $\|\mathbf{w}^i - \mathbf{w}^{i-1}\| < \varepsilon_w$ then
            Exit the inner loop.
        end if
    end for
    Store local mode: $(\hat{\mathbf{M}}^t, \hat{\mathbf{w}}^t) = (\mathbf{M}^i, \mathbf{w}^i)$.
end for
Find the best mode: $\ell = \arg\min_{1\leqslant t\leqslant n_T} \varphi(\hat{\mathbf{M}}^t, \hat{\mathbf{w}}^t)$.
Output: MAP estimate $(\hat{\mathbf{M}}, \hat{\mathbf{w}}) = (\hat{\mathbf{M}}^\ell, \hat{\mathbf{w}}^\ell)$ or (if desired) all modes $(\hat{\mathbf{M}}^1, \hat{\mathbf{w}}^1), \ldots, (\hat{\mathbf{M}}^{n_T}, \hat{\mathbf{w}}^{n_T})$.
```

The prior distributions for both M and W involve the knowledge about the number of strains n, i.e. the MOI. As mentioned, this number may be unknown in many practical applications. An alternative to the method described above is to resort to the discrepancy principle with an approach that resembles the so-called 'regularization by discretization' technique [33, 34]. To this end, let $(\hat{\mathbf{M}}(n), \hat{\mathbf{w}}(n))$ denote the MAP estimate for a given $n \geqslant 1$. Now the goal is to find n such that the discrepancy between the measurement and the reconstruction is approximately equal to the magnitude of noise, which is still assumed to be known. More precisely, we start from $n = 1$ and keep increasing n until

$$d(n) := \|\hat{\mathbf{M}}(n)\hat{\mathbf{w}}(n) - \mathbf{d}\|_2^2 \leqslant \sum_{i=1}^{q} \gamma_i^2.$$

Note that d is a non-increasing function if the MAP estimates are global minimizers of the objective φ.

3.5. MAP estimation as a convex mixed-integer quadratic program

Although the block coordinate descent method is computationally simple and efficient in finding local minima, there is no guarantee that it yields a global minimum. Even with numerous trials, one may end up finding only local minima. However, we can reformulate the problem as a *mixed-integer quadratic program* (MIQP) with a convex objective function and linear constraints in addition to the binary restriction on the matrix $\mathbf{M}$. For moderate-sized instances, this class of program can be efficiently solved to global optimality by a commercial off-the-shelf solver such as Gurobi [30] or CPLEX [35].

In short, we reformulate the problem by replacing the bilinear term $\mathbf{M}_{i,j}\mathbf{w}_j$ with its so-called McCormick envelope for $i = 1, \ldots, q$ and $j = 1, \ldots, n$. We then use the fact that $\mathbf{M}$ is binary and every component of $\mathbf{w}$ is bounded between 0 and 1 to prove the equivalence between the two formulations. See [36] for a general treatment of this technique and [37, 38] for two examples of applications.

To write the McCormick envelopes, we define the auxiliary variables

$$\mathbf{Z}_{i,j} := \mathbf{M}_{i,j}\mathbf{w}_j, \qquad i = 1, \ldots, q, \quad j = 1, \ldots, n.$$

Using these new variables, problem (3) can be equivalently written as

$$\begin{array}{ll} \min & \displaystyle\sum_{i=1}^{q} \frac{1}{\gamma_i^2} \left(\sum_{j=1}^{n} \mathbf{Z}_{i,j} - \mathbf{d}_i \right)^2 \\ \text{subject to} & \mathbf{Z}_{i,j} = \mathbf{M}_{i,j}\mathbf{w}_j, \qquad i = 1, \ldots, q, \quad j = 1, \ldots, n, \\ & \mathbf{M} \in \Omega_M, \\ & \mathbf{w} \in \Omega_W. \end{array} \tag{9}$$

Notice that in (9) the objective is convex and all the non-convexity comes from the bilinear constraints defining $\mathbf{Z}$. Next, we replace each bilinear constraint by a convex envelope given by the McCormick's inequalities [39]:

$$\begin{array}{ll} \min & \displaystyle\sum_{i=1}^{q} \frac{1}{\gamma_i^2} \left(\sum_{j=1}^{n} \mathbf{Z}_{i,j} - \mathbf{d}_i \right)^2 \\ \text{subject to} & \mathbf{Z}_{i,j} \geqslant 0, \\ & \mathbf{Z}_{i,j} \leqslant \mathbf{M}_{i,j}, \\ & \mathbf{Z}_{i,j} \leqslant \mathbf{w}_j, \\ & \mathbf{Z}_{i,j} \geqslant \mathbf{M}_{i,j} + \mathbf{w}_j - 1, \qquad i = 1, \ldots, q, \quad j = 1, \ldots, n, \\ & \mathbf{M} \in \Omega_M, \\ & \mathbf{w} \in \Omega_W. \end{array} \tag{10}$$

We claim that the problems (9) and (10) are equivalent. Since both problems have the same objective, it is enough to show that both problems have the same set of feasible solutions. Let S_1 and S_2 denote the set of feasible solutions of problem (9) and problem (10), respectively. By construction, $S_1 \subseteq S_2$. Next we show that $S_1 \supseteq S_2$. Let $(\mathbf{M}, \mathbf{w}, \mathbf{Z}) \in S_2$. Then, for all $i = 1, \ldots, q$ and $j = 1, \ldots, n$, there are two cases:

(i) $\mathbf{M}_{i,j} = 0$: In this case, the McCormick's inequalities imply that $\mathbf{Z}_{i,j} = 0$ and $0 \leqslant \mathbf{w}_j \leqslant 1$. Thus, $\mathbf{Z}_{i,j} = \mathbf{M}_{i,j}\mathbf{w}_j$ in this case.

(ii) $\mathbf{M}_{i,j} = 1$: In this case, the McCormick's inequalities imply that $0 \leqslant \mathbf{Z}_{i,j} \leqslant 1$ and $\mathbf{Z}_{i,j} = \mathbf{w}_j$. Thus, $\mathbf{Z}_{i,j} = \mathbf{M}_{i,j}\mathbf{w}_j$ in this case as well.

Therefore, $(\mathbf{M}, \mathbf{w}, \mathbf{Z}) \in S_1$, and we conclude that $S_1 = S_2$ and hence the problems are equivalent.

Following, e.g. a branch-and-bound strategy, we fix some of the binary-constrained entries $\mathbf{M}_{i,j}$ to be either 0 or 1, and relax the binary restrictions from the remaining entries (see, e.g. [27] for a general overview of MINLP solvers). This way (10) becomes a convex quadratic program with linear constraints and continuous variables. Solving this relaxed problem is straightforward and similar to (8). By alternating the fixed entries in a tree-like fashion and comparing the minima of the relaxed problems, we establish upper and lower bounds for the problem (10) and thus for the minimum of the original target function φ. This tree is traversed until a desired gap between the upper and lower bounds is achieved.

3.6. Integrating the posterior density

Next, we discuss how to compute integrals that include the posterior density (2). This becomes useful when one wants to compute conditional moments such as mean or variance of the random variables M and W, given the observations D.

Let f be a function that satisfies

$$f(\mathbf{M},\mathbf{w}) = \sum_{k=1}^{m} f_k(\mathbf{M}_k,\mathbf{w}), \tag{11}$$

that is, f_k depends only on the kth block of $\mathbf{M}$ in addition to $\mathbf{w}$. Examples of such functions include $f(\mathbf{M},\mathbf{w}) = \mathbf{M}$ and $f(\mathbf{M},\mathbf{w}) = \mathbf{w}$, as well as the entrywise powers of $\mathbf{M}$ and $\mathbf{w}$. Here we present an integration scheme for computing the posterior mean of f. More precisely, we consider the integral

$$\begin{aligned}\mathbb{E}[f(M,W) \mid D] &= \int_{\Omega} f(\mathbf{M},\mathbf{w})\pi(\mathbf{M},\mathbf{w} \mid \mathbf{d})\,\mathrm{d}\mathbf{M}\mathrm{d}\mathbf{w}\\ &= \int_{\Omega_W} \sum_{\mathbf{M}\in\Omega_M} f(\mathbf{M},\mathbf{w})\pi(\mathbf{M},\mathbf{w} \mid \mathbf{d})\,\mathrm{d}\mathbf{w}.\end{aligned} \tag{12}$$

Naïvely summing over all possible binary matrices is impractical even for moderate parameter values, since $|\Omega_M| = p^{mn}$. Thus, we suggest a more efficient approach that exploits the separability of the posterior in the same fashion as (6).

First, note that the posterior density can be written as

$$\pi(\mathbf{M},\mathbf{w} \mid \mathbf{d}) = C\exp\left(\sum_{i=1}^{m} g_i(\mathbf{M}_i)\right) = C\prod_{i=1}^{m}\exp\left(g_i(\mathbf{M}_i)\right), \tag{13}$$

where C is a constant that depends on $\mathbf{d}$ and that can be computed by considering the case $f = 1$, and

$$g_i(\mathbf{M}_i) := -\frac{1}{2}\left\|\mathbf{M}_i\mathbf{w} - \mathbf{d}_i\right\|_{\Gamma_i}^2.$$

Next, the sum in (12) can be decomposed by iterating through the blocks separately, that is,

$$\sum_{\mathbf{M}\in\Omega_M} f(\mathbf{M},\mathbf{w})\pi(\mathbf{M},\mathbf{w} \mid \mathbf{d}) = \sum_{\mathbf{M}_1\in\tilde{\Omega}_M} \cdots \sum_{\mathbf{M}_m\in\tilde{\Omega}_M} f(\mathbf{M},\mathbf{w})\pi(\mathbf{M},\mathbf{w} \mid \mathbf{d}).$$

By substituting the expressions (11) and (13) into this sum, we can write the integrand in (12) as

$$
\begin{aligned}
&C \sum_{\mathbf{M}_1 \in \tilde{\Omega}_M} \cdots \sum_{\mathbf{M}_m \in \tilde{\Omega}_M} \prod_{i=1}^{m} \exp\left(g_i(\mathbf{M}_i)\right) \sum_{k=1}^{m} f_k(\mathbf{M}_k, \mathbf{w}) \\
=&C \sum_{k=1}^{m} \sum_{\mathbf{M}_1 \in \tilde{\Omega}_M} \exp\left(g_1(\mathbf{M}_1)\right) \cdots \sum_{\mathbf{M}_m \in \tilde{\Omega}_M} \exp\left(g_m(\mathbf{M}_m)\right) f_k(\mathbf{M}_k, \mathbf{w}).
\end{aligned}
\tag{14}
$$

For each k in the outermost sum, f_k can be put inside the sum that corresponds to the kth block of the binary matrix. As a result, the integrand in (12) becomes a sum where each summand is a product of sums.

The rapid decay of the exponential function can easily trigger underflows when using floating point arithmetics. Therefore, a numerically more stable version of the integration technique is outlined in algorithm 2. We see that for each quadrature node $\mathbf{w} \in \Omega_W$, the evaluation of the integrand has a computational complexity of $\mathcal{O}(mp^n)$. For the integral over Ω_W, we could apply some deterministic quadrature rule if n is small, but in the numerical experiments in the next section, we will instead use Monte Carlo method for simplicity [40].

Algorithm 2. Computing the conditional mean of a function f.

```
Input: Measurements $\mathbf{d} \in \mathbb{R}^q$, function $f$ of the separable form (11), quadrature nodes and
weights $\{\mathbf{w}^{(s)}, \zeta^{(s)}\}_{s=1}^{S} \subset \Omega_W \times \mathbb{R}$, number of strains $n$, number of measurement locations $m$,
number of classes $p$.
for $s = 1, \ldots, S$ do
    for $k = 1, \ldots, m$ do
        for $j = 1, \ldots, p^n$ do
            Let $\mathbf{M}^{(j)}$ be the $j$th element of $\tilde{\Omega}_M$ (in some arbitrary but fixed order)
            $L_{k,j} = -\frac{1}{2}\|\mathbf{M}^{(j)}\mathbf{w}^{(s)} - \mathbf{d}_k\|_{\Gamma_k}^2$
            $F_{k,j} = f_k(\mathbf{M}^{(j)}, \mathbf{w}^{(s)})$
        end for
        $U_k = \max_j L_{k,j}$ (for numerical stability)
        $P_k = \sum_j \exp(L_{k,j} - U_k)$
        $G_k = \sum_j F_{k,j} \exp(L_{k,j} - U_k)$
    end for
    Compute sum (14): $J_s(f) = G_1 P_2 P_3 \cdots P_m + P_1 G_2 P_3 \cdots P_m + \ldots + P_1 \cdots P_{m-1} G_m$
    $J_s(1) = \prod_{k=1}^{m} P_k$ (corresponds to $f = 1$)
    $\lambda_s = \sum_{k=1}^{m} U_k$
end for
Compute unnormalized integrals with the quadrature rule:
$I(f) = \sum_s \zeta_s J_s(f) \exp(\lambda_s - \max_\ell \lambda_\ell)$
$I(1) = \sum_s \zeta_s J_s(1) \exp(\lambda_s - \max_\ell \lambda_\ell)$
Output: Posterior mean of $f$ as $I(f)/I(1)$.
```

4. Numerical experiments

We validate our computational techniques using both synthetic and real data with known ground truths. In section 4.2 we illustrate the problem of non-uniqueness with some simple examples. We then introduce more realistic reconstruction problems in section 4.3 and study how different parameter values affect the accuracy of the reconstruction. Finally, in section 4.4 we apply our methods to experimental data.

4.1. Implementation

Our computational experiments are performed using the Julia [41] programming environment. We have implemented a module for strain identification that includes the block coordinate descent method and the numerical integration technique. For the convex problem formulation, we use Gurobi through an interface of the Julia module JuMP [42] to solve the problem with a global optimization method. Our implementation is freely available at https://github.com/lruthotto/StrainRecon.jl/

4.2. Identifiability

In this section we demonstrate with a few simple examples that the solution to the MAP estimation problem can be either unique or non-unique. The examples are chosen to be small such that full enumeration of the binary matrices are possible. That is, the results can be confirmed by solving $|\Omega_M| = p^{mn}$ quadratic programming problems (8). For larger problems this would not be feasible. Therefore, we also illustrate how the possible ambiguity can be seen in the posterior statistics.

First, let us choose $m = 3$ and $n = p = 2$, and consider the following example that illustrates the concept of bi-independency from section 3.4. Let $\mathbf{w}^{(1)} = (0.6, 0.4)^\top$, $\mathbf{w}^{(2)} = (0.5, 0.5)^\top$, and let $\mathbf{M} \in \{0, 1\}^{3\times 2}$ be

$$\mathbf{M} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Note that $\mathbf{c}^\top \mathbf{w}^{(2)} = 0$ for $\mathbf{c} = (1, -1)^\top$. The data is given by

$$\mathbf{d}^{(1)} = \mathbf{M}\mathbf{w}^{(1)} = \begin{pmatrix} 0.4 \\ 0.6 \\ 1.0 \end{pmatrix} \quad \text{and} \quad \mathbf{d}^{(2)} = \mathbf{M}\mathbf{w}^{(2)} = \begin{pmatrix} 0.5 \\ 0.5 \\ 1.0 \end{pmatrix}.$$

One can readily verify that there are no other pairs in Ω that would yield the first data vector $\mathbf{d}^{(1)}$, therefore the inverse problem has a unique solution. In contrast, $\mathbf{d}^{(2)}$ can be obtained by choosing $\mathbf{w}^{(2)}$ as above and any matrix in Ω_M which has row sums of 1 for the first two rows and 2 for the third row. Obviously, there are four such matrices.

We use the integration technique presented in section 3.6 to compute the posterior mean (i.e. conditional mean) and standard deviation for the unknowns $\mathbf{M}$ and $\mathbf{w}$. Throughout this section, we assume that the noise covariance matrix in (2) is $\Gamma = \gamma^2 I$ for some standard deviation $\gamma > 0$. For large values of γ we expect to see larger standard deviation in the posterior and the posterior mean is expected to approach the mean of the prior. However, already with $\gamma = 10^{-2}$ the posterior means $\mathbf{M}_{\mathrm{CM}}^{(1)}$ and $\mathbf{w}_{\mathrm{CM}}^{(1)}$ corresponding to $\mathbf{d}^{(1)}$ are practically indistinguishable from the true values and the posterior standard deviation $\mathbf{M}_{\mathrm{std}}^{(1)}$ is numerically zero. On the other hand, the standard deviation $\mathbf{w}_{\mathrm{std}}^{(1)}$ of the frequency vector is approximately $(0.007, 0.007)^\top$. As a comparison, for $n = 2$, the standard deviation of the essentially 1-dimensional uniform distribution on Ω_W is approximately 0.14.

As already seen above, $\mathbf{d}^{(2)}$ has more uncertainty in the reconstruction. This can also be verified by computing the posterior moments, since now we obtain

$$\mathbf{M}_{\mathrm{CM}}^{(2)} \approx \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{M}_{\mathrm{std}}^{(2)} \approx \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0 & 0 \end{pmatrix},$$

which are in line with the earlier observations. The frequency vector, however, has less uncertainty in the second case. The posterior mean is close to $(0.5, 0.5)^\top$, as expected, and the standard deviation is only $\mathbf{w}_{\mathrm{std}}^{(2)} \approx (0.004, 0.004)^\top$.

As a next example, we consider the identification problem for $(m, n, p) = (4, 3, 2)$ from the data

$$\mathbf{d} = (0.1, 0.3, 0.5, 0.6)^\top. \tag{15}$$

No exact solution for the inverse problem can be found, but four global minima for (3) can be obtained. These correspond to two different frequency vectors and four different binary matrices. The left side of figure 1 shows $\min_{\mathbf{M}\in\Omega_M} \varphi(\mathbf{M}, \mathbf{w})$ for different frequency vectors with $\gamma = 10^{-2}$. The global minima can be seen at $\mathbf{w} = (0.52, 0.36, 0.12)^\top$ and $\mathbf{w} = (0.56, 0.32, 0.12)^\top$. In addition, there is at least one local minimum at $\mathbf{w} = (0.45, 0.30, 0.25)^\top$, which we occasionally obtain as the output of our block coordinate descent algorithm.

The posterior mean and standard deviation for the binary matrix $\mathbf{M}$ are

$$\mathbf{M}_{\mathrm{CM}} \approx \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0.5 & 0.5 & 0.5 \\ 1 & 0 & 0.5 \end{pmatrix} \quad \text{and} \quad \mathbf{M}_{\mathrm{std}} \approx \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.5 & 0.5 & 0.5 \\ 0 & 0 & 0.5 \end{pmatrix},$$

respectively. The uncertainty in the third row results from having the value 0.5 in $\mathbf{d}$ (see section 3.4), whereas two different values in the lower right corner of $\mathbf{M}$ correspond to two different frequency vectors. For the frequency vector we obtain $\mathbf{w}_{\mathrm{CM}} \approx (0.54, 0.34, 0.12)^\top$ and $\mathbf{w}_{\mathrm{std}} \approx (0.022, 0.021, 0.008)^\top$.

The right hand side of figure 1 shows the entropy of the distribution $\pi(\mathbf{M} \mid \mathbf{w}, \mathbf{d})$, defined as a function of the frequency vector $\mathbf{w}$

$$E_M(\mathbf{w}) := \sum_{\mathbf{M}\in\Omega_M} \pi(\mathbf{M} \mid \mathbf{w}, \mathbf{d}) \log_2 \Big(\pi(\mathbf{M} \mid \mathbf{w}, \mathbf{d})\Big). \tag{16}$$

The entropy clearly indicates areas of Ω_W where the matrix minimizer of $\varphi(\mathbf{M}, \mathbf{w})$ is highly non-unique. For example, for $\mathbf{w} = (0.5, 0.3, 0.2)^\top$ there are 12 binary matrices $\mathbf{M}$ that result in the same minimal value of $\varphi(\mathbf{M}, \mathbf{w})$. Note that the entropy (16) can be efficiently computed also for larger examples by using the same row-decoupling technique as in (6).

Finally, let us mention that both MAP estimation techniques introduced in sections 3.4 and 3.5 reliably find a global minimizer in all previous examples, except in the last example where a local minimizer is sometimes returned by the block coordinate descent method if n_T is small. Which of the global minimizers is found depends on the starting points $\mathbf{w}^0$ and the implementation details; for example, how the minimizing matrix in (4) is chosen in case it is not unique.

4.3. Accuracy of the MAP estimation

We validate the accuracy of the block coordinate descent method (see algorithm 1) and the convex MIQP formulation (see section 3.5) using 10 000 randomly generated data sets with different noise levels.

Before quantifying the accuracies of our reconstruction methods, we have to define a meaningful distance function between the ground truth $(\mathbf{M}, \mathbf{w})$ and the reconstruction $(\hat{\mathbf{M}}, \hat{\mathbf{w}})$, or more generally, a distance between any two pairs in Ω. We first note that every misclassification in the binary matrix should have equal impact on the distance, that is, it should not matter whether we identify the first class as the third class or the third class as the second class, and so on. To ensure this property holds true, both binary matrices are augmented to matrices in $\{0, 1\}^{mp\times n}$ by adding the missing first row of each block so that the column sums of each block become exactly 1. We denote this modification of a matrix $\mathbf{M}$ by $\tau(\mathbf{M})$. Another observation is that the order of the strains does not matter; the requirement that the frequencies are in non-increasing order is for computational purposes only. As an extreme example, if we had

$$(\mathbf{M}, \mathbf{w}) = \left(\begin{pmatrix}1 & 0\\ 1 & 0\end{pmatrix}, \begin{pmatrix}0.51\\ 0.49\end{pmatrix}\right), \quad (\hat{\mathbf{M}}, \hat{\mathbf{w}}) = \left(\begin{pmatrix}0 & 1\\ 0 & 1\end{pmatrix}, \begin{pmatrix}0.51\\ 0.49\end{pmatrix}\right),$$

the strains would be identified perfectly and their frequencies would be reconstructed accurately, so we would expect a small distance. Therefore, we minimize over all possible permutations of the strains before computing the distance.

The distance, or reconstruction error e, can now be defined as

$$e(\mathbf{M}, \mathbf{w}, \hat{\mathbf{M}}, \hat{\mathbf{w}}) := \min_{\mathbf{P}\in\sigma(n)} \|\tau(\mathbf{M})\mathrm{diag}(\mathbf{w}) - \tau(\hat{\mathbf{M}})\mathrm{diag}(\hat{\mathbf{w}})\mathbf{P}\|_1, \tag{17}$$

where $\sigma(n)$ is the set of permutation matrices of size $n \times n$ and $\|\cdot\|_1$ denotes the entrywise 1-norm, i.e. the usual ℓ^1-norm after vectorization. Enumerating all permutations is feasible in our examples due to the small number of strains, n.

Let us study the distribution of the reconstruction error by sampling realizations $(\mathbf{M}, \mathbf{w}) \in \Omega$ and computing the corresponding MAP estimates $(\hat{\mathbf{M}}, \hat{\mathbf{w}})$ with algorithm 1, where $n_T = 20$, $\varepsilon_w = 10^{-3}$ and $n_I = 10$. For given values of m, n, and p, we draw 10 000 independent samples from the uniform prior distribution $\pi(\mathbf{M}, \mathbf{w})$ with the additional restriction that the matrix $\mathbf{M}$ must not contain duplicate columns. Before computing the MAP estimate, the data $\mathbf{d} = \mathbf{Mw}$ is contaminated with independent zero mean Gaussian noise with standard deviation $\gamma > 0$. For comparison, the reconstructions for the same noisy data are also computed after converting the objective function to convex form as described in section 3.5. The resulting MIQP problems are solved using Gurobi software with 'MIPGap' tolerance parameter set to 10^{-6}.

Figure 2 shows the reconstruction errors e for $m = 10$, $n \in \{3, 4\}$, $p \in \{2, 4\}$ and $\gamma \in \{10^{-2}, 10^{-3}\}$. For clarity, all reconstruction errors are sorted in ascending order. The average distance e between two random samples is shown for each case by a horizontal line. The first observation is that both reconstruction methods perform significantly better than just randomly drawing the reconstruction, even with uniform priors. We also notice that the reconstruction error increases when the number of strains, n, is increased, but decreases when the number of classes, p, is increased.

Comparing the two MAP estimation methods, we see that solving the convex MIQP problem yields smaller statistical error in all cases, compared to the block coordinate descent. For example, when $(n, p) = (4, 2)$, the former produces negligible error in almost two thirds of the samples, whereas for the latter, fewer than half of the samples are reconstructed with such

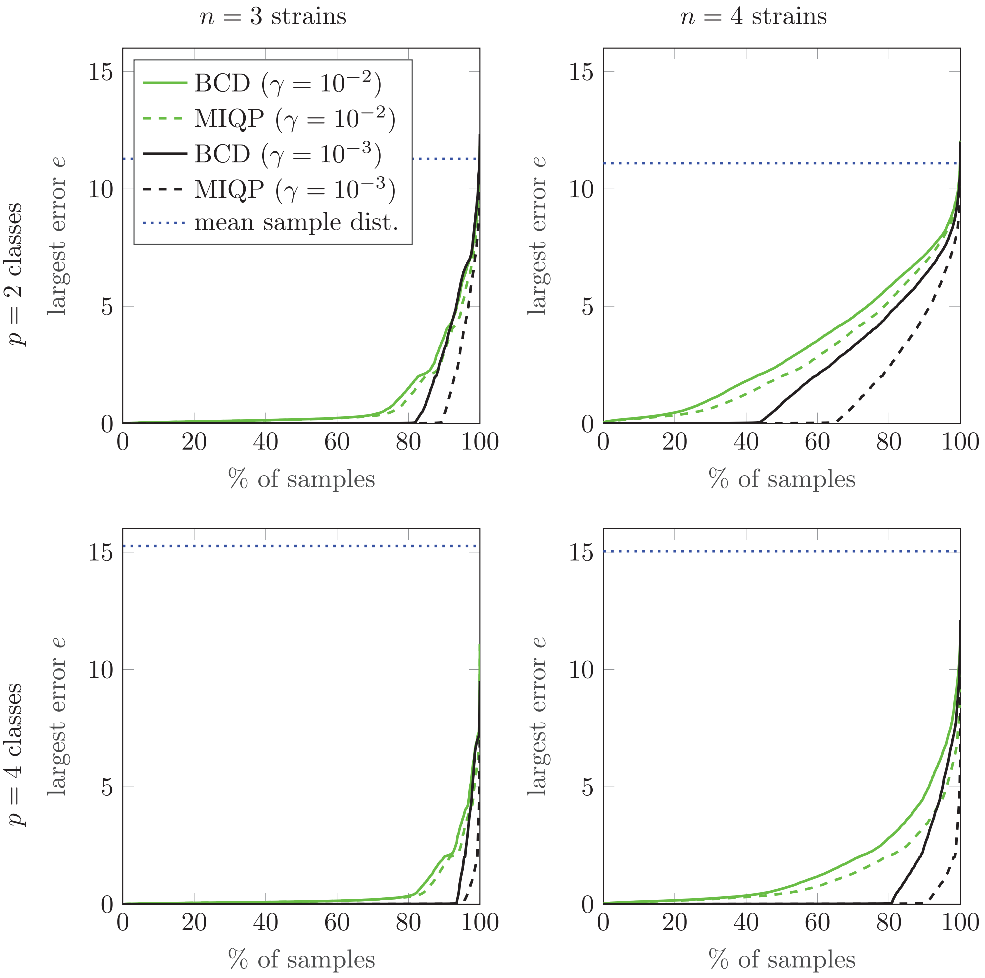

Figure 2. Sorted reconstruction errors for $m = 10$ measurement sites and Gaussian measurement error with standard deviation $\gamma = 10^{-2}$ (green) and $\gamma = 10^{-3}$ (black). The solid line depicts the block coordinates descent; the dashed line corresponds to the convex MIQP. The horizontal dashed line shows the average distance between the random samples.

accuracy. However, it is expected that increasing the number of trials, n_T, in algorithm 1 would make the block coordinate descent method perform better.

Unsurprisingly, the reconstruction errors become larger when the measurement noise is increased. In addition, with $\gamma = 10^{-2}$ the difference between the two reconstruction methods is less evident than with the smaller noise level.

4.4. Experimental data and uncertainty quantification

The initial motivation for the strain reconstruction was to tackle the practical challenge of disambiguating malarial strains. We now apply our algorithms to the open experimental dataset previously analyzed by Zhu *et al* [3]. The dataset is generated from lab-mixed *in vitro* samples of DNA from four laboratory parasite strains (3D7, Dd2, HB3, and 7G8) that are mixed in 27

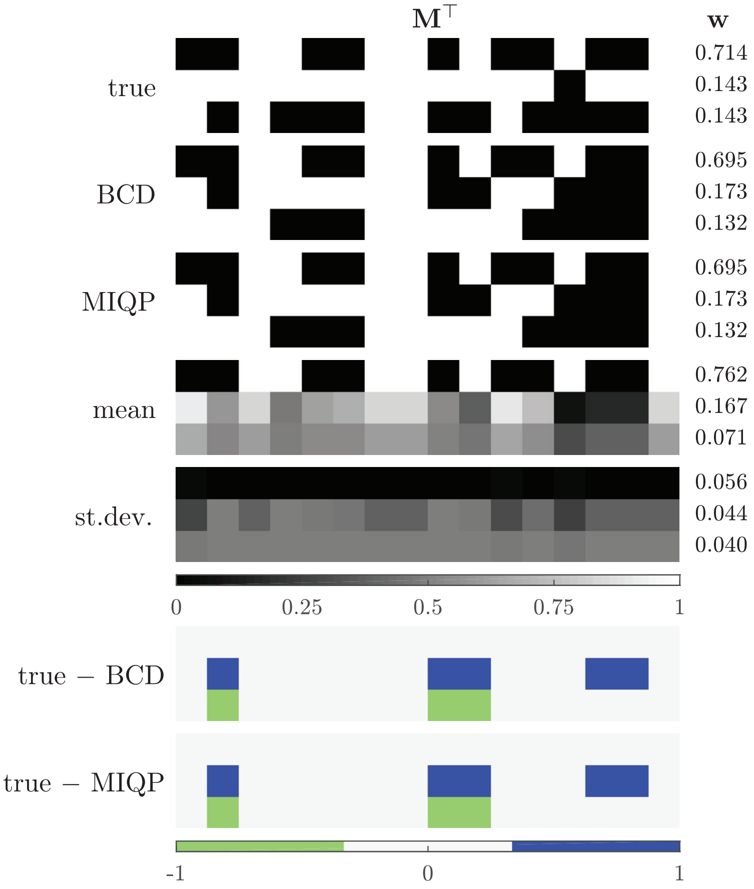

Figure 3. Ground truth and reconstructed $\mathbf{M}$ and $\mathbf{w}$ from using experimental measurement of $\mathbf{d}$ with $m = 16, p = 2, n = 3$ and assumed Gaussian random noise with zero mean and standard deviation $\gamma = 10^{-1}$. The relative strain frequencies are sorted from highest to lowest and shown to the right of their corresponding SNP barcodes. From top to bottom: $\mathbf{M}$ and $\mathbf{w}$ corresponding to the ground truth, two MAP estimates from sections 3.4 and 3.5, and the conditional mean and standard deviation.

different proportions. In a process similar to that of section 2, each of the 27 samples is sent to the MalariaGEN pipeline[6] and genotyped with an Illumina sequencing platform to produce a measurement vector. Since the mixture proportions are controlled, the underlying ground truth strain barcodes and frequencies $(\mathbf{M}, \mathbf{w})$ are known.

Only three of the 27 samples contained $n = 3$ strains, specifically PG0395-C, PG0396-C, and PG0397-C, and the remaining samples contained either a single strain or two strains. PG0395-C is a mixture of three parasite strains in near equal proportion, representing the edge case for identifiability where our algorithm has no basis for disambiguation (see discussion on bi-dependency in section 3.4).

To illustrate the power of our algorithm in a challenging scenario, we focus on sample PG0397-C that contains $n = 3$ strains in proportions 1:1:5. We compute $\mathbf{d}$ for each of the

[6] The Malaria Genomic Epidemiology Network: www.malariagen.net/.

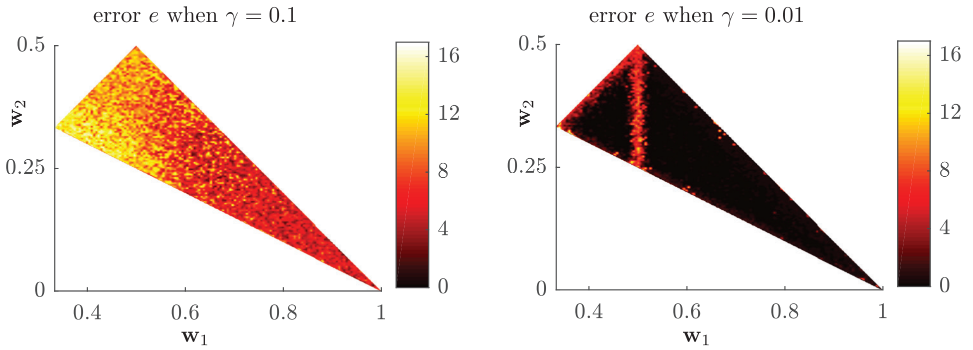

Figure 4. Reconstruction error e (17) from block coordinate descent MAP estimation for varying $\mathbf{w}$ given a fixed $\mathbf{M}$ and two noise levels. For noise level $\gamma = 10^{-1}$, higher reconstruction errors are observed in comparison to the same instances of $\mathbf{w}$ for the noise level $\gamma = 10^{-2}$. Cases where a component of $\mathbf{w}$ can (approximately) be expressed as a sum of one or two other components are reflected in higher reconstruction errors for the noise level $\gamma = 10^{-2}$ at red areas.

17 420 biallelic SNP sites of the sample by parsing the genomic sequences (reads) output by the Illumina sequencer in the VCF (Variant Call Format) format. At each SNP site, we used the proportion alternate read/(alternate read + reference read) for the $\mathbf{d}$ vector component, where alternate read and reference read refer to the number of reads that support the alternate allele ('non-reference') or reference allele present at the given SNP site, respectively.

Field samples often have low parasite copy numbers, and the number of SNP locations with recoverable allele frequencies will not necessarily reflect full genome coverage. We choose to compare the 16 SNP sites recovered from the Daniels *et al* [4] 24 SNP barcoding scheme in this sample set to allow direct evaluation of the two approaches. In our algorithms, we thus set $m = 16$ with $p = 2$.

The sample standard deviation of the error $\mathbf{d} - \mathbf{Mw}$ in the data is about 0.05. In our experiment, we assume the noise vector $\mathbf{n}$ to be a Gaussian random variable with zero mean and standard deviation $\gamma = 10^{-1}$. While more elaborate noise models may be used in practice, our goal is to demonstrate that a simple noise model works sufficiently well with real data when the standard deviation parameter is chosen appropriately.

The strain reconstructions $\hat{\mathbf{M}}$ and $\hat{\mathbf{w}}$ from the block coordinate descent ($n_T = 200$, $\varepsilon_w = 10^{-3}$ and $n_I = 10$) and the convex MIQP formulation are identical, as shown in figure 3. The figure also shows the ground truth $(\mathbf{M}, \mathbf{w})$, the conditional means and the posterior standard deviations. As expected with a 1:1:5 mixture, we observe larger standard deviations and reconstruction errors in the SNP barcodes of the two less prominent strains using either of the MAP estimation methods. This experiment also highlights the problem of identifiability mentioned in sections 3.4 and 4.2, as the two less prominent strains have an equal true relative frequency of 0.143. In contrast, we see a perfect reconstruction of the SNP barcode associated with the most prominent strain which has a true relative frequency of 0.714.

Finally, we also consider varying $\mathbf{w}$ when $\mathbf{M}$ is kept fixed at its true value. We generate $\mathbf{d}$ by using (1) and adding Gaussian noise with zero mean and standard deviation $\gamma = 10^{-1}$ or $\gamma = 10^{-2}$. For both noise levels, the reconstruction errors $e(\mathbf{M}, \mathbf{w}, \hat{\mathbf{M}}, \hat{\mathbf{w}})$, based on the block coordinate descent MAP estimation method ($n_T = 200$, $\varepsilon_w = 10^{-3}$ and $n_I = 10$), are shown side by side in figure 4. As expected, larger reconstruction errors can be seen for the noise level $\gamma = 10^{-1}$ in comparison to the case where the noise level is $\gamma = 10^{-2}$. It is worth noting two instances where large reconstruction errors are present for $\gamma = 10^{-2}$, namely when one component of $\mathbf{w}$ can be expressed as the sum of two other components, as represented by the vertical bright area at $\mathbf{w}_1 = 0.5$, and when two components are equal, as shown by the bright areas near the edges of the depicted triangle. The large reconstruction errors reflect the problem of bi-dependency as described in section 3.4.

5. Discussion and conclusion

In this paper, we present a mathematical formulation and computational framework for identifying strains of target microorganisms using PCR measurements from mixed samples. Extracting information about strains from mixed samples has the potential to reduce bias, time-to-results, and laboratory costs, and thus is critical for efficient screening. Our method alleviates the need for culturing and isolating pathogens to produce detailed genetic information, which makes it attractive for public health applications involving samples composed of multiple strains of the same microorganism. Epidemiological surveillance relies on the identification of microorganisms in samples, however, distinguishing multiple strains in mixed samples currently requires linking of locations [13] or a prior dictionary of known strains [1, 3]. Our methods do not require these limiting assumptions and are thus more broadly applicable.

Our main contribution is the mathematical formulation of strain identification as an inverse problem that estimates a binary matrix encoding the strains and a vector modeling their relative contributions to the measured data. The resulting problem is highly underdetermined and, also due to the presence of the binary constraints, challenging to solve. We propose several efficient methods inspired by structurally similar problems such as blind source separation [16, 19, 20, 28], non-negative matrix factorization [22] and blind deconvolution [23–25] but also leveraging result from mixed-integer programming.

Following a Bayesian approach, we derive a posterior density where prior information is incorporated to limit the underdetermined nature of the problem. The prior on the frequency vector enforces the non-negativity and sum-to-one properties, as well as a decreasing order to limit ambiguity. The prior on the strain matrix represents the binary constraints.

We propose efficient computational methods for exploring the posterior distribution. First, using block coordinate descent, we approximately solve the nonlinear mixed-integer problem arising in MAP estimation from different starting guesses to identify local and global modes. We exploit the fact that the optimization problem for the binary matrix decouples across rows to obtain a scheme whose complexity is linear with respect to the number of measurements and exponential in the number of strains to be recovered. Since the latter is relatively small in the target application, we can use full enumeration in this step. Second, we derive a convex re-formulation of the problem. This approach is less scalable but provides a lower bound for the negative log-likelihood that can be used to certify the optimality. Third, we propose an efficient numerical integration technique for estimating the conditional mean and standard deviation of the posterior.

As shown in our numerical examples on synthetic and experimental data with available ground truths, these methods allow one to discover the ambiguity of the problem at hand and capture uncertainty in the solution. Developing more scalable and accurate techniques to quantify the uncertainty by sampling from the multimodal posterior is a subject of future work.

Our work paves the way for fast and inexpensive species-specific differentiation of strains of targeted microorganisms through DNA barcoding and whole genome multilocus sequence typing, enabling epidemiologists and public health officials to conduct more granular tracking of pathogens and surveillance of infectious diseases.

Acknowledgments

We thank Ya Ping Shi at the Malaria Branch, Division of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention (CDC), U.S.A. for introducing us the need of new approach for molecular strain identification and providing initial funding. The authors would also like to thank Felix Lucka (CWI, Netherlands and UCL, UK) for fruitful discussions and his useful suggestions. This work is partially supported by the Advanced Molecular Detection fund from the CDC and the National Science Foundation (NSF), awards DMS#1522599 and CNS#1553579.

ORCID iDs

Lauri Mustonen https://orcid.org/0000-0002-8390-2151
Lars Ruthotto https://orcid.org/0000-0003-0803-3299

References

[1] Galinsky K *et al* 2015 COIL:a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data *Malaria J.* **14** 4

[2] Tenover F, Arbeit R and Goering R 1997 How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists *Infection Control Hosp. Epidemiol.* **18** 426–39

[3] Zhu S, Almagro-Garcia J and McVean G 2018 Deconvolution of multiple infections in *Plasmodium falciparum* from high throughput sequencing data *Bioinformatics* **34** 9–15

[4] Daniels R *et al* 2008 A general SNP-based molecular barcode for *Plasmodium falciparum* identification and tracking *Malaria J.* **7** 223

[5] Nadon C *et al* 2017 PulseNet international: vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance *Eurosurveillance* **22** 30544

[6] Maiden M, van Rensburg M J, Bray J, Earle S, Ford S, Jolley K and McCarthy N 2013 MLST revisited: the gene-by-gene approach to bacterial genomics *Nat. Rev. Microbiol.* **11** 728–36

[7] Quainoo S, Coolen J, van Hijum S, Huynen M, Melchers W, van Schaik W and Wertheim H 2017 Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis *Clin. Microbiol. Rev.* **30** 1015–63

[8] Singer R, Mayer A, Hanson T and Isaacson R 2009 Do microbial interactions and cultivation media decrease the accuracy of Salmonella surveillance systems and outbreak investigations? *J. Food Prot.* **72** 707–13

[9] Döpfer D, Buist W, Soyer Y, Munoz M, Zadoks R, Geue L and Engel B 2008 Assessing genetic heterogeneity within bacterial species isolated from gastrointestinal and environmental samples: how many isolates does it take? *Appl. Environ. Microbiol.* **74** 3490–6

[10] Jacob M, Almes K, Shi X, Sargeant J and Nagaraja T 2011 *Escherichia coli* O157: H7 genetic diversity in bovine fecal samples *J. Food Prot.* **74** 1186–8

[11] Sacchi C *et al* 2011 Incorporation of real-time PCR into routine public health surveillance of culture negative bacterial meningitis in São Paulo, Brazil *PLoS One* **6** e20675

[12] Langley G *et al* 2015 Effect of culture-independent diagnostic tests on future emerging infections program surveillance *Emerg. Infectious Dis.* **21** 1582–8

[13] Vollmers J, Wiegand S and Kaster A-K 2017 Comparing and evaluating metagenome assembly tools from a microbiologist's perspective—not only size matters! *PLoS One* **12** e0169662

[14] Kaipio J and Somersalo E 2006 *Statistical and Computational Inverse Problems* (*Applied Mathematical Sciences* vol 160) (New York: Springer)

[15] Calvetti D and Somersalo E 2007 *An Introduction to Bayesian Scientific Computing* (*Ten Lectures on Subjective Computing* vol 2) (New York: Springer)

[16] Diamantaras K and Chassioti E 2000 Blind separation of *N* binary sources from one observation: a deterministic approach *Int. Workshop on Independent Component Analysis and Blind Signal Separation* pp 93–8

[17] Diamantaras K 2006 A clustering approach for the blind separation of multiple finite alphabet sequences from a single linear mixture *Signal Process.* **86** 877–91

[18] Belouchrani A, Abed-Meraim K, Cardoso J-F and Moulines E 1997 A blind source separation technique using second-order statistics *IEEE Trans. Signal Process.* **45** 434–44

[19] Talwar S, Viberg M and Paulraj A 1994 Blind estimation of multiple co-channel digital signals using an antenna array *IEEE Signal Process. Lett.* **1** 29–31

[20] Behr M and Munk A 2017 Identifiability for blind source separation of multiple finite alphabet linear mixtures *IEEE Trans. Inf. Theory* **63** 5506–17

[21] Lin C-J 2007 Projected gradient methods for nonnegative matrix factorization *Neural Comput.* **19** 2756–79

[22] Slawski M, Hein M and Lutsik P 2013 Matrix factorization with binary components *Advances in Neural Information Processing Systems* pp 3210–8

[23] You Y and Kaveh M 1996 A regularization approach to joint blur identification and image restoration *IEEE Trans. Image Process.* **5** 416–28

[24] Chan T and Wong C 1998 Total variation blind deconvolution *IEEE Trans. Image Process.* **7** 370–5

[25] Chan T and Wong C 2000 Convergence of the alternating minimization algorithm for blind deconvolution *Linear Algebra Appl.* **316** 259–85

[26] Laird C, Biegler L and van Bloemen Waanders B 2006 Mixed-integer approach for obtaining unique solutions in source inversion of water networks *J. Water Resour. Plan. Manage.* **132** 242–51

[27] Belotti P, Kirches C, Leyffer S, Linderoth J, Luedtke J and Mahajan A 2013 Mixed-integer nonlinear optimization *Acta Numer.* **22** 1–131

[28] Talwar S, Viberg M and Paulraj A 1996 Blind separation of synchronous co-channel digital signals using an antenna array. I. Algorithms *IEEE Trans. Signal Process.* **44** 1184–97

[29] Mahajan A, Leyffer S, Linderoth J, Luedtke J and Munson T 2017 Minotaur: a mixed-integer nonlinear optimization toolkit ANL/MCS-P8010-0817, Argonne National Lab

[30] Gurobi Optimization, Inc. 2016 Gurobi optimizer reference manual

[31] Li Y, Cichocki A and Zhang L 2003 Blind separation and extraction of binary sources *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **86** 580–9

[32] Nocedal J and Wright S 2006 *Numerical Optimization* (*Springer Series in Operations Research and Financial Engineering*) (New York: Springer)

[33] Hämarik U, Kaltenbacher B, Kangro U and Resmerita E 2016 Regularization by discretization in Banach spaces *Inverse Problems* **32** 035004

[34] Hansen P 2010 *Discrete Inverse Problems* (*Fundamentals of Algorithms* vol 7) (Philadelphia, PA: SIAM)

[35] IBM Corp. 2016 IBM ILOG CPLEX optimization studio: CPLEX user's manual

[36] Gupte A, Ahmed S, Cheon M and Dey S 2013 Solving mixed integer bilinear problems using MILP formulations *SIAM J. Optim.* **23** 721–44

[37] Gupte A, Ahmed S, Dey S and Cheon M 2017 Relaxations and discretizations for the pooling problem *J. Glob. Optim.* **67** 631–69

[38] Schumacher K, Chen R-Y and Cohn A 2017 Transmission expansion with smart switching under demand uncertainty and line failures *Energy Syst.* **8** 549–80

[39] McCormick G 1976 Computability of global solutions to factorable nonconvex programs: part I—convex underestimating problems *Math. Program.* **10** 147–75

[40] Jacod J and Protter P 2004 *Probability Essentials* (Berlin: Springer)

[41] Bezanson J, Edelman A, Karpinski S and Shah V B 2017 Julia: a fresh approach to numerical computing *SIAM Rev.* **59** 65–98

[42] Dunning I, Huchette J and Lubin M 2017 JuMP: a modeling language for mathematical optimization *SIAM Rev.* **59** 295–320