Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Multivariate Analysis 97 (2006) 1221–1250
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

Parameter estimation under ambiguity and contamination with the spurious model

María Teresa Gallegos, Gunter Ritter*

Fakultät für Mathematik und Informatik, Universität Passau, D-94030 Passau, Germany

Received 17 June 2004
Available online 19 September 2005

Abstract

Recently, we proposed variants as a statistical model for treating ambiguity. If data are extracted from an object with a machine then it might not be able to give a unique *safe* answer due to ambiguity about the correct interpretation of the object. On the other hand, the machine is often able to produce a finite number of alternative feature sets (of the same object) that contain the desired one. We call these feature sets *variants* of the object. Data sets that contain variants may be analyzed by means of statistical methods and all chapters of multivariate analysis can be seen in the light of variants. In this communication, we focus on point estimation in the presence of variants and outliers. Besides robust parameter estimation, this task requires also selecting the regular objects and their valid feature sets (*regular variants*). We determine the mixed MAP–ML estimator for a model with spurious variants and outliers as well as estimators based on the integrated likelihood. We also prove asymptotic results which show that the estimators are nearly consistent.

The problem of variant selection turns out to be computationally hard; therefore, we also design algorithms for efficient approximation. We finally demonstrate their efficacy with a simulated data set and a real data set from genetics.

AMS 1991 subject classification: 62H12; 62F10; 62F15

Keywords: Point estimation under ambiguity; Variant selection; Outliers; Motif discovery in genetics

* Corresponding author. Fax: +49 851 509.
E-mail address: ritter@fmi.uni-passau.de (G. Ritter).

0047-259X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2005.08.002

1. Introduction

1.1. Variants—the general idea

We are constantly facing and effortless resolving a large amount of ambiguity without even noticing it; see Mumford [21] for an interesting exposition on this subject. However, ambiguity poses a major problem when objects such as acoustic signals, images, or geometric figures are to be processed with an automatic system, for example [27]. To this end, the automaton usually begins with extracting characteristic features from the objects. Feature extraction means abstraction and, thus, a reduction of complexity. Even if a machine is carefully programmed to react properly as soon as it faces ambiguities, it is often not able to produce a *unique* set of features in a safe way since feature extraction from a complex object depends also on the correct interpretation of the object. In complex situations, the interpretation may be ambiguous and finding the correct one may need the features to be extracted—a deadlock arises. In such cases, one may resort to extracting several (structurally equal) feature sets from the same object, one for each possible interpretation. We call these feature sets *variants* of the object. The *regular variant* comes from the (yet unknown) correct interpretation, the others are the *irregular variants*. The latter appear as outliers relative to the object w.r.t. the task and the former, only, are of interest. Doing this for several objects (of the same kind) results in a data set that represents some or all objects by more than one row. That is, whereas an ordinary data set consists of one row per object, a data set with variants contains as many rows per object as there are variants extracted from it, cf. Table 1. The more ambiguous the object, the more variants may be expected.

While the complexity is now reduced, the ambiguity is still conserved in the data set due to the presence of the variants. This paper is about how to resolve it. Since we do not expect the regular variants of the objects to be equal, this task is somewhat fuzzy and offers itself

Table 1
Three-dimensional data sets (a) without and (b) with variants. In (b), only one of the two variants of object 1 and one of the three variants of object 2 is a valid representative of its object

(a)			
object_1	5.37	1.62	2.45
object_2	4.11	2.21	2.13
object_3	3.34	4.54	5.46
object_4	8.35	6.76	7.78
object_5	1.36	2.48	1.41
object_6	5.76	7.61	2.15
(b)			
object_1	2.37	3.62	4.41
object_1	1.14	1.21	3.12
object_2	3.30	5.62	7.33
object_2	8.11	6.29	7.13
object_2	3.11	4.21	3.13
object_3	6.54	5.22	8.46

to statistical treatment. The regular variants are expected to be more densely concentrated than the irregular ones and it seems natural to resolve the ambiguity by selecting the most homogeneous variants across the objects as regular variants, thus determining the correct interpretations and characteristic feature sets. Thus, one faces the task of

- selecting a section across given groups of observations (the variants of the same object) that is as *homogeneous and compact* as possible.

In other words, one wishes to

- select *representatives* from the groups that match as far as possible the selected representatives of all other groups. [1]

Variants have recently been proposed [26,29,30] as a model for treating ambiguities. To our knowledge, the general problem of variant selection has rarely been studied systematically from a theoretical and statistical point of view, before. The study of variants leads into the field of multivariate analysis but the concept itself seems to be new there. Variants unify and extend several, apparently very different situations. By way of explaining their meaning and range, we describe next four situations where they are of benefit.

1.2. Motivating examples

(a) *Approximate substring identification*: Let there be given n strings of arbitrary lengths $\geqslant d$ over an alphabet $\mathbb{A}$. We ask for a pattern of length d that is, at least approximately, shared by all the strings. Each substring of length d of a string can be considered as a potential representative that flows from a specific interpretation of its string. In this sense, each string is an ambiguous object. This task can be embedded in the framework of variants by extracting from each string (object) the group of *all* its (overlapping) substrings of length d (the variants). Any method that finds the regular variants solves the problem since these are just defined to have the desired property. If we require exact instead of approximate matching, possibly with wild cards, then this problem is well known in computer science under the name "substring" or "factor matching" and possesses efficient solutions by dynamic optimization, see [8]. The problem of approximate substring matching is more complex and suggests to apply statistical methods.

An interesting and important application of this task is *motif discovery* in unaligned genetic and polypeptide sequences. In the former case, the alphabet $\mathbb{A} = \{a, g, c, t\}$ consists of the four nucleotides adenine, guanine, cytosine, and thymine, in the latter of the 20 amino acids that make up natural proteins. We will discuss this example in more detail in Section 4.2 showing how our method works on a specific genetic data set.

Motif discovery makes sense also in structures other than strings, for example, in graphs [10,37], where motifs are used for representing complex graphs.

(b) *Analysis of polarity and shape*: A second type of applications employs variants in order to handle *ambiguities* when a visual object or an acoustic signal is to be interpreted, in particular by a machine. The following example from biomedical image processing

[1] This concept is contrary to that of a centroid which represents best the members of *its* group.

Fig. 1. Four acrocentric chromosomes in metaphase (*Homo sapiens*, class 21). The two chromosomes on the left need two variants each, the others even three.

actually triggered our interest in variants. A cell of a eukaryotic organism contains a number of chromosomes constant within its species. A chromosome, visible under a microscope during the metaphase of mitosis, is a usually oblong object with a symmetry axis that divides it lengthwise into two so-called chromatids. Characteristics of a chromosome are, among others, its *area*, its *length*, and the *cosine coefficients* of the profile associated with its band pattern made visible by staining. Measuring the area is uncritical and easily performed by counting. In order to extract the remaining features with an automatic system, the symmetry axis has first to be determined by shape-analytical methods. It allows to measure the length and the cosine coefficients of even order. The odd ones depend on polarity which is not easily available from the shape alone, a first ambiguity, see the two chromosomes on the left of Fig. 1. Our method suggests to continue the analysis by extracting two variants from the object, one for each polarity.

The situation may be even more complex. One of the tasks of a chromosome is DNA replication; it accomplishes this by splitting into its two chromatids. Shortly before they are completely separated, they are joint merely at a region called centromere. Acrocentric chromosomes, whose centromeres are close to one tip, appear Y-shaped at this stage allowing even three possible shape interpretations, one for each branch of the Y, see the two chromosomes on the right of Fig. 1. Elementary approaches to resolving the ambiguity at this stage are prone to errors, see [24]. Our method proposes to first extract the corresponding variants and to combine the selection and estimation processes later by means of statistical methods. Acro- and telocentric chromosomes give rise to additional complications since, at the stage shortly before division, they may be confused with bent chromosomes.

Application of variant analysis to automatic classification of segmented metaphase cells [32–34] has substantially reduced the error rate compared with more classical approaches.

(c) *Segmentation problems*: The idea of extracting several variants from one object has been applied for some time in signal or image segmentation under the name "segmentation hypotheses." In fact, their consideration marks one of the main achievements in this field in the last two decades, see [3]. As a simple example, look at Fig. 2. It shows on the left side an object composed of two chromosomes in a human metaphase cell. Assume that 45 components have been found in the cell and that the other 44 have been identified as clear chromosomes. Since a normal human cell contains 46 chromosomes, the fact that this component is heavily bent leads to the conjecture that it might be composed of two. It is, however, not clear what the two components should be and the correct decomposition cannot be decided from the shape of the object alone. But the shape suggests two reasonable interpretations that are also shown in Fig. 2. They are the two variants to be considered in this example for further analysis.

Fig. 2. A component of a metaphase as observed under a microscope (left). This object is ambiguous since it can be interpreted in two different ways as composed of two chromosomes if its rectangular shape, only, is taken into account (center and right). This gives rise to creating two variants of the object. The interpretation on the right is correct.

Variants may, of course, also be applied when *linear* structures or random *sequences* have to be segmented and interpreted. Such structures appear in speech recognition, optical character recognition [3], and computational biology, [9,15]. A very popular method applicable to linear structures is the hidden Markov model (HMM) [25]. It may also be viewed as a method for generating and resolving segmentation variants.

(d) *A sociological example*: A very simple example is this: suppose that we wish to compare the quality of instruction between all school classes of a fixed grade in some state. The survey is to be carried out by testing just one student from each class. Now, the result of education depends both on instruction and on the students' properties, e.g. on their IQ's and on their backgrounds. Therefore, in order to reduce dependence on the latter, one should choose representatives whose properties and backgrounds are as homogeneous as possible. Here, the objects are the school classes and the variants are their students.

In cases (a) and (d), all possible variants are taken into account, in cases (b) and (c), the variants are created by means of special procedures from image processing. In all four cases, feature sets can now be extracted from the generated variants. If the meaningful one is among them, the object is regular and an outlier, otherwise. The original problem of analyzing the objects has thus been transformed to the statistical problem of analyzing a data set with variants and outliers. Statistical methods offer a reasonable way to select the correct interpretation thus resolving the ambiguity problem. It is the purpose of this paper to discuss this point in the context of parameter estimation.

1.3. Parameter estimation in the presence of variants and outliers

In a recent paper [30], we set out to study variants from a mathematical and statistical perspective dealing first with pure *variant selection*. Assuming the distribution of the *regular* population to be *known*, we studied the question whether the regular variant could be detected among all variants by a method that uses essentially this distribution alone. We called such a method a *simple selector*.

As we were investigating the pure selection problem, we realized that variants added a new perspective to multivariate analysis. A substantial part thereof, such as parameter estimation, discriminant analysis, clustering, regression analysis, factor analysis, and ro-

bustness can be viewed in the light of variants. Some of these topics will be treated in further communications. In the present paper, we deal with estimation of the *unknown* distribution parameters of the regular variants in the presence of irregular variants and outliers. If the positions of the regular variants among all variants were known for all objects then estimation of their parameters would be a classical task. Insofar our method extends classical point estimation. The main novelty is that these positions are hidden (or latent) and, in order to achieve our goal, we have to simultaneously estimate the positions of the regular variants *and* the parameters. There is a parallel to statistical clustering where both, class assignments and parameters, have to be estimated at the same time.

Almost all real data sets contain outliers in the sense of observations that are far away from the regular population. The classical point estimators usually break down in the presence of outliers. Some protection against outliers is, therefore, necessary if the estimators are to be useful in practice [5,28]. Since irregular variants may be viewed as outlying observations within an object, they offer a simple and natural way of defining statistical outliers as objects that lack the regular variant. This is our view here.

Thus, we assume that each object contains at most one regular variant and introduce the spurious-variants and -outliers model, Section 2. It treats the irregular variants as unpredictable—each irregular variant may obey its own distribution; they may or may not depend on the regular variant of their objects if there is one. We identify weighted maximum likelihood estimators, MAP w.r.t. variant selection and ML w.r.t. the parameters and, for some distributional assumptions, MAP-estimators of the selection based on the integrated likelihood w.r.t. the population parameters. The clear decision-theoretic background of such estimators leads to reasonable results in applications if the distributional assumptions are met. Interestingly, we find trimming estimators that turn out to be extensions of Rousseeuw's [35] minimum covariance determinant estimator, MCD, to variants. It does not seem to be well known that MCD is an ML-estimator w.r.t. some statistical model, a fact that was noticed by Pesch [23]. This estimator is known [17] to have the maximum asymptotic breakdown point of 50% that an affine equivariant point estimator can achieve.

Due to the presence of the irregular variants, our estimators are not exactly consistent. However, in Section 3, we establish some sample asymptotic results for normal regular populations that guarantee some kind of near consistency.

Although the criteria obtained are intuitively appealing they lead to computationally hard optimization problems as a consequence of the combinatorial problems of selection and trimming. Therefore, it is important to design efficient algorithms for approximate solutions and a major part of the paper is devoted to this point. The spurious-variants model leads to some straightforward algorithms, see Sections 2.3–2.6. The algorithms are efficient if the underlying statistical model allows simple expressions for the maximum likelihood estimates (m.l.e.) of the parameters. Examples are Gaussian and coin-tossing models for which we formulate efficient reduction steps as least units in the estimation process, Section 3. Section 4 contains a simulation study and an analysis of a well-known real data set from genetics. Both demonstrate the efficacy of the methods. The appendices contain some implementational issues and the proofs.

Other models with specific distributional assumptions on the irregular variants are, of course, possible and will be studied on later occasions.

1.4. General notation and preliminaries

The symbol E stands for a general sample space and f_X (f_μ, f_γ) signifies the density function of a random variable $X : \Omega \to E$ (probability distribution μ or μ_γ on E) w.r.t. some reference measure on E. We denote the m.l.e. of a parameter ϑ given the observation $\mathbf{x}$ by $\mathrm{ML}_\vartheta(\mathbf{x})$; this is in general a subset of the parameter set. Given an m-tuple $(x_1, \ldots, x_m)$, we often abbreviate $x_1^m = (x_j)_{j=1}^m = (x_1, \ldots, x_m)$ and $x_{\widehat{k}} = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_m)$. The symbol $\mathcal{S}_b$ denotes the group of all permutations of the index set $1..b$.

Our basic data set is an n-tuple $\mathbf{x} = (x_1, \ldots, x_n)$ of *multiple observations* $x_i = (x_{i,1}, \ldots, x_{i,b_i}) \in E^{b_i}$. The attribute "multiple" means that each object i, $1 \leqslant i \leqslant n$, is observed by the $b_i \geqslant 1$ variants $x_{i,1}, \ldots, x_{i,b_i}$; the number b_i may depend on the object. Given some sequence $\pi \in \prod_i \mathcal{S}_{b_i}$ of permutations, we write $\mathbf{x}_{\pi^{-1}} = \Big((x_{1,\pi_1^{-1}(k)})_{k=1}^{b_1}, \ldots, (x_{n,\pi_n^{-1}(k)})_{k=1}^{b_n}\Big)$; this is the data set with the variants reordered according to π.

Owing to the outliers, our estimators require an input parameter $r \leqslant n$, $r > 0$, to be chosen in advance. Its meaning is a lower bound on the number of regular objects contained in the data set. Our methods will work smoothly only if the actual number of outliers does not exceed $n - r$. We comment in Section 2.6(c) on how to choose r.

Central to our investigation is the notion of a (variant) *selection* $\mathbf{h}$ and its support $\mathbf{l}$. The selection is a partial function defined on r objects in $1..n$, the *support* of $\mathbf{h}$. We will view the latter as a function $\mathbf{l} : 1..n \to 0..1$ that marks r objects in $1..n$ as regular (1) and call any such function a support. The selection $\mathbf{h} = (h_i)_{l_i=1}$ determines h_i as the site of the regular variant of the regular object i. A selection implicitly contains the information about its support, that is, about its regular objects. For example, if $r = 2$, then ([object_1, 1], [object_2, 3]) is a selection in the data set (b) of Table 1. This selection considers object 3 as an outlier. Our main objective is estimating the "true" selection.

We denote the cross section $(x_{i,h_i})_{l_i=1}$ specified by a selection $\mathbf{h}$ by $x_{\mathbf{h}}$; it is a classical data set with one row per object i s.th. $l_i = 1$. The cross section of the selection above is

object_1	2.37	3.62	4.41
object_2	3.11	4.21	3.13

We write the letter b to denote the maximum number of variants, $\max_{1 \leqslant i \leqslant n} b_i$, taken over all objects of the data set and $\mathcal{F}_j$, $1 \leqslant j \leqslant b$, stands for the set of all objects i that contain *at least* j variants, $b_i \geqslant j$. Plainly, $\mathcal{F}_1 = 1..n$ and objects $i \in \mathcal{F}_2$ contain irregular variants.

2. Spurious outliers and variants

In this section, we assume that the regular variants come from an unknown member of some dominated, parametric statistical model $(\mu_\gamma)_{\gamma \in \Gamma}$ on E, while the irregular variants and outliers are *spurious* [4], comparable with gross outliers which obey no statistical law. We feel that the best way of handling this idea in a statistical (!) framework is by assuming that each irregular variant and outlier comes from its own population. This model is appropriate in situations where the irregular variants are unpredictable. The main aim is to estimate the parameter of the *regular* population while the irregular variants are considered as irrelevant, containing no information on the regular ones, and the outliers are ignored.

(a) *Regular and irregular variants*: We first introduce the *ordered* model of a *regular* object i, $l_i = 1$. Let $Z_i = (Z_{i,1}, \ldots, Z_{i,b_i}) : (\Omega, P) \to E^{b_i}$ denote its b_i variants in some natural order with the regular variant $Z_{i,1}$ in front. The law of $Z_{i,1}$ is a member of the parametric family $(\mu_\gamma)_{\gamma\in\Gamma}$. The irregular variants of an object $i \in \mathcal{F}_2$ make up the vector $Z_{i,\widehat{1}} = (Z_{i,2}, \ldots, Z_{i,b_i}) \in E^{b_i-1}$. The following is the main assumption on the spurious irregular variants.

($\mathrm{SV_r}$) The irregular variants $Z_{i,\widehat{1}}$ of an object $i \in \mathcal{F}_2$ obey a parametric model with parameter $\psi_i \in \Psi_i$ such that the likelihood integrated w.r.t. some prior measure τ_i on Ψ_i, satisfies

$$\int_{\Psi_i} f_{Z_{i,\widehat{1}}}[z_{i,\widehat{1}} \mid Z_{i,1} = z_{i,1}, \psi_i]\tau_i\,(\mathrm{d}\psi_i) = 1, \tag{1}$$

that is, it does not depend on z_i.

There are two important and sufficiently general situations where this is true.

(A) The state space $E = \mathbb{R}^d$ is Euclidean, $\Psi_i = E^{b_i-1}$, the irregular variants obey a *location model*

$$Z_{i,\widehat{1}} = U_i + \psi_i$$

with some (unknown) random noise $U_i : (\Omega, P) \to E^{b_i-1}$, and τ_i is Lebesgue measure on Ψ_i. Indeed, in this case, the conditional Lebesgue density is

$$f_{Z_{i,\widehat{1}}}[z_{i,\widehat{1}} \mid Z_{i,1} = z_{i,1}, \psi_i] = f_{U_i}[z_{i,\widehat{1}} - \psi_i \mid Z_{i,1} = z_{i,1}]$$

and, hence,

$$\int_{\Psi_i} f_{Z_{i,\widehat{1}}}[z_{i,\widehat{1}} \mid Z_{i,1} = z_{i,1}, \psi_i]\,\mathrm{d}\psi_i = 1.$$

(B) The parameter set Ψ_i is singleton, the irregular variants of an object are independent of the regular one, and the distribution of $Z_{i,\widehat{1}}$ is taken as the reference measure for its density. This case includes the idea of irregular variants "uniformly distributed" on some domain.

(b) *Outliers*: We consider outliers as objects that lack a regular variant. The counterpart of ($\mathrm{SV_r}$) for the outliers reads

($\mathrm{SV_o}$) The outlier $Z_i \in E^{b_i}$, $l_i = 0$, obeys a parametric model with parameter $\psi_i \in \Psi_i$ such that the likelihood integrated w.r.t. some prior measure τ_i on Ψ_i satisfies

$$\int_{\Psi_i} f_{Z_i}[z_i \mid \psi_i]\tau_i\,(\mathrm{d}\psi_i) = 1. \tag{2}$$

The conditions (A) and (B) carry over to corresponding conditions for outliers.

We assume that the sequence of objects $(Z_i)_{i=1}^n$ is statistically independent but not necessarily i.i.d., not even in cases where all b_i's are equal and $r = n$. Let $\mathbf{l}$ be some support. By the product formula, the likelihood for the data set $\mathbf{z} = (z_1, \ldots, z_n)$ (of ordered objects) is

$$f_Z(\mathbf{z} \mid \mathbf{l}, \gamma, \psi) = \prod_{i:l_i=1} f_\gamma(z_{i,1}) \prod_{i\in\mathcal{F}_2} f_{Z_{i,\widehat{1}}}[z_{i,\widehat{1}} \mid Z_{i,1} = z_{i,1}, \psi_i] \prod_{i:l_i=0} f_{Z_i}[z_i \mid \psi_i] \tag{3}$$

and, by (1) and (2), the likelihood integrated over the parameters of the irregular variants and outliers w.r.t. to the prior measures τ_i is

$$f_Z(\mathbf{z} \mid \mathbf{l}, \gamma) = \prod_{i:l_i=1} f_\gamma(z_{i,1}). \tag{4}$$

(c) *Permutations and trimming*. What we actually observe are the random vectors Z_i with their variants in (unobservable) disorder. We, thus, model the ith observation x_i as a realization of a random vector $(Z_{i,T_i(1)}, Z_{i,T_i(2)}, \ldots, Z_{i,T_i(b_i)})$ with random permutations T_i of $1..b_i$. Let us denote the random support by $L : \Omega \to \{\mathbf{l} \in (0..1)^n \mid \sum_i l_i = r\}$. We assume that it is uniformly distributed and that, given L, the T_i's are independent and so are T_1^n and Z_1^n. We also assume that the random variables T_i and $L_{\widehat{i}}$ are independent given L_i, $1 \leqslant i \leqslant n$.

Finally, we assume *regularity* of the permutations T_i as defined in [30, p. 318]. This means that, for each regular object i ($L_i = 1$), the probability $P[T_i = \pi_i | L_i = 1]$, $\pi_i \in \mathcal{S}_{b_i}$, depends only on the site $h_i := \pi_i^{-1}(1)$ of its regular variant. It may depend on the object i itself. It is this regularity that, in combination with the condition ($\mathrm{SV_r}$), allows to estimate the positions of the regular variants without having to care about the permutations. Correspondingly, if i is an outlier, we say that T_i is regular if each permutation is equally likely. Let $q_{i,k}$ denote the prior probability for the regular variant of the ith object to be found at position k. Without loss of generality, we may assume $q_{i,k} > 0$. Abbreviating $c_{\mathbf{h}} = \prod_{i:l_i=1} b_i q_{i,h_i}$, we conclude the description of our model with the following lemma.

2.1 Lemma. *We have* $P[T = \pi \mid L = \mathbf{l}] = C c_{\mathbf{h}}$ *with the constant* $C = \prod_{i=1}^n \frac{1}{b_i!}$.

We next propose several estimators for $\mathbf{h}$ and γ, a combined MAP–ML-estimator in the context of a general population and MAP estimators based on the likelihood integrated w.r.t. γ in special cases, see Sections 2.7 and 3. In view of an MAP–ML-estimator, we combine a maximum likelihood approach for estimating $\gamma \in \Gamma$ with maximum a posteriori inference for estimating the regular objects and selecting the regular variants. This means maximizing the joint conditional density $f_{L,T,X}[\mathbf{l}, \pi_1^n, \mathbf{x} \mid \gamma]$ w.r.t. $\mathbf{l}$, $\pi_1^n \in \prod_{i=1}^n \mathcal{S}_{b_i}$, and $\gamma \in \Gamma$. Given a selection $(h_i)_{l_i=1}$, we will abbreviate $f_\gamma(x_{\mathbf{h}}) = f_{\mathbf{l},\gamma}(x_{\mathbf{h}}) = \prod_{i:l_i=1} f_\gamma(x_{i,h_i})$. By the independences given L postulated above and by Lemma 2.1 and (4), it equals

$$\begin{aligned} f_{L,T,X}[\mathbf{l}, \pi_1^n, \mathbf{x} \mid \gamma] &= f_{L,T,Z_\pi}[\mathbf{l}, \pi_1^n, \mathbf{x} \mid \gamma] = f_{L,T,Z}[\mathbf{l}, \pi_1^n, \mathbf{x}_{\pi^{-1}} \mid \gamma] \\ &= P[L = \mathbf{l}] P[T = \pi_1^n \mid L = \mathbf{l}] f_Z[\mathbf{x}_{\pi^{-1}} \mid \mathbf{l}, \gamma] \\ &= \text{const} \cdot c_{\mathbf{h}} \prod_{l_i=1} f_\gamma(x_{i,h_i}) = \text{const} \cdot c_{\mathbf{h}} f_{\mathbf{l},\gamma}(x_{\mathbf{h}}). \end{aligned} \tag{5}$$

We call the last expression a *weighted likelihood*. It depends on γ and $\mathbf{h}$ and, therefore, also on $\mathbf{l}$. Its maximizer w.r.t. $\mathbf{l}$, $\mathbf{h}$, and γ, which we call the *weighted m.l.e.*, may be obtained by the Principle of Dynamic Optimization in three steps. In order to ensure its applicability, we require the data $\mathbf{x}$ to satisfy the condition

($\mathrm{GP_1}$) the ML-estimate of γ for the cross section $x_{\mathbf{h}}$ exists for any selection $(h_i)_{l_i=1} \in \prod_{l_i=1} 1..b_i$.

In the case of a Euclidean space E, this is often a condition on the affine geometry of the data $\mathbf{x}$. Denoting the ML-estimate of γ for a given selection $\mathbf{h}$ by $\gamma(\mathbf{h})$ $(= \mathrm{ML}_\gamma(x_\mathbf{h}))$, we have the following theorem.

2.2 Theorem (*The weighted m.l.e. for the spurious model*). *Assume the conditions* $(\mathrm{SV_r})$, $(\mathrm{SV_o})$, *and* $(\mathrm{GP_1})$, *regularity of all permutations* T_i, *and all other assumptions made above*.

(a) *The weighted m.l.e. of the selection* $\mathbf{h} \in \prod_{l_i=1} 1..b_i$ *of the positions of the regular variants is determined by the maximum of the criterion*

$$c_\mathbf{h} f_{\mathbf{l},\gamma(\mathbf{h})}(x_\mathbf{h}) = c_\mathbf{h} \max_\gamma f_{\mathbf{l},\gamma}(x_\mathbf{h}) = c_\mathbf{h} \max_\gamma \prod_{l_i=1} f_\gamma(x_{i,h_i}), \tag{6}$$

taken over all selections $\mathbf{h}$.

(b) *If the maximizer is denoted by* $\mathbf{h}^\star$ (*if it is not unique*, *choose one*) *then the weighted m.l.e. of the parameter* γ *of the regular variants is* $\gamma(\mathbf{h}^\star)$.

Eq. (6) shows that it is the maximum *value* $\max_\gamma f_{\mathbf{l},\gamma}(x_\mathbf{h})$ of the likelihood function that is needed in order to determine the optimal selection. However, in general, this will require the maximizer $\gamma(\mathbf{h})$.

If all prior probabilities q_i are uniform, then $c_\mathbf{h}$ does not depend on $\mathbf{h}$ and the ML-estimates can be given another interpretation. Given i.i.d. random variables X_i, $1 \leqslant i \leqslant r$, distributed according to an unknown "true" distribution μ, the arithmetic means

$$\frac{1}{r}\sum_{i=1}^{r} -\ln f_\mu(X_i) \quad \text{and} \quad \frac{1}{r}\sum_{i=1}^{r} \ln \frac{f_\mu}{f_\gamma}(X_i)$$

converge to the entropy $-E \ln f_\mu(X_1)$ of μ and to the Kullback–Leibler divergence $E \ln \frac{f_\mu}{f_\gamma}(X_1)$ of μ and γ, respectively, P-a.s. Hence, given a finite sequence $x_1, \ldots, x_r$ of observations, the means

$$\frac{1}{r}\sum_{i=1}^{r} -\ln f_\mu(x_i) \quad \text{and} \quad \frac{1}{r}\sum_{i=1}^{r} \ln \frac{f_\mu}{f_\gamma}(x_i)$$

are sample versions of these quantities. Neither of the two can be computed since f_μ is unknown, but their sum $\frac{1}{r}\sum_{i=1}^{r} -\ln f_\gamma(x_i)$ is an expression of γ alone. Theorem 2.2 says that the m.l.e. of $\mathbf{h}$ and γ minimizes this sum applied to $x_\mathbf{h}$. In other words, the m.l.e. chooses a selection with a small sample entropy for which there exists at the same time a parameter γ with small sample divergence. If the parameter can be chosen in such a way that the divergence vanishes (this is possible, for instance, if E is discrete) then the minimum is no larger than the sample entropy of the regular variants selected.

Theorem 2.2 reduces the problem of estimating selections and parameters to a combinatorial optimization problem and to maximizing likelihood functions. Now, there are astronomically many selections, $\sum_{C \in \binom{1..n}{r}} \prod_{i \in C} b_i$; enumerating all is not feasible except for small instances and approximation algorithms are desirable. In the remainder of this section, we design and substantiate a number of such algorithms.

2.3. Local search

Our first algorithm is a local descent method called *Glauber dynamics* in statistical physics. It is based on the criterion itself. Define a neighborhood structure on the set of all selections by declaring two selections as neighboring if they either differ in the regular variant of one regular object *or* if one outlier is declared as regular (with some regular variant) and vice versa. We, thus, obtain the following reduction step. It is useful only if the statistical model of the regular variants allows an efficient update formula for $f_{\gamma(\mathbf{h}')}(x_{\mathbf{h}'})$ from $f_{\gamma(\mathbf{h})}(x_{\mathbf{h}})$ for any two neighboring selections $\mathbf{h}$ and $\mathbf{h}'$; cf. Section 3.

The local reduction step
// Input: A selection $\mathbf{h}$ and its corresponding parameters $\gamma(\mathbf{h})$;
// Output: A selection $\mathbf{h}_{\text{new}}$ with larger Criterion (6) and $\gamma(\mathbf{h}_{\text{new}})$
or the response "local maximum."

(i) search for a neighbor $\mathbf{h}'$ of $\mathbf{h}$ such that

$$c_{\mathbf{h}'} f_{\gamma(\mathbf{h}')}(x_{\mathbf{h}'}) > c_{\mathbf{h}} f_{\gamma(\mathbf{h})}(x_{\mathbf{h}}) \tag{7}$$

(e.g. the first occurrence, or the smallest value, or something in between);
(ii) *if* there is such an $\mathbf{h}'$ then return $\mathbf{h}_{\text{new}} = \mathbf{h}'$ together with $\gamma(\mathbf{h}_{\text{new}})$;
// this value has been computed in (i)
else return "local maximum."

Now, starting from an initial selection $\mathbf{h}{=}\mathbf{h}^{(0)}$ and iterating local reduction steps, we obtain a sequence $(\mathbf{h}^{(t)})_{t=0}^{N}$ of selections such that Criterion (6) increases, i.e.,

$$c_{\mathbf{h}^{(t+1)}} f_{\gamma(\mathbf{h}^{(t+1)})}(x_{\mathbf{h}^{(t+1)}}) > c_{\mathbf{h}^{(t)}} f_{\gamma(\mathbf{h}^{(t)})}(x_{\mathbf{h}^{(t)}}), \quad t < N.$$

Since the number of possible selections is finite, this iterative process must reach a local maximum after a finite number, say N, of steps; it is detected in step $N+1$. We will discuss in Section 2.6, how to improve the local optima. However, if n is large, the problem of finding a *global* optimum is inherently hard and, like clustering, the problem of finding an optimal string (the variant selection in our context) belongs to the class of intractable problems well known in computer science.

The local reduction step has two disadvantages: first, it needs an update of the density $f_{\gamma(\mathbf{h})}$ for each trial in (i) even if it is unsuccessful and, second, it does not allow the simultaneous swapping of more than one variant or more than one object. We, therefore, present two more reduction steps that do not suffer from these shortcomings. The main idea of the subsequent algorithms is contained in the next proposition.

2.4 Proposition. *Assume* (GP_1) *and let* $\mathbf{h}$ *and* $\mathbf{h}_{\text{new}}$ *be two selections such that*

$$\sum_{l_{\text{new,i}}=1} \left\{ \ln b_i q_{i,h_{\text{new,i}}} + \ln f_{\gamma(\mathbf{h})}(x_{i,h_{\text{new,i}}}) \right\} \geqslant \sum_{l_i=1} \left\{ \ln b_i q_{i,h_i} + \ln f_{\gamma(\mathbf{h})}(x_{i,h_i}) \right\}. \tag{8}$$

(a) *Then* $c_{\mathbf{h}_{\text{new}}} f_{\gamma(\mathbf{h}_{\text{new}})}(x_{\mathbf{h}_{\text{new}}}) \geqslant c_{\mathbf{h}} f_{\gamma(\mathbf{h})}(x_{\mathbf{h}})$; *cf. Criterion* (6).
(b) *If there is strict inequality in* (8) *then there is also strict inequality in* (a).

(c) *Let there be equality in* (a). *If, for any set of observations, the m.l.e.* $\mathrm{ML}(\gamma)$ *is unique then we have* $\gamma(\mathbf{h}_{\mathrm{new}}) = \gamma(\mathbf{h})$. *If, in addition,* $\max_\gamma f_\gamma(z)$ *depends on* $z = (z_1, \ldots, z_n)$, $z_i \in E$, *only by way of* $\mathrm{ML}_\gamma(z)$ *then we have also* $c_{\mathbf{h}_{\mathrm{new}}} = c_{\mathbf{h}}$.

2.5. Reduction steps based on Proposition 2.4

Proposition 2.4 is the basis for reduction steps more efficient than the local one. Given some selection $\mathbf{h}$, the weights

$$u_{\mathbf{h}}(i,k) := -\ln b_i q_{i,k} - \ln f_{\gamma(\mathbf{h})}(x_{i,k}), \quad i \in 1..n, \quad k \in 1..b_i. \tag{9}$$

play a key role. Condition (8) is equivalent to

$$\sum_{l_{\mathrm{new,i}}=1} u_{\mathbf{h}}(i, h_{\mathrm{new,i}}) \leqslant \sum_{l_i=1} u_{\mathbf{h}}(i, h_i). \tag{10}$$

Note that, in contrast to (7), the parameters of the same selection $\mathbf{h}$ appear on both sides of the estimate. Let us call the pair (i, h_i) *inconsistent* with $\mathbf{h}$ if there exists a neighbor (i', k) such that $u_{\mathbf{h}}(i', k) < u_{\mathbf{h}}(i, h_i)$. If a selection $\mathbf{h}$ possesses an inconsistency and some or all of them are improved with result $\mathbf{h}_{\mathrm{new}}$ then (10) holds with strict inequality and Proposition 2.4 assures us that $\mathbf{h}_{\mathrm{new}}$ improves Criterion (6). It is possible to improve *one* or *several* inconsistencies at a time. Consequently, we formulate two more reduction steps. Like the local reduction step, the first one looks for a better neighbor.

The single-point reduction step
// Input: A selection $\mathbf{h}$;
// Output: A selection $\mathbf{h}_{\mathrm{new}}$ with larger Criterion (6) *or* the response "stop."

(i) Compute the estimate $\gamma(\mathbf{h})$ (using update formulae);
(ii) search for a neighbor $\mathbf{h}'$ of $\mathbf{h}$ defined by a regular object i and a variant (j, k) (see Section 2.3) such that $u_{\mathbf{h}}(j,k) < u_{\mathbf{h}}(i, h_i)$ (e.g. the first occurrence *or* $u_{\mathbf{h}}(j,k) - u_{\mathbf{h}}(i, h_i)$ minimum and < 0);
(iii) *if* there are such a variant then return $\mathbf{h}_{\mathrm{new}} = \mathbf{h}'$;
else "stop."

The other extreme is to remove all inconsistencies at a time. This means minimizing the sum

$$\sum_i u_{\mathbf{h}}(i, k_i) \tag{11}$$

over all selections $\mathbf{k} = (k_i)$. Summation runs over all objects i regular w.r.t. the selection $\mathbf{k}$. In algorithmic terms, the corresponding reduction step reads as follows.

The multi-point reduction step
// Input: A selection $\mathbf{h}$;
// Output: A selection $\mathbf{h}_{\mathrm{new}}$ with larger Criterion (6) *or* the response "stop."

(i) Compute the estimate $\gamma(\mathbf{h})$;
(ii) for each object $i \in \mathcal{F}_2$, determine an element $h_{\mathrm{new,i}} \in \mathrm{argmin}_{k \in 1..b_i} u_{\mathbf{h}}(i, k)$;

(iii) determine the r objects i with minimum values $u_{\mathbf{h}}(i, h_{\text{new,i}})$ and call the corresponding selection $\mathbf{h}_{\text{new}}$;
(iv) *if* (11) has improved with $\mathbf{k} = \mathbf{h}_{\text{new}}$ then return $\mathbf{h}_{\text{new}}$;
else "stop."

Now, starting from an initial selection $\mathbf{h}^{(0)}$ and iterating reduction steps, we obtain a sequence $(\mathbf{h}^{(t)})_{t \geqslant 0}$ of selections such that $\sum_i u_{\mathbf{h}^{(t)}}(i, h_i^{(t)})$ decreases; by Proposition 2.4(a), Criterion (6) increases. Since the number of selections is finite, the criterion must reach equality after a finite number of steps and Proposition 2.4(b) shows that at latest then, maybe earlier, the signal "stop" will appear. Let this happen in step $N + 1$,

$$\sum_i u_{\mathbf{h}^{(N)}}(i, h_i^{(N+1)}) = \sum_i u_{\mathbf{h}^{(N)}}(i, h_i^{(N)}). \tag{12}$$

The selection $\mathbf{h}^{(N)}$ is one approximation to the optimal solution. It is "self-consistent" in the sense that (8) and (10) with $\mathbf{h} = \mathbf{h}^{(N)}$ cannot hold with strict inequality. It is also a simple selector [29,30] for all regular objects i selected w.r.t. the estimated parameters.

The appearance of the stop signal in the single- and multi-point reduction steps does not mean that a local maximum w.r.t. the selection graph is reached. If $\mathbf{h}^{(N+1)} \neq \mathbf{h}^{(N)}$, Proposition 2.4 does, of course, not affirm $c_{\mathbf{h}^{(N+1)}} f_{\gamma(\mathbf{h}^{(N+1)})}(x_{\mathbf{h}^{(N+1)}}) = c_{\mathbf{h}^{(N)}} f_{\gamma(\mathbf{h}^{(N)})}(x_{\mathbf{h}^{(N)}})$. As indicated above, we may well have $c_{\mathbf{h}^{(N+1)}} f_{\gamma(\mathbf{h}^{(N+1)})}(x_{\mathbf{h}^{(N+1)}}) > c_{\mathbf{h}^{(N)}} f_{\gamma(\mathbf{h}^{(N)})}(x_{\mathbf{h}^{(N)}})$, i.e., a local reduction step might improve the criterion. In other words, if $\mathbf{h}^{(N+1)}$ is not unique then selections equivalent w.r.t. the weights (9) may differ in the Criterion (6) and we might try to continue the iteration. [2] However, this option is not taken into account since, in the case of continuous distributions, this case has probability zero and since it would require computation of the criterion as soon as the stop signal has appeared; this does not harmonize with the philosophy of the algorithms designed in the present Section 2.5.

2.6. Overall algorithms and randomization

(a) The reduction steps discussed so far may, and often will, get stuck in some non-optimal selection. A simple example with of a local, non-global optimum is this: let $d = 1$, $r = n = 2$, $b_1 = b_2 = 2$, $x_1 = (-1, 3)$, $x_2 = (0, 5)$. Here, besides the global solution $\mathbf{h} = (1, 1)$, the selection $\mathbf{h}' = (2, 2)$ is a local maximum at a much lower level. Therefore, some optimization method that overcomes local maxima must be employed. It turns out that, in the present case, the application of *multistart optimization* to the foregoing iterative reduction steps is sufficient; the limit selection with the best criterion is the proposed approximation.

(b) The reduction steps proposed so far are *greedy* in the sense that they iterate a certain move that optimally exploits but local information. This move may be shortsighted since it cannot take into account later developments. It is well known that, while greedy algorithms serve as a basis for optimization also in the presence of hard problems, their convergence

[2] An example of non-uniqueness is the following: let the regular population be Gaussian, $d = 2$, $r = n = 4$, $b_i = 1$ for $i \neq 4$, $b_4 = 2$, and let $x_{1,1} = (1, 0)$, $x_{2,1} = (-1, 0)$, $x_{3,1} = (0, -1)$, $x_{4,1} = (0, 1)$, $x_{4,2} = \frac{\sqrt{2}}{2}(1, 1)$. Furthermore, let $q_{4,1} = q_{4,2}$. There are two selections; start with $h_4^{(0)} = 1$. Simple computations show that both selections are possible outputs of the reduction step with input $\mathbf{h}^{(0)}$ and that the other has a larger Criterion (6).

is usually too slow in these cases. More efficient procedures for attaining low values add a chaotic component that allows them to violate greediness. One way of accelerating the convergence is randomization. Nowadays standard methods are the Metropolis algorithm [20] and Gibbs sampling [6]. Both are theoretically well established. The former has the drawback of being rather sensitive to the choice of the "temperature" parameter, a positive real number. In many cases, there is a best value that depends on the instance. It seems that there is no practicable rule, let alone an applicable theory, how to find it and the value is usually determined by trial and error. We noticed that the local algorithm combined with Metropolis and multistart yields good results. Another way of introducing a chaotic component is to just omit the current object when the parameters are updated in the single-point reduction step. This modification, too, substantially accelerates the multistart method. It is akin to Gibbs sampling.

(c) The number r_0 of regular elements contained in the data set and, hence, a reasonable input parameter r may be chosen by a test procedure. One performs the estimation for sufficiently many values of $r \leqslant n$. The number of regular objects, r_0, is chosen by validation with a goodness-of-fit test that compares the selection found with the estimated distribution, for instance χ^2 in the normal case, cf. Gallegos and Ritter [4]. We recommend the largest r_0 so that there is sufficient fit. If there is no r_0 that fits well enough then the distributional assumption on the regular population is questionable.

2.7. Using the integrated likelihood

Instead of estimating the combinatorial structure and the parameter γ simultaneously one may first integrate the likelihood w.r.t. a suitable prior measure σ on the parameter space Γ. It remains to maximize the integrated likelihood $c_{\mathbf{h}} \int_\Gamma f_{\mathbf{l},\gamma}(x_{\mathbf{h}})\sigma\,(\mathrm{d}\gamma)$ w.r.t. all selections $\mathbf{h}$. This may be performed by local search combined with multistart and/or Metropolis as above. After the selection has been determined, the parameters may be directly estimated from the regular variants by classical methods. In some cases, we recover the m.l.e. 2.2 in others something else, see the special distributions treated in the following section.

3. Special regular populations

In order to concretize the foregoing theory, it is interesting to discuss it with respect to standard models such as Gaussian families, elliptical symmetries, exponential families, and coin tossing. The symbol $\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}^d$ denotes d-dimensional Euclidean space, and $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. We denote the trace and determinant of a square matrix A by $\operatorname{tr} A$ and $\det A$, respectively. The symbol $N^d_{m,V}$ or just $N_{m,V}$ stands for the d-variate normal distribution with mean vector m and covariance matrix V and also for its Lebesgue density. We introduce also the notation

$$m_1(\mathbf{h}) = \frac{1}{r}\sum_i x_{i,h_i}$$

and

$$V_1(\mathbf{h}) = \frac{1}{r}\sum_i (x_{i,h_i} - m_1(\mathbf{h}))(x_{i,h_i} - m_1(\mathbf{h}))^T$$

for the sample mean vector and sample covariance matrix, respectively, of the cross section $x_{\mathbf{h}}$. As before, summation runs over all objects i regular w.r.t. the selection $\mathbf{h}$.

3.1. The full Gaussian model $N^d_{m_1,V_1}$

Here, (GP_1) means that, for all selections $\mathbf{h}$, the set $\{x_{i,h_i} \mid l_i = 1\} \subseteq \mathbb{R}^d$ contains at least $d+1$ elements in general position. As parameter γ, the pair consisting of the expectation m_1 and the covariance matrix V_1 is appropriate. Moreover, the theory of normal parameter estimation shows that, up to a constant factor, $\max_{(m_1,V_1)} \prod_i N^d_{m_1,V_1}(x_{i,h_i})$ equals $\det V_1(\mathbf{h})^{-\frac{r}{2}}$ with unique optimal estimates $m_1(\mathbf{h})$ and $V_1(\mathbf{h})$. Thus, Criterion (6) becomes $\max_{\mathbf{h}} c_{\mathbf{h}} \det V_1(\mathbf{h})^{-\frac{r}{2}}$. By the properties of the determinant, the optimum selection $\mathbf{h}^\star$ is invariant with respect to affine transformations and the estimates $m_1(\mathbf{h}^\star)$ and $V_1(\mathbf{h}^\star)$ are equivariant. In the absence of irregular variants, but presence of outliers, the present criterion becomes Rousseeuw's [35] minimum covariance determinant, MCD, and the multi-point reduction step reduces to the alternating algorithm proposed in [36,22].

Update formulae for $m_1(\mathbf{h})$ and $V_1(\mathbf{h})^{-1}$ useful in the first parts of the local and single-point reduction steps are presented in Appendix A. We have

$$u_{\mathbf{h}}(i,k) = \text{const}_{\mathbf{h}} - \ln\, b_i q_{i,k} + \tfrac{1}{2}(x_{i,k} - m_1(\mathbf{h}))^T V_1(\mathbf{h})^{-1}(x_{i,k} - m_1(\mathbf{h}))$$

with a constant that depends only on $V_1(\mathbf{h})$. For Part (ii) of the single-point reduction step, the formula

$$\begin{aligned} &u_{\mathbf{h}}(i,k) - u_{\mathbf{h}}(i,h_i) \\ &\quad = \ln\frac{q_{i,h_i}}{q_{i,k}} + \frac{1}{2}\Big((x_{i,k} - m_1(\mathbf{h}))^T V_1(\mathbf{h})^{-1}(x_{i,k} - m_1(\mathbf{h})) \\ &\qquad - (x_{i,h_i} - m_1(\mathbf{h}))^T V_1(\mathbf{h})^{-1}(x_{i,h_i} - m_1(\mathbf{h}))\Big) \\ &\quad = \ln\frac{q_{i,h_i}}{q_{i,k}} + \frac{1}{2}(x_{i,k} - x_{i,h_i})^T V_1(\mathbf{h})^{-1}(x_{i,k} + x_{i,h_i} - 2m_1(\mathbf{h})) \end{aligned}$$

is useful.

We next determine the Bayesian estimator of the selection w.r.t. Jeffreys' invariant prior measure $\sigma(\mathrm{d}m_1, \mathrm{d}V_1) = \mathrm{d}m_1 \det V_1^{-\frac{d+2}{2}} \mathrm{d}V_1$. According to (5),

$$f_{L,T,X}(\mathbf{l}, \pi_1^n, \mathbf{x}) = C c_{\mathbf{h}} \int f_{m_1,V_1}(x_{\mathbf{h}})\sigma(\mathrm{d}m_1, \mathrm{d}V_1).$$

If W denotes the scatter matrix of $z_1, \ldots, z_r \in \mathbb{R}^d$, standard computations using Steiner's formula first show

$$f_{m_1,V_1}(z_1^r) = r^{-\frac{d}{2}} \det(2\pi V_1)^{-\frac{r}{2}+\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\operatorname{tr} V_1^{-1} W\right\} N^d_{0,V_1/r}(\bar{z} - m_1).$$

Integration w.r.t. m_1 removes the last factor so that, after multiplication with $\det V_1^{-\frac{d+2}{2}}$ and up to a constant factor, the Lebesgue density of an inverted Wishart distribution with parameter r remains. Integration w.r.t. V_1 thus shows that the Bayesian estimator coincides with the weighted m.l.e. $\operatorname{argmax}_{\mathbf{h}} c_{\mathbf{h}} \det V_1(\mathbf{h})^{-r/2}$ derived from Theorem 2.2(a) for the present statistical model.

A simple example with two variants and $Z_i \sim N_{0,1} \otimes \mathrm{UNI}_{[-10,10]}$ convinces the reader that the m.l.e. of the parameters of the regular population is not consistent. However, the inconsistency is controlled and vanishes gradually as the population of irregular variants is more and more spread in space and the actual number of outliers does not exceed $n - r$. These are the contents of our next theorem which applies to the location model (A) stated at the beginning of Section 2.

3.2 Theorem. *Assume that regular objects i are defined by $Z_{i,1} \sim N^d_{m_1,V_1}$ and $Z_{i,\widehat{1}} = \kappa_1 U_i + \psi_i$ with some centered, square integrable, Lebesgue continuous random vector $U_i : \Omega \to \mathbb{R}^{(b_i-1)d}$ and that irregular objects i are defined by $Z_i = \kappa_0 U_i + \psi_i$ with analogous random vectors $U_i : \Omega \to \mathbb{R}^{b_i d}$.*

(a) *Let $3d+1 \leqslant r \leqslant n$ and let Z_i, $1 \leqslant i \leqslant n$, be an independent array of regular and irregular objects. If it contains at least r regular objects then, P-a.s., the m.l. selector $\mathbf{H}^\star(n)$ w.r.t. the general normal model selects r regular objects and their first variants as regular if $\kappa_1, \kappa_0 \geqslant \kappa(n, \omega)$ are large enough.*

(b) *Let Z_i, $i \geqslant 1$, be an independent sequence of regular and irregular objects. If the number $r_0(n)$ of regular objects among $1..n$ tends to infinity as $n \to \infty$ and if the parameter $r = r_0(n)$ is chosen for estimation with the objects $1..n$ for all n then, P-a.s.,*

$$\lim_{n\to\infty} \lim_{\kappa_1,\kappa_0\to\infty} m_1(\mathbf{H}^\star(n)) = m_1 \quad \text{and} \quad \lim_{n\to\infty} \lim_{\kappa_1,\kappa_0\to\infty} V_1(\mathbf{H}^\star(n)) = V_1.$$

(c) *If Z_i, $i \geqslant 1$, is an independent sequence of regular objects and the estimation is understood without outliers (that is, $r = n$) then the limits in* (b) *are again valid.*

The reader may wonder why we require $r = r_0(n)$ instead of $r < r_0(n)$ in part (b). The hypothesis $r = r_0(n)$ there guarantees that the supports of the selections form an increasing sequence as n grows. Therefore, the strong law may be applied to their union. In the case $r < r_0(n)$, the supports would be random r-element subsets of $1..n$.

Theorem 3.2 may be generalized by replacing the parameter κ_1 with d by d diagonal matrices D_h with diagonal entries $\kappa_{1,h,1}, \ldots, \kappa_{1,h,d}$, $h \in 2..b$. Put

$$Z_{i,h} = D_h U_{i,h} + \psi_{i,h}, \quad h \in 2..b_i.$$

The theorem remains true if we require $r \geqslant (b+2)d + 1$ and that the diagonal elements of each matrix converge to infinity with their quotients bounded.

3.3. The spherical Gaussian model $N^d_{m_1,v_1\mathbf{I}_d}$

We next assume that the regular variants belong to a normal population with unknown mean vector $m_1 \in \mathbb{R}^d$ and *spherical* covariance matrix $v_1\mathbf{I}_d$ of unknown size v_1 (> 0). Here, (GP_1) is satisfied if and only if any cross section of r elements in the data set contains at least two different variants. As parameter γ, the pair consisting of the expectation

$m_1 \in \mathbb{R}^d$ and the size v_1 of the covariance matrix is appropriate. Moreover, the theory of normal parameter estimation shows that, up to a constant factor, $\max_{(m_1,v_1)} \prod_i N^d_{m_1,v_1\mathbf{I}_d}(x_{i,h_i})$ equals $\operatorname{tr} V_1(\mathbf{h})^{-\frac{rd}{2}}$ with unique optimal parameters $m_1(\mathbf{h})$ and $v_1(\mathbf{h}) = \operatorname{tr} V_1(\mathbf{h})/d$. Thus, Criterion (6) becomes $c_\mathbf{h} \operatorname{tr} V_1(\mathbf{h})^{-\frac{rd}{2}}$, here. An update formula for $m_1(\mathbf{h})$ useful in the first parts of the local and single-point reduction steps appears in Appendix A. From (14), it also follows

$$\begin{aligned}\operatorname{tr} V_1(\mathbf{h}') - \operatorname{tr} V_1(\mathbf{h}) &= \frac{1}{r}\left(\|x_{j,k}\|^2 - \|x_{i,h_i}\|^2\right) - \left(\|m_1(\mathbf{h}')\|^2 - \|m_1(\mathbf{h})\|^2\right)\\ &= \frac{1}{r}\left(\|x_{j,k}\|^2 - \|x_{i,h_i}\|^2\right)\\ &\quad - \left(2m_1(\mathbf{h}) + \frac{1}{r}(x_{j,k} - x_{i,h_i}), \frac{1}{r}(x_{j,k} - x_{i,h_i})\right),\end{aligned}$$

where $\mathbf{h}'$ is the neighbor of $\mathbf{h}$ defined by inserting $k \neq h_i$ as the regular variant of the regular object $i = j$ *or* by declaring the outlier j as regular with regular variant k and i as an outlier. This is an efficient way of updating the trace $\operatorname{tr} V_1(\mathbf{h})$ since the quantity $\frac{1}{r}(x_{j,k} - x_{i,h_i})$ is the increment of the mean already computed and the norms may be computed at the very beginning and stored. The computational cost is about $3d$ elementary operations. We also have

$$u_\mathbf{h}(i,k) = \text{const}_\mathbf{h} - \ln b_i q_{i,k} + \frac{d}{2\operatorname{tr} V_1(\mathbf{h})}\|x_{i,k} - m_1(\mathbf{h})\|^2$$

with a constant that depends only on $v_1(\mathbf{h})$. The expression

$$\begin{aligned}&u_\mathbf{h}(i,k) - u_\mathbf{h}(i,h_i)\\ &\quad= \ln\frac{q_{i,h_i}}{q_{i,k}} + \frac{d}{2\operatorname{tr} V_1(\mathbf{h})}\left[\|x_{i,k} - m_1(\mathbf{h})\|^2 - \|x_{i,h_i} - m_1(\mathbf{h})\|^2\right]\\ &\quad= \ln\frac{q_{i,h_i}}{q_{i,k}} + \frac{d}{2\operatorname{tr} V_1(\mathbf{h})}(x_{i,k} - x_{i,h_i})^T(x_{i,k} + x_{i,h_i} - 2m_1(\mathbf{h}))\end{aligned}$$

for the weight difference is useful in Part (ii) of the single-point reduction step.

As in the full Gaussian model, the Bayesian estimator of the selection w.r.t. Jeffreys' invariant prior measure

$$\sigma(\mathrm{d}m_1, \mathrm{d}v_1) = \mathrm{d}m_1 v_1^{-\frac{d}{2}-1}\,\mathrm{d}v_1$$

coincides with the weighted m.l.e. derived from Theorem 2.2. It reads here

$$\operatorname{argmax}_\mathbf{h} c_\mathbf{h} \operatorname{tr} V_1(\mathbf{h})^{-rd/2}.$$

The following theorem is the analogue of Theorem 3.2 for the spherical normal model.

3.4 Theorem. *Theorem* 3.2 *remains valid if* $v_1\mathbf{I}_d$ *is substituted for* V_1 *and estimation is w.r.t. the spherical normal model. In part* (a), $r \geqslant 4$ *is sufficient.*

Roughly speaking, Theorems 3.2(c) and 3.4(c) state that, if n is kept fixed and large enough, then the two estimates are close to the correct parameters if the irregular populations are sufficiently diffuse in space, the degree depending on n. The following theorem

differs from these statements in that the order of the two limits is reversed. It states that, if the diffusion of the irregular population is kept fixed at a sufficiently high level, then the estimates are close to the correct values for all sufficiently large n. We restrict matters to data sets without outliers, $r = n$, here.

3.5 Theorem. *Let $(Z_i)_{i\geqslant 1}$ be an independent sequence of regular objects satisfying* ($\mathrm{SV_r}$), *and assume that the sequence $(b_i)_i$ of numbers of variants is generated by tossing a (possibly) biased, b-sided coin distributed according to $(s_1, \ldots, s_b)$. Assume that $Z_{i,1}$, $i \geqslant 1$, is d-dimensional standard normal and that all subsequences $(Z_i)_{b_i=j}$, $j \in 1..b$, are identically distributed. Then, for all $\varepsilon > 0$, there exists $\eta > 0$ such that*

$$\limsup_n \|m_1(\mathbf{H}^*(n))\| \leqslant \varepsilon \quad \text{and} \quad 1-\varepsilon \leqslant \liminf_n v_1(\mathbf{H}^*(n)) \leqslant \limsup_n v_1(\mathbf{H}^*(n)) \leqslant 1,$$
$$P\text{-a.s.},$$

P-a.s., if $f_{Z_{i\widehat{1}}}[z_{i\widehat{1}} \mid Z_{i1} = z_{i1}] \leqslant \eta$ for all $z \in \mathbb{R}^{b_i d}$.

3.6. Elliptical symmetry

The tail of the normal distribution is too light to fit some real-world distributions occurring in practical applications; cf., e.g., [31]. A remedy is the use of elliptically symmetric distributions. They are specified by three quantities: a radial function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ describing the tail decay, usually strictly decreasing, the mean vector m_1, and the covariance matrix V_1. The radial function is normalized and assumed to be given. The density is

$$f_{Z_{i,1}}(z) = \det V_1^{-1/2} \varphi\left(\sqrt{(z-m_1)^T V_1^{-1}(z-m_1)}\right).$$

Three cases may be distinguished: $-\ln\varphi$ convex, linear, or concave. In the normal case, $-\ln\varphi(s) = s^2/2 + \text{const}$ is convex, in the exponential case, $-\ln\varphi(s) = \lambda s + \text{const}$ is linear, and for Pearson's type-VII family, e.g., $-\ln\varphi(s) = \frac{\lambda}{2}\ln(1+s^2/\eta)$ is concave. The last-mentioned family has the heaviest tail.

For the parameter γ, we choose again the pair (m_1, V_1) as in the full normal case. Maronna [19] showed that the ML-estimators of these two parameters exist under the same assumption made for the normal family. In general, its computation is much more demanding and requires a fixed-point algorithm. Since the computational cost of ML_{m_1} and ML_{V_1} cannot be neglected, the naive estimators $m_1(\mathbf{h})$ and $V_1(\mathbf{h})$, which result from the method of moments, may be employed as a heuristic, instead.

As in the normal case, the optimum selection $\mathbf{h}^\star$ is invariant with respect to affine transformations and the estimates of m_1 and V_1 are equivariant. Finally, we have

$$u_{\mathbf{h}}(i,k) = \text{const}_{\mathbf{h}} - \ln b_i q_{i,k} - \ln\varphi\left(\sqrt{(x_{i,k}-\mathrm{ML}_{m_1}(x_{\mathbf{h}}))^T \mathrm{ML}_{V_1^{-1}}(x_{\mathbf{h}})(x_{i,k}-\mathrm{ML}_{m_1}(x_{\mathbf{h}}))}\right).$$

with a constant that depends only on $V_1(\mathbf{h})$.

3.7. Exponential families

A Lebesgue-continuous exponential family with natural parameter $\gamma \in \Gamma \subseteq \mathbb{R}^q$ is specified by the Lebesgue densities

$$f_\gamma(z) = c_\gamma^{-1} e^{-\gamma\cdot S(z)},$$

where $S : \mathbb{R}^d \to \mathbb{R}^q$ is the generating statistic of the family and $c_\gamma = \int_{\mathbb{R}^d} e^{-\gamma \cdot S(z)}\, \mathrm{d}z$ is the partition function, the normalizing constant. For details, we refer the reader to the monography [2].

All members f_γ of an exponential family are measure-theoretically equivalent. Let R be their joint support and assume that the exponential family has full rank. The closed convex hull $C = \overline{\mathrm{conv}}\, R$ of R plays a key role. The ML-estimate of γ for z exists if and only if $S(z)$ lies in its interior. Therefore, (GP_1) is satisfied if and only if, for any selection $\mathbf{h}$, the cross section $x_\mathbf{h}$ belongs to this interior. If, moreover, the cumulant transform $\ln c(\gamma)$ is steep then the m.l.e. $\gamma(\mathbf{h})$ is determined by the moment equation $E_{\gamma(\mathbf{h})} S = S(x_\mathbf{h})$. Moreover, we have

$$u_\mathbf{h}(i,k) = \mathrm{const}_\mathbf{h} - \ln b_i q_{i,k} + \gamma(\mathbf{h}) S(x_{i,k})$$

with a constant that depends only on $\gamma(\mathbf{h})$.

The Bayesian estimator of the selection $\mathbf{h}$ w.r.t. the conjugate prior density with parameter $s \in \mathbb{R}^q$, $f_s(\gamma) = c_\gamma^{-1} e^{-\gamma \cdot s}$, is given by the criterion

$$\mathrm{argmax}_\mathbf{h}\, c_\mathbf{h} \int_\Gamma c_\gamma^{-2} f_{s+S(x_\mathbf{h})}(\gamma)\, \mathrm{d}\gamma.$$

3.8. A coin-tossing model

We next consider a discrete model with sample space $E = (1..s)^d$. The regular variants are generated by tossing d independent, possibly biased, s-sided coins. Here, condition (GP_1) is always satisfied. The parameter γ is an $s \times d$ table $\mathbf{p}$ of real numbers $p_{y,m} \geqslant 0$ whose columns sum to 1. Each variant $x \in E$ generates a path in this table that visits each column exactly once. Its probability is the product of the entries along the path,

$$f_\mathbf{p}(x) = \prod_{m=1}^{d} p_{x_m,m}.$$

Given a selection $\mathbf{h}$, let $n_{y,m}(\mathbf{h}) = \#\{i \mid x_{i,h_i,m}{=}y\}$ be the frequency of the outcome y at entry m taken over the r selected variants. The frequencies sum up to rd. The m.l.e. of the table consists of the relative frequencies $n_{y,m}(\mathbf{h})/r$, $y \in 1..s$, $m \in 1..d$, and, up to a multiplicative constant, the maximum value of the log-likelihood function, $\max_\mathbf{p} \ln f_\mathbf{p}(x_\mathbf{h})$, equals their negative entropy. Thus, the ML-estimate of $\mathbf{h}$ is the selection that minimizes this entropy. The quantities $u_\mathbf{h}(i,k)$ become

$$u_\mathbf{h}(i,k) = -\ln b_i q_{i,k} - \sum_m \ln \frac{n_{x_{i,k,m},m}(\mathbf{h})}{r}.$$

Reasonable methods for optimizing the likelihood are proposed in 2.6(b), in particular multistart iteration of local reduction steps combined with Metropolis. It means replacing in $u_\mathbf{h}$ the relative frequencies $n_{y,m}/r$ with $n'_{y,m}/(r-1)$, where the prime indicates omission of the "current" line. Note that the probability estimates are now based on $r-1$ observations. It becomes even more efficient if, instead of the maximum likelihood, Laplace's Law of Succession is used for estimating the probabilities $p_{y,m}$. Then the numbers $(n'_{y,m} + 1)/$

$(r-1+s)$ replace the relative frequencies $n'_{y,m}/(r-1)$. In the extreme case of a data set consisting of one line one obtains the unbiased prior $1/s$.

Let us compare the criterion obtained in Theorem 2.2 with the MAP-criterion based on the integrated likelihood. In view of (5), we have with the notation above

$$f_{\mathbf{p}}(x_{\mathbf{h}}) = \prod_{i=1}^{n}\prod_{m=1}^{d} p_{x_{i,h_i,m},m} = \prod_{m=1}^{d}\prod_{y\in 1..s} p_{y,m}^{n_{y,m}(\mathbf{h})}. \tag{13}$$

Geometrically, the parameter space is the d-fold product of $(s-1)$-dimensional unit simplices. It is reasonable to endow it with the uniform prior. Up to the factor $s!^d$, the integral of (13) w.r.t. this prior equals

$$\int f_{\mathbf{p}}(x_{\mathbf{h}})\,\mathrm{d}\mathbf{p} \propto \prod_{m=1}^{d}\int\prod_{y=1}^{s} p_{y,m}^{n_{y,m}(\mathbf{h})}\,\mathrm{d}p_m.$$

Integration on the right-hand side extends over the $(s-1)$-dimensional unit simplex $\{p_m \mid \sum_y p_{y,m}\leqslant 1,\ \ p_{y,m}\geqslant 0\}$. Iterated integration of the beta function w.r.t. $p_{y,m}$ shows that these integrals equal $\frac{1}{(r+s)!}\prod_{y=1}^{s} n_{y,m}(\mathbf{h})!$. Hence, the integrated likelihood is

$$f_{T,X}(\pi_1^n, \mathbf{x}) \propto c_{\mathbf{h}}\int f_{\mathbf{p}}(x_{\mathbf{h}})\,\mathrm{d}\mathbf{p} \propto c_{\mathbf{h}}\prod_{m=1}^{d}\prod_{y=1}^{s} n_{y,m}(\mathbf{h})!.$$

This criterion differs from Theorem 2.2 which, in the present case, is the exponential of the negative entropy.

4. Two studies

4.1. A simulation study with normal data

This simulation study applies the multi-point algorithm to four simulated five-dimensional data sets, each consisting of 500 objects with up to four variants. The regular variants are drawn from $N^5_{m_1,\mathbf{I}_5}$ with $m_1 = 0,\ 2e_1$ and the irregular variants are deterministic and defined by $x_{i,h,k} = \alpha\sin(ihk)$, $i \in 1..500$, $h \in 1..b_i$, $k \in 1..5$, and $\alpha = 3,\ 5$. In all cases, the best target value is reached in one iteration of three or four reduction steps. Our implementation in C++ requires 0.01 s on a 1 GHz processor. No more than 10% of the estimated regular variants were actually generated as irregular ones; they happened to lie close to the center of the regular population. The parameters of the estimated regular variants are shown in Table 2.

4.2. A problem from functional genetics

We consider the well-known problem of *motif discovery* in genetics. It has received much attention since the genomes of a number of species, among them *Homo sapiens*, have become known in recent years. The genomes of pro- and eukaryotic species are DNA

Table 2
Means and scatter matrices of the optimal solutions attained for the test runs described in Section 4.1

	$m_1 = [0 \quad 0 \quad 0 \quad 0 \quad 0]$	$m_1 = [2 \quad 0 \quad 0 \quad 0 \quad 0]$
	$[0.061 \quad 0.014 \quad -0.013 \quad 0.062 \quad -0.046]$	$[1.977 \quad 0.007 \quad -0.013 \quad 0.050 \quad -0.043]$
$\alpha = 3$	$\begin{bmatrix} 0.884 & & & & \\ 0.023 & 0.942 & & & \\ 0.087 & 0.007 & 0.986 & & \\ -0.043 & 0.042 & -0.047 & 0.955 & \\ 0.037 & -0.013 & 0.075 & -0.019 & 0.992 \end{bmatrix}$	$\begin{bmatrix} 1.074 & & & & \\ 0.020 & 0.970 & & & \\ 0.058 & 0.002 & 1.001 & & \\ -0.051 & 0.035 & -0.048 & 0.946 & \\ 0.014 & -0.013 & 0.038 & -0.020 & 0.953 \end{bmatrix}$
	$[0.047 \quad -0.002 \quad -0.014 \quad 0.046 \quad -0.047]$	$[2.000 \quad 0.002 \quad -0.013 \quad 0.042 \quad -0.047]$
$\alpha = 5$	$\begin{bmatrix} 0.945 & & & & \\ 0.031 & 0.979 & & & \\ 0.067 & 0.005 & 0.998 & & \\ -0.065 & 0.032 & -0.049 & 0.959 & \\ 0.035 & -0.002 & 0.046 & -0.024 & 0.970 \end{bmatrix}$	$\begin{bmatrix} 1.040 & & & & \\ 0.025 & 0.992 & & & \\ 0.056 & 0.007 & 1.015 & & \\ -0.046 & 0.021 & -0.038 & 0.965 & \\ 0.022 & 0.002 & 0.035 & -0.013 & 0.963 \end{bmatrix}$

strings made up of many copies of the purine-pyrimidine base pairs (adenine, thymine) (a,t) and (guanine, cytosine) (g,c). They contain the genes as special substrings. The genes hold the coding information for the production of specific proteins. Protein synthesis involves an intermediate step, the *transcription* of DNA into "messenger" RNA. It is catalyzed by an enzyme, the RNA polymerase. The spaces between genes are believed to have mainly regulatory function. The regions preceding the genes, the promoters, contain small specific gene precursor segments that serve as binding sites for special proteins, transcription factors that enhance or inhibit transcription and for the RNA polymerases. Among these segments are the classical boxes, for instance, the Pribnow box TATAAT in prokaryotes located 10 base pairs upstream of the initiation point of transcription and the perfect palindrome [3] TATA in eukaryotes. In few instances, these motifs are fully reproduced, in the others at least to a high degree. The monography [7] offers a good account of the genetic background.

The binding of a biomolecule and a DNA segment is a complex process governed by the 3D molecular shapes, by van der Waals, electrostatic, and hydrophobic forces and by hydrogen bonds. It is surprising that the binding sites can be captured to a high degree by simple statistical models. About 20 years ago, a motif in the genome of *Escherichia coli* aroused the interest of some statisticians and computational biologists: the binding sites for the cyclic AMP receptor protein (CRP) in the promoters of 18 genes of *E. coli*, [38]. Experimental methods have identified 24 CRP-binding sites of length $d = 22$, all approximate representatives of their consensus pattern ***tgtga******tcaca***, a perfect palindrome, see Table 3. The central segment of length 16 contains two highly conserved regions of length five, each. The asterisks indicate positions of little sequence preference. Although these sites cannot be discerned by the human eye, it is possible to characterize and locate them by statistical and computational means.

[3] A genetic pattern is called a palindrome if its reverse equals its complement defined by swapping a $\leftrightarrow$ t, c $\leftrightarrow$ g.

Table 3
The regions of length 105 base pairs upstream of 18 genes in E. coli

t a a t g t t t g t g c t g g t **t t t t g t g g c a t c g g g c g a g a a t** a g c g c g t g g t g t g a a a g a c t g t **t t t t t t g a t c g t t t t c a c a a a a** a t g g a a g t c c a c a g t c t t g a c a g
g a c a a a a a a c g c g t a a c **a a a a g t g t c t a t a a t c a c g g c a** g a a a a g t c c a c a t t g a **t t a t t t g c a c g g c g t c a c a c t t** t g c t a t g c c a t a g c a t t t t t a t c c a t a a g
a c a a a t c c c a a t a a c t t a a t t a t t g g g a t t t g t t a t a t a t a a c t t t a t a a a t t c c t a a a a t t a c a c a a a g t t a a t **a a c t g t g a g c a t g g t c a t a t t t** t t a t c a a t
c a c a a a g c g a a a g c t a t g c t a a a a c a g t c a g g a t g c t a c a g t a a t a c a t t g a t g t a c t g c a t **g t a t g c a a a g g a c g t c a c a t t a** a c c g t g c a g t a c a g t t g a t a g c
a c g g t g c t a c a c t t g t a t g t a g c g c a t c t t t c t t t a c g g t c a a t c a g c a **t g g t g t t a a a t t g a t c a c g t t t t** t a g a c c a t t t t t t c g t c g t g a a a c t a a a a a a a a c c
a g t g a a **t t a t t t g a a c c a g a t c g c a t t a** c a g t g a t g c a a a c t t g t a a g t a g a t t t c c t t **a a t t g t g a t g t g t a t c g a a g t g** t g t t g c g g a g t a g a t g t t a g a a t a
g c g c a t a a a a a a c g g c t a a a t t c t t g t g t a a a c g a t t c c a c **t a a t t t a t t c c a t g t c a c a c t t t** t c g c a t c t t t g t t a t g c t a t g g t t a t t t c a t a c c a t a a g c c
g c t c c g g c g g g g g t t t t t t g t t a t c t g c a a t t c a g t a c a **a a a c g t g a t c a a c c c c t c a a t t t** t c c c t t t g c t g a a a a a t t t t c c a t t g t c t c c c c t g t a a a g c t g t
a a c g c a a t **t a a t g t g a g t t a g c t c a c t c a t** t a g g c a c c c c a g g c t t t a c a c t t t a t g c t t c c g g c t c g t a t g t t g t g t g **g a a t t g t g a g c g g a t a a c a a t t** t c a c
a c a t t a c c g c c a a **t t c t g t a a c a g a g a t c a c a c a a** a g c g a c g g t g g g g c g t a g g g g c a a g g a g g a t g g a a a g a g g t t g c c g t a t a a a g a a a c t a g a g t c c g t t t a
g g a g g a g g c g g g a g g a t g a g a a c a c g g c **t t c t g t g a a c t a a a c c g a g g t c** a t g t a a g g a a t **t t t c g t g a t g t t g c t t g c a a a a** a t c g t g g c g a t t t t a t g t g c g c a
g a t c a g c g t c g t t t t a g g t g a g t t g t t a a t a a a g a t t t g g **a a t t g t g a c a c a g t g c a a a t t c** a g a c a c a t a a a a a a a c g t c a t c g c t t g c a t t a g a a a g g t t t c t
g c t g a c a a a a a a g a t t a a a c a t a c c t t a t a c a a g a c t t t t t t t t c a t **a t g c c t g a c g g a g t t c a c a c t t** g t a a g t t t t c a a c t a c g t t g t a g a c t t t a c a t c g c c
t t t t t t a a a c a t t a a a a t t c t t a c g t a a t t t a t a a t c t t t a a a a a a a g c a t t t a a t a t t g c t c c c g a a c **g a t t g t g a t t c g a t t c a c a t t t** a a a c a a t t t c a g a
c c c a t g a g a g t g a a a t **t g t t g t g a t g t g g t t a a c c c a a** t t a g a a t t c g g g a t t g a c a t g t c t t a c c a a a a g g t a g a a c t t a t a c g c c a t c t c a t c c g a t g c a a g c
c t g g c t t a a c t a t g c g g c a t c a g a g c a g a t t g t a c t g a g a g t g c a c c a t a t g **c g g t g t g a a a t a c c g c a c a g a t** g c g t a a g g a g a a a a t a c c g c a t c a g g c g c t c
c t g t g a c g g a a g a t c a c t t c g c a g a a t a a a t a a a t c c t g g t g t c c c t g t t g a t a c c g g g a a g c c c t g g g c c a a c t t t t g g c g a **a a a t g a g a c g t t g a t c g g c a c g**
g a t t t t t a t a c t t t a a c t t g t t g a t a t t t a a a g g t a t t t a a t t g t a a t a a c g a t a c t c t g g a a a g t a t t g a a a g t t a **a t t t g t g a g t g g t c g c a c a t a t** c c t g t t

The data set contains 24 experimentally determined binding sites of length 22 bases (shown in boldface) for the transcription factor CRP that helps to control the gene's expression in the cell. The original version is from [38], this extended table is reproduced from [14]. (The first two base pairs of the first substring of line 17 are missing and the spacing between the two highly conserved regions is 7 bp instead of 6; see also the comment in [38]).

Table 4
Table of estimated probabilities of the central 16 bases for the ML- and the MAP-estimator with integrated likelihood

a	0.00	0.00	0.06	0.11	0.89	0.33	0.33	0.06	0.39	0.06	0.28	0.00	0.06	0.78	0.06	0.78
c	0.11	0.11	0.06	0.06	0.11	0.28	0.22	0.22	0.00	0.22	0.22	0.00	0.83	0.00	0.83	0.11
g	0.00	0.72	0.00	0.78	0.00	0.17	0.28	0.33	0.22	0.56	0.17	0.22	0.00	0.22	0.06	0.06
t	0.89	0.17	0.89	0.06	0.00	0.22	0.17	0.39	0.39	0.17	0.33	0.78	0.11	0.00	0.06	0.06

As indicated in the Introduction, this problem may be viewed in the light of variant selection. We estimate the starting positions of the central segments of length 16 of the binding sites. It is sufficient to extract all 90 substrings of length 16 from each of the 18 strings as variants. We use the standard model of inhomogeneous, independent, biased four-sided coins discussed in Section 3.8 in a slightly more general context for the regular variants. Since there is no prior information on the positions of the binding sites, a natural choice for the priors q_i is uniformity. Each of the algorithms of Sections 2.3 and 2.5 may be applied in order to estimate the 4 by 16 table of parameters

$$\begin{pmatrix} p_{a,1} \cdots p_{a,16} \\ p_{c,1} \cdots p_{c,16} \\ p_{g,1} \cdots p_{g,16} \\ p_{t,1} \cdots p_{t,16}. \end{pmatrix}.$$

Each column sums to 1 and corresponds to a coin and an entry in the motif. Our model actually finds those substrings as regular variants that produce the minimum empirical entropy. A similar model was already used in this context in [38]. However, our algorithms are different. They, too, readily find 16 of the experimentally determined binding sites in 18 rows, missing two and, of course, the double occurrences. Note that our algorithms are also applicable to strings of different lengths.

The estimated regular variants w.r.t. ML and the MAP-criterion with the integrated likelihood coincide. Hence, the same is true for the probability tables. The common table is Table 4. The maximum entries in each column yield the correct consensus sequence. Both estimators erroneously determine the two patterns starting from base pairs 54 and 23 in lines 7 and 8, respectively, as binding sites. Dedicated methods based on finer statistical models that take into account also the background noise identify between 16 and all observed binding sites. The authors of [11–13,16] propose Gibbs-sampling strategies, Bailey and Elkan [1] design a clustering algorithm, and an overview is contained in [14].

After the table of parameters has been estimated, other occurrences of the motif can be located as segments of high probability.

Acknowledgments

We thank the referees, whose comments much improved the paper. We are particularly grateful to the referee who pointed out the genetic example to us.

Appendix A. General update formulae

We compile here some update formulae that are basic for the efficiency of the single-point algorithms. Let $\mathbf{x} = (x_1, \ldots, x_r)$ be a data set and let $\mathbf{x}_y = (x_1, \ldots, x_r, y)$ be obtained from $\mathbf{x}$ by attaching one more point y. It is well known and easily verified that the update of the sample mean is $m(\mathbf{x}_y) = m(\mathbf{x}) + \frac{1}{r+1}(y - m(\mathbf{x}))$ and that of the SSP matrix is $W(\mathbf{x}_y) = W(\mathbf{x}) + \frac{r+1}{r}(y - m(\mathbf{x}_y))(y - m(\mathbf{x}_y))^T$. Let $\mathbf{h}'$ be the neighbor of $\mathbf{h}$ in the selection graph, Section 2.3, defined by inserting $k \neq h_i$ as the regular variant of the regular object $i = j$ *or* by declaring the outlier j as regular with regular variant k and i as an outlier. Two applications of each of these formulae yield the updates of the sample mean and the sample covariance matrix for the transition from $\mathbf{h}$ to a $\mathbf{h}'$.

(a) *Sample mean*:

$$m_1(\mathbf{h}') = m_1(\mathbf{h}) + \frac{1}{r}(x_{j,k} - x_{i,h_i}). \tag{14}$$

(b) *Sample covariance*:

$$\begin{aligned} &V_1(\mathbf{h}') - V_1(\mathbf{h}) \\ &\quad = \frac{1}{r-1}(x_{j,k} - m_1(\mathbf{h}'))(x_{j,k} - m_1(\mathbf{h}'))^T \\ &\qquad - \frac{1}{r-1}(x_{i,h_i} - m_1(\mathbf{h}))(x_{i,h_i} - m_1(\mathbf{h}))^T. \end{aligned} \tag{15}$$

(c) *Inverse of sample covariance*:

Let A be a regular d by d matrix, let $b \in \mathbb{R}^d$, and let $\gamma \in \mathbb{R}$. According to [18], A.2.4(V), if the inverse of $(A + \gamma bb^T)^{-1}$ exists, then it is computed from A^{-1} and b in the following way:

$$(A + \gamma bb^T)^{-1} = A^{-1} - \frac{1}{\frac{1}{\gamma} + b^T(A^{-1}b)}(A^{-1}b)(A^{-1}b)^T. \tag{16}$$

The number of elementary operations (+,–,*,/) needed for (16) is $3\,d^2 + \mathcal{O}(d)$. According to (15), a double application of this identity yields an update formula for the inverse of the new sample covariance matrix $V_1(\mathbf{h}')$. Therefore, the cost of computing $V_1(\mathbf{h}')^{-1}$ from $V_1(\mathbf{h})^{-1}$ is $6\,d^2 + \mathcal{O}(d)$.

(d) *Determinant of sample covariance*:

The update is obtained from (15) and (16) again by a double application of the identity

$$\det(A + \gamma bb^T) = (1 + \gamma b^T(A^{-1}b)) \det A.$$

There is essentially no additional computational cost after the inverse has been computed.

Appendix B. Proofs

Proof of Lemma 2.1. Regularity of T_i for a regular object i implies

$$P[T_i = \pi_i \mid L_i = 1] = \frac{q_{i,h_i}}{(b_i - 1)!}$$

for all π_i such that $\pi_i^{-1}(1) = h_i$. On the other hand,

$$P[T_i = \pi_i \mid L_i = 0] = \frac{1}{b_i!}.$$

By independence of the family $(T_i)_i$ given L and by the independence of the permutations T_i and the $L_{\hat{i}}$'s given L_i, we have

$$\begin{aligned} P[T = \pi \mid L = \mathbf{l}] &= \prod_{i=1}^{n} P[T_i = \pi_i \mid L = \mathbf{l}] = \prod_{i=1}^{n} P[T_i = \pi_i \mid L_i = l_i] \\ &= \prod_{i:l_i=1} \frac{q_{i,h_i}}{(b_i - 1)!} \prod_{i:l_i=0} \frac{1}{b_i!} = C\, c_{\mathbf{h}}. \qquad \square \end{aligned}$$

Proof of Theorem 2.2. Since there are only finitely many sequences $\mathbf{h}$ and by assumption (GP_1), we may apply the Principle of Dynamic Optimization to the weighted likelihood (5) to write

$$\max_{\mathbf{l},\mathbf{h},\gamma} c_{\mathbf{h}} f_{\mathbf{l},\gamma}(x_{\mathbf{h}}) = \max_{\mathbf{l},\mathbf{h}} c_{\mathbf{h}} \max_{\gamma} f_{\mathbf{l},\gamma}(x_{\mathbf{h}}) = \max_{\mathbf{l},\mathbf{h}} c_{\mathbf{h}} f_{\mathbf{l},\gamma(\mathbf{h})}(x_{\mathbf{h}}). \qquad \square$$

Proof of Proposition 2.4. The first inequality in the chain

$$c_{\mathbf{h}_{\text{new}}} f_{\gamma(\mathbf{h}_{\text{new}})}(x_{\mathbf{h}_{\text{new}}}) \geqslant c_{\mathbf{h}_{\text{new}}} f_{\gamma(\mathbf{h})}(x_{\mathbf{h}_{\text{new}}}) \geqslant c_{\mathbf{h}} f_{\gamma(\mathbf{h})}(x_{\mathbf{h}}) \tag{17}$$

follows from m.l. estimation $f_{\gamma(\mathbf{h}_{\text{new}})}(x_{\mathbf{h}_{\text{new}}}) \geqslant f_{\gamma(\mathbf{h})}(x_{\mathbf{h}_{\text{new}}})$ and the second is just inequality (8). Hence, Parts (a) and (b).

If there is equality in (a) then the first inequality in (17) is an equality. Hence, $f_{\gamma(\mathbf{h}_{\text{new}})}(x_{\mathbf{h}_{\text{new}}}) = f_{\gamma(\mathbf{h})}(x_{\mathbf{h}_{\text{new}}})$ and the first part of Claim (c) follows from the uniqueness assumption. If $f_{\text{ML}_\gamma(z)}(z) = \max_\gamma f_\gamma(z)$ depends only on $\text{ML}_\gamma(z)$ then $\gamma(\mathbf{h}_{\text{new}}) = \gamma(\mathbf{h})$ implies $f_{\gamma(\mathbf{h}_{\text{new}})}(x_{\mathbf{h}_{\text{new}}}) = f_{\gamma(\mathbf{h})}(x_{\mathbf{h}})$ and, hence, the remaining claim. $\square$

Proof of Theorem 3.2. (a) Let $r_0 \geqslant r$ be the number of regular objects actually contained in the data set of n elements. Without loss of generality, they are the first r_0 objects. Let $\mathbf{h} = (h_i)_{l_i=1}$ be some fixed selection. Let $P_{-1} := \{i \in 1..r_0 \mid l_i = 1, h_i = 1\}$, $P_1 := \{i \in 1..r_0 \mid l_i = 1, h_i \neq 1\}$, $P_0 := \{i \in (r_0+1)..n \mid l_i = 1\}$, a partition of the support of $\mathbf{h}$. If $\#P_{-1} > 0$, let $\overline{Z}_{-1}$ be the random (sample) mean of all $Z_{i,1}$ such that $i \in P_{-1}$ and, if $\#P_j > 0$, $j \in 0..1$, let $\overline{Z}_j$, $\overline{U}_j$ and $\overline{\psi}_j$ be the random (sample) mean of all Z_{i,h_i}, U_{i,h_i} and ψ_{i,h_i} such that $i \in P_j$. Our proof is based on the identity

$$\begin{aligned} V_1(\mathbf{h}) = &\frac{1}{r} \sum_{j=-1}^{1} \sum_{i \in P_j} (Z_{i,h_i} - \overline{Z}_j)(Z_{i,h_i} - \overline{Z}_j)^T \\ &+ \sum_{-1 \leqslant j < k \leqslant 1} \frac{\#P_j}{r} \frac{\#P_k}{r} (\overline{Z}_j - \overline{Z}_k)(\overline{Z}_j - \overline{Z}_k)^T, \end{aligned}$$

($\sum_\emptyset := 0$) which follows from [4, Lemma A.3], applied to the partition $\{P_{-1}, P_1, P_0\}$ of the support of $\mathbf{h}$. In terms of the random vectors $Z_{i,1}$ and U_i, $1 \leqslant i \leqslant n$, this

equality reads

$$
\begin{aligned}
V_1(\mathbf{h}) = {} & \frac{1}{r}\sum_{i\in P_{-1}}(Z_{i,1}-\overline{Z}_{-1})(Z_{i,1}-\overline{Z}_{-1})^T \\
& +\frac{1}{r}\sum_{j=0}^{1}\sum_{i\in P_j}\Big[\kappa_j(U_{i,h_i}-\overline{U}_j)+\psi_{i,h_i}-\overline{\psi}_j\Big]\Big[\kappa_j(U_{i,h_i}-\overline{U}_j)+\psi_{i,h_i}-\overline{\psi}_j\Big]^T \\
& +\frac{\#P_{-1}}{r}\sum_{j=0}^{1}\frac{\#P_j}{r}\Big(\overline{Z}_{-1}-\kappa_j\overline{U}_j-\overline{\psi}_j\Big)\Big(\overline{Z}_{-1}-\kappa_j\overline{U}_j-\overline{\psi}_j\Big)^T \\
& +\frac{\#P_1}{r}\frac{\#P_0}{r}\Big[(\kappa_1\overline{U}_1-\kappa_0\overline{U}_0)+(\overline{\psi}_1-\overline{\psi}_0)\Big]\Big[(\kappa_1\overline{U}_1-\kappa_0\overline{U}_0)+(\overline{\psi}_1-\overline{\psi}_0)\Big]^T.
\end{aligned}
$$

With the abbreviation $\widehat{\psi}_{i,h_i} := \psi_{i,h_i}-\overline{\psi}_j$, it implies the following two estimates valid for $j\in 0..1$:

$$
\begin{aligned}
rV_1(\mathbf{h}) &\geqslant \sum_{i\in P_j}\Big[\kappa_j(U_{i,h_i}-\overline{U}_j)+\widehat{\psi}_{i,h_i}\Big]\Big[\kappa_j(U_{i,h_i}-\overline{U}_j)+\widehat{\psi}_{i,h_i}\Big]^T \\
&\geqslant \kappa_j\Bigg[\kappa_j\sum_{i\in P_j}(U_{i,h_i}-\overline{U}_j)(U_{i,h_i}-\overline{U}_j)^T \\
&\quad +\sum_{i\in P_j}\Big(\widehat{\psi}_{i,h_i}U_{i,h_i}^T+U_{i,h_i}\widehat{\psi}_{i,h_i}^T\Big)\Bigg] \qquad (18)
\end{aligned}
$$

and

$$
\begin{aligned}
V_1(\mathbf{h}) &\geqslant \frac{1}{r}\sum_{i\in P_{-1}}(Z_{i,1}-\overline{Z}_{-1})(Z_{i,1}-\overline{Z}_{-1})^T \\
&\quad +\frac{\#P_{-1}\#P_j}{r^2}\Big(\overline{Z}_{-1}-\kappa_j\overline{U}_j-\overline{\psi}_j\Big)\Big(\overline{Z}_{-1}-\kappa_j\overline{U}_j-\overline{\psi}_j\Big)^T. \qquad (19)
\end{aligned}
$$

We need two important consequences of Lebesgue continuity and independence of the random vectors U_i, $1\leqslant i\leqslant n$. If $j\in 0..1$ is such that $\#P_j\geqslant d+1$ then we have $\sum_{i\in P_j}(U_{i,h_i}-\overline{U}_j)(U_{i,h_i}-\overline{U}_j)^T > 0$, P-a.s. and, if $\#P_j\geqslant 1$, then $\overline{U}_j\neq 0$, P-a.s. These assertions are also true for $Z_{i,1}$ instead of U_{i,h_i}.

Now, assume that $\mathbf{h}$ does not have the desired property, that is, assume $\#P_{-1} < r$. If $\#P_{-1}\leqslant d$ then $\#P_j\geqslant d+1$ for $j=0$ or $j=1$ by the assumption $r\geqslant 3d+1$ and by the pigeon hole principle. Due to the first consequence above, the first matrix on the right in (18) converges to infinity as $\kappa_j\to\infty$, P-a.s. Therefore, in this case, we have $V_1(\mathbf{h})\to\infty$ as $\kappa_j\to\infty$, P-a.s. In the opposite case ($\#P_{-1}\geqslant d+1$), equality (19), P-a.s. regularity of the random (sample) covariance matrix S_{-1} of all $Z_{i,1}$ such that $i\in P_{-1}$, and equality (16) in [4] imply

$$
\begin{aligned}
\det\, V_1(\mathbf{h}) \geqslant \det\Big(\frac{\#P_{-1}}{r}S_{-1}\Big)\Bigg[1+\frac{\#P_j}{r}\Big(\overline{Z}_{-1}-\kappa_j\overline{U}_h-\overline{\psi}_j\Big)^T \\
\times S_{-1}^{-1}\Big(\overline{Z}_{-1}-\kappa_j\overline{U}_j-\overline{\psi}_j\Big)\Bigg], \quad P\text{-a.s.} \qquad (20)
\end{aligned}
$$

Now, by assumption, we have $\#P_j \geqslant 1$ for some $j \in 0..1$ and the second consequence above implies $\overline{U}_j \neq 0$ P-a.s. Therefore, the (random) vector $Y := \overline{Z}_{-1} - \kappa_j \overline{U}_j - \overline{\psi}_j$ tends to infinity P-a.s as $\kappa_j \to \infty$. Furthermore, denoting the largest eigenvalue of S_{-1} by $\rho > 0$, we have $Y^T S_{-1}^{-1} Y \geqslant \rho^{-1} \|Y\|^2 \to \infty$ and (20) $\to \infty$ as $\kappa_j \to \infty$, P-a.s. This completes our proof since $V_1(\mathbf{1}_r)$ is finite and $\det V_1(\mathbf{1}_r) < \infty$, P-a.s., by square integrability of $Z_{i,1}$.

The assumption $r = r_0(n)$ of part (b) guarantees that, at stage n, estimation of m_1 and V_1 is carried out with all regular objects up to n. Therefore, its proof is a standard application of the strong law of large numbers. Part (c) follows from (b). □

A proof of Theorem 3.4 proceeds in the same way as that of Theorem 3.2 with the trace replacing the determinant. For a proof of Theorem 3.5, we first state and prove three lemmas.

Lemma 1. *If Y_1 is d-dimensional standard normal and if $\alpha > 0$ then the function*

$$m \mapsto P[\|Y_1 - m\| \leqslant \alpha] = P_{Y_1}(B_\alpha(m)), \quad m \in \mathbb{R}^d$$

decreases continuously and strictly to zero as $\|m\| \to \infty$.

Proof. For $d = 1$, the claims follow from continuity, symmetry, and monotonicity of the normal density function. The continuity being plain for all d, it remains to prove the decreasing property for $d \geqslant 2$. Let $Y_1 = (X, U)$, X d-1-dimensional, U scalar (both standard normal and independent). By point symmetry, it is sufficient to compare $m_1, m_2 \in \mathbb{R}^d$ of the form $m_k = (0, \ldots, 0, m_k')$, $0 \leqslant m_1' < m_2'$. By independence of X and U, we have

$$\begin{aligned} P[\|(X, U) - (0, \ldots, 0, m_2')\| \leqslant \alpha] &= P[\|X\|^2 + |U - m_2'|^2 \leqslant \alpha^2] \\ &= \int_0^{\alpha^2} P[|U - m_2'|^2 \leqslant \alpha^2 - t] P_{\|X\|^2}(\mathrm{d}t). \end{aligned} \quad (21)$$

Using the claim for $d = 1$, we may conclude

$$(21) < \int_0^{\alpha^2} P[|U - m_1'|^2 \leqslant \alpha^2 - t] P_{\|X\|^2}(\mathrm{d}t) = P[\|(X, U) - (0, \ldots, 0, m_1')\| \leqslant \alpha]. \qquad \square$$

Lemma 2. *Let Y_1 be d-dimensional standard normal. For any $\varepsilon > 0$, there exists $\alpha_0 > 0$ such that, for all $m \in \mathbb{R}^d$, we have*

$$\int_0^{\alpha_0} P[\|Y_1 - m\|^2 > \alpha] \, \mathrm{d}\alpha \geqslant d - \varepsilon.$$

Proof. The claim follows from Lemma 1 and

$$\int_0^{\alpha_0} P[\|Y_1 - m\|^2 > \alpha] \, \mathrm{d}\alpha \geqslant \int_0^{\alpha_0} P[\|Y_1\|^2 > \alpha] \, \mathrm{d}\alpha \underset{\alpha_0 \to \infty}{\longrightarrow} E\|Y_1\|^2 = d. \qquad \square$$

Lemma 3. *Let $Y = (Y_1, \ldots, Y_b)$ be a bd-dimensional random vector such that Y_1 is d-dimensional standard normal. For all $\varepsilon > 0$, there exists $\eta > 0$ such that, for all $m \in \mathbb{R}^d$,*

we have

$$E \min_h \|Y_h - m\|^2 \geqslant d - \varepsilon$$

if $f_{Y_{\widehat{1}}}[y_{\widehat{1}} \mid Y_1 = y_1] \leqslant \eta$ *for all* $y \in \mathbb{R}^{bd}$.

Proof. We first compute ($\wedge$ indicates the minimum of two numbers)

$$\begin{aligned} &E\left(\|Y_1 - m\|^2 \wedge \min_{h\geqslant 2} \|Y_h - m\|^2\right) \\ &= \int_0^\infty P[\|Y_1 - m\|^2 \wedge \min_{h\geqslant 2} \|Y_h - m\|^2 > \alpha]\,\mathrm{d}\alpha \\ &= \int_0^\infty E\left(\mathbf{1}_{[\|Y_1-m\|^2>\alpha]} P\left[\min_{h\geqslant 2} \|Y_h - m\|^2 > \alpha \mid Y_1\right]\right)\mathrm{d}\alpha. \end{aligned} \tag{22}$$

Now, let α_0 be as in Lemma 2 and choose $\eta > 0$ so small that $P[\min_{h\geqslant 2} \|Y_h - m\|^2 \leqslant \alpha_0 \mid Y_1] < \varepsilon$ for all $m \in \mathbb{R}^d$. It follows $P[\min_{h\geqslant 2} \|Y_h - m\|^2 \leqslant \alpha \mid Y_1] < \varepsilon$ for all $\alpha \leqslant \alpha_0$ and, hence,

$$\begin{aligned} (22) &\geqslant \int_0^{\alpha_0} E\left(\mathbf{1}_{[\|Y_1-m\|^2\geqslant\alpha]} P\left[\min_{h\geqslant 2} \|Y_h - m\|^2 > \alpha \mid Y_1\right]\right)\mathrm{d}\alpha \\ &\geqslant (1-\varepsilon)\int_0^{\alpha_0} P[\|Y_1 - m\|^2 > \alpha]\,\mathrm{d}\alpha \geqslant (1-\varepsilon)(d-\varepsilon) \end{aligned}$$

for all $m \in \mathbb{R}^d$. □

Proof of Theorem 3.5. Given a selection $\mathbf{h} \in \prod_{i=1}^\infty 1..b_i$ and a vector $m \in \mathbb{R}^d$, let

$$v_1^{(n)}(\mathbf{h}, m) = \frac{1}{nd}\sum_{i=1}^n \|Z_{i,h_i} - m\|^2,$$

the trace divided by d of the scatter matrix about m of the first n variants selected. Let $\mathcal{G}_j = \{i \in 1..n \mid b_i = j\}$ and let $n_j = \#\mathcal{G}_j$. Plainly, $\lim_n n_j/n = s_j$. If $\mathbf{h}^*$ is optimal for $(Z_1, \ldots, Z_n)$, we have for all m

$$\begin{aligned} v_1^{(n)}(\mathbf{h}^*, m) &= \frac{1}{nd}\sum_{i=1}^n \|Z_{i,h_i^*} - m\|^2 = \sum_{j=1}^b \frac{n_j}{n}\frac{1}{n_j d}\sum_{i\in\mathcal{G}_j} \|Z_{i,h_i^*} - m\|^2 \\ &= \sum_{j=1}^b \frac{n_j}{n}\frac{1}{n_j d}\sum_{i\in\mathcal{G}_j} \min_h \|Z_{i,h} - m\|^2 \\ &\underset{n\to\infty}{\longrightarrow} \frac{1}{d}\sum_{j=1}^b s_j E \min_h \|Z_{i,h} - m\|^2 \end{aligned} \tag{23}$$

P-a.s. by the strong law. According to Lemma 3, there exists $\eta > 0$ s.th. the right side is at least $1 - \varepsilon$ for all $m \in \mathbb{R}^d$ if the conditional density $f_{Z_{i\widehat{1}}}[z_{i\widehat{1}} \mid Z_{i1} = z_{i1}]$ is less than η.

Since $v_1^{(n)}(\mathbf{H}^*(n)) = \min_m v_1^{(n)}(\mathbf{h}, m)$, we have shown the lower estimate in the claim on the variance. The upper estimate follows from $\limsup_n v_1(\mathbf{H}^*(n)) \leqslant \lim_n v_1(\mathbf{1}_n) = 1$.

The claim on the mean value follows from the result just proved and Steiner's formula. Indeed, from (23) applied with $m = 0$, we first infer

$$d\, v_1^{(n)}(\mathbf{h}^*, 0) \underset{n\to\infty}{\longrightarrow} \sum_{j=1}^{b} s_j E \min_h \|Z_{i,h}\|^2 \leqslant E\|Z_{i,1}\|^2 = d$$

and conclude

$$\begin{aligned}\|m_1(\mathbf{H}^*(n))\|^2 &= \frac{1}{n}\sum_{i=1}^{n} \|Z_{i,h_i^*}\|^2 - \frac{1}{n}\sum_{i=1}^{n} \|Z_{i,h_i^*} - m_1(\mathbf{H}^*(n))\|^2 \\ &= d v_1^{(n)}(\mathbf{h}^*, 0) - d v_1(\mathbf{H}^*(n)) \leqslant d + \varepsilon - (d - \varepsilon) = 2\varepsilon\end{aligned}$$

if n is large enough. □

References

[1] T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, in: R. Altman, D. Brutlag, P. Karp, R. Lathrop, D. Searls (Eds.), Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, CA, 1994, pp. 28–36.

[2] O. Barndorff-Nielsen, Information and Exponential Families in Statistical Theory, Wiley, Chichester, New York, Brisbane, Toronto, 1979.

[3] R.G. Casey, E. Lecolinet, A survey of methods and strategies in character segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 18 (1996) 690–706.

[4] M.T. Gallegos, G. Ritter, A robust method for cluster analysis, Ann. Statist. 33 (2005) 347–380.

[5] U. Gather, B.K. Kale, Maximum likelihood estimation in the presence of outliers, Comm. Statist.—Theory Methods 17 (1988) 3767–3784.

[6] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1984) 609–628.

[7] A.J.F. Griffiths, J.H. Miller, D.T. Suzuki, R.C. Lewontin, W.M. Gelbart, An Introduction to Genetic Analysis, Freeman and Company, New York, 1993.

[8] D. Gusfield, Algorithms on Strings, Trees, and Sequences; Computer Science and Computational Biology, Cambridge University Press, Cambridge, 1997.

[9] A. Krogh, M. Brown, I.S. Mian, K. Sjölander, D. Haussler, Hidden Markov models in computational biology, applications to protein modeling, J. Mol. Biol. (235) (1994) 1501–1531.

[10] M. Kuramochi, G. Karypis, Frequent subgraph discovery, in: Proceedings of the 2001 IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, 2001, pp. 313–320.

[11] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, J.C. Wootton, Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, Science 262 (1993) 208–214.

[12] C.E. Lawrence, A.A. Reilly, A expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences, PROTEINS: Structure, Function, Genetics 7 (1990) 41–51.

[13] J.S. Liu, The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem, JASA 89 (1994) 958–966.

[14] J.S. Liu, M. Gupta, X. Liu, L. Mayerhofer, C.E. Lawrence, Statistical models for biological sequence motif discovery, Case Studies in Bayesian Statistics VI, Lecture Notes in Statistics, vol. 167, Springer, New York, 2002, pp. 3–32, with a discussion by Michael A. Newton and rejoinder.

[15] J.S. Liu, C.E. Lawrence, Bayesian inference on biopolymer models, Bioinformatics 15 (1999) 38–52.

[16] J.S. Liu, A.F. Neuwald, C.E. Lawrence, Bayesian models for multiple local sequence alignment and Gibbs sampling strategies, JASA 90 (1995) 1156–1169.
[17] H.P. Lopuhaä, P.J. Rousseeuw, Breakdown points of affine equivariant estimators of multivariate location and covariance matrices, Ann. Statist. (1991) 229–248.
[18] K.V. Mardia, T. Kent, J.M. Bibby, Multivariate Analysis, sixth ed., Academic Press, London, New York, Toronto, Sydney, San Francisco, 1997.
[19] R.A. Maronna, Robust *M*-estimators of multivariate location and scatter, Ann. Statist. 4 (1976) 51–67.
[20] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equations of state calculations by fast computing machines, J. Chem. Phys. 21 (1953) 1087–1091.
[21] D. Mumford, Pattern theory: the mathematics of perception, in: LI Tatsien (Ed.), Proceedings of the International Congress on Mathematics, Beijing, vol. I, World Scientific, Singapore, 2002, pp. 401–422.
[22] C. Pesch, Computation of the minimum covariance determinant estimator, in: W. Gaul, H. Locarek-Junge (Eds.), Classification in the Information Age, Proceedings of the 22nd Annual GfKl Conference, Dresden 1998, Springer, Berlin, 1999, pp. 225–232.
[23] C. Pesch, Eigenschaften des gegenüber Ausreissern robusten MCD-Schätzers und Algorithmen zu seiner Berechnung, Ph.D. Thesis, Universität Passau, Fakultät fur Mathematik und Informatik, 2000.
[24] J. Piper, E. Granum, On fully automatic feature measurement for banded chromosome classification, Cytometry 10 (1989) 242–255.
[25] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (1989) 257–286.
[26] G. Ritter, Classification and clustering of objects with variants, in: W. Gaul, O. Opitz, M. Schader (Eds.), Data Analysis, Scientific Modeling and Practical Application, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, Heidelberg, 2000, pp. 41–50.
[27] G. Ritter, System zur Erkennung und Klassifizierung von Objekten, German patent 10037742, August 2000.
[28] G. Ritter, M.T. Gallegos, Outliers in statistical pattern recognition and an application to automatic chromosome classification, Pattern Recognition Lett. 18 (1997) 525–539.
[29] G. Ritter, M.T. Gallegos, A Bayesian approach to object identification in pattern recognition, in: A. Sanfeliu, et al. (Eds.), Proceedings of the 15th International Conference on Pattern Recognition, vol. 2, Barcelona, 2000, pp. 418–421.
[30] G. Ritter, M.T. Gallegos, Bayesian object identification: variants, J. Multivariate Anal. 81 (2002) 301–334.
[31] G. Ritter, M.T. Gallegos, K. Gaggermeier, Automatic context-sensitive karyotyping of human chromosomes based on elliptically symmetric statistical distributions, Pattern Recognition 28 (1995) 823–831.
[32] G. Ritter, C. Pesch, Polarity-free automatic classification of chromosomes, Comput. Statist. Data Anal. 35 (2001) 351–372.
[33] G. Ritter, G. Schreib, Profile and feature extraction from chromosomes, in: A. Sanfeliu, et al. (Eds.), Proceedings of the 15th International Conference on Pattern Recognition, vol. 2, Barcelona, 2000, pp. 287–290.
[34] G. Ritter, G. Schreib, Using dominant points and variants for profile extraction from chromosomes, Pattern Recognition 34 (2001) 923–938.
[35] P.J. Rousseeuw, Multivariate estimation with high breakdown point, in: W. Grossmann, G.Ch. Pflug, I. Vincze, W. Wertz (Eds.), Mathematical Statistics and Applications, vol. 8B, Reidel, Dordrecht, Boston, Lancaster, Tokyo, 1985, pp. 283–297.
[36] P.J. Rousseeuw, K. Van Driessen, A fast algorithm for the minimum covariance determinant estimator, Technometrics 41 (1999) 212–223.
[37] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: Simple building blocks of complex networks, Science 298 (2002) 824–827.
[38] G.D. Stormo, G.W. Hartzell III, Identifying protein-binding sites from unaligned DNA fragments, Proc. Natl. Acad. Sci. 86 (1989) 1183–1187.