**REGULAR ARTICLE**

# Parameters not empirically identifiable or distinguishable, including correlation between Gaussian observations

## Christian Hennig[1]

## Abstract

It is shown that some theoretically identifiable parameters cannot be empirically identified, meaning that no consistent estimator of them can exist. An important example is a constant correlation between Gaussian observations (in presence of such correlation not even the mean can be empirically identified). Empirical identifiability and three versions of empirical distinguishability are defined. Two different constant correlations between Gaussian observations cannot even be empirically distinguished. A further example are cluster membership parameters in $k$-means clustering. Several existing results in the literature are connected to the new framework. General conditions are discussed under which independence can be distinguished from dependence.

**Keywords** Independence testing · Random effect · Estimability · K-means clustering · Model assumptions

**Mathematics Subject Classification** 62F99 · 62G99

## 1 Introduction

Meaningful statistical inference is only possible if the target of inference (parameter) is identifiable, meaning that if parameter values differ, the parameterised distributions should also differ. There are several versions of identifiability definitions, and many identifiability and non-identifiability results, see, e.g., Yakowitz and Spragins (1968), Rothenberg (1971), Prakasa Rao (1992), Ho and Rosen (2017).

Here situations are treated in which parameters are identifiable according to this classical definition, yet the parameters cannot be identified from observed data. Rothenberg (1971), Hsiao (1983), Prakasa Rao (1992) define identifiability with

✉ Christian Hennig
christian.hennig@unibo.it

[1] Dipartimento di Scienze Statistiche "Paolo Fortunati", University of Bologna, Via delle belle arti, 40126 Bologna, Italy

explicit reference to observable data but do not cover the issues that are treated here. Regarding the results in Rothenberg (1971), there is no difference between classical identifiability and identifiability from observations. Simple examples for classical identifiability issues are the non-identifiability of linear regression parameters in case of collinear explanatory variables, and identifiability of the parameters of mixture distributions, which can be guaranteed under certain assumptions (Yakowitz and Spragins 1968), particularly ruling our label switching, but counterexamples exist (Prakasa Rao 1992, Chapter 8). Hsiao (1983), Prakasa Rao (1992) also study situations in which issues occur because certain modelled random variables are unobservable, such as the true value of a variable in errors-in-variables models. Some examples in Sect. 4 are also of this kind, but there are further reasons why the observed data may not allow for identification of classically identifiable parameters, which are explored here.

Some such situations have already appeared in the literature, see, e.g., Neyman and Scott (1948), Bahadur and Savage (1956), Donoho (1988), Spirtes et al. (1993), Robins et al. (2003), Molenberghs et al. (2008), Almeida and Mouchart (2014). Section 4 gives more details on these works, and how they fit into the unified terminology introduced here.

It turns out that there are different possible levels of information about identifiable parameters in the data, and therefore various definitions are introduced. Consistent estimators may or may not exist ("empirical identifiability"). Sets that can distinguish two parameter values for a finite sample size may or may not exist ("empirical distinguishability" with weaker and stronger versions).

The concept of empirical identifiability is closely connected to the concept of estimability, which is also stronger than classical identifiability. Once more there are several versions around. The concept mostly focuses on what can be estimated with a give finite sample, and its connection with classical identifiability is investigated, see, e.g., Bunke and Bunke (1974), Jacquez and Greif (1985), Maclaren and Nicholson (2020).

This work was motivated by the discovery that data hold no information about distinguishing i.i.d. Gaussian observations from Gaussian data with a constant correlation between any two observations. This will be used as a guiding example. Section 2 derives a key result regarding this situation. Section 3 presents the main definitions, some of their implications and some more examples. Section 4 reviews results from the literature that fit into the framework of Sect. 3. Section 5 uses this framework to discuss the general problem of telling apart dependence and independence in situations in which potential dependence is not governed by the observation order or observable external information. This is relevant in many situations that require independence assumptions. Section 6 presents another example in some detail, namely the empirical identification of parameters indicating the cluster memberships of every single point in $k$-means clustering. Section 7 concludes the paper. All proofs are in the Appendix.

## 2 Constant correlation between Gaussian observations

This work was motivated by the following example, which in itself should be of strong interest.

**Example 1** A model assumption for much standard statistical inference is to assume independently identically distributed (i.i.d.) Gaussian $X_1, \ldots, X_n$, $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ (model M0). Now consider Gaussian $X_1, \ldots, X_n$ with correlation $\text{Cor}(X_i, X_j) = \rho > 0$ constant for any $i \neq j$ (model M1; M01 denotes the model with $\rho \geq 0$ assumed).

This would be a problem for inference about $\mu$, because in the latter situation, for the arithmetic mean $\bar{X}_n$:

$$\mathcal{L}(\bar{X}_n) = \mathcal{N}\left(\mu, \frac{(1-\rho)\sigma^2}{n} + \rho\sigma^2\right) \to_{n\to\infty} \mathcal{N}\left(\mu, \rho\sigma^2\right).$$

This means that the mean is inconsistent for $\mu$ as long as $\rho\sigma^2 > 0$; confidence intervals and tests computed based on the i.i.d. assumption will be biased, possibly dramatically so.

Although $\rho$ is identifiable in the classical sense, it turns out to not be empirically identifiable from observed data. It is not even possible to empirically distinguish any two $\rho_1 \neq \rho_2$ (in fact, plots of data generated from model M01 with different values for $\rho$ including the i.i.d. case $\rho = 0$ do not reveal any features by which these distributions could be distinguished). $\mu$ in model M01 is not empirically identifiable either, but two $\mu_1 \neq \mu_2$ are empirically distinguishable, see Sect. 3.

The following lemma shows that in model M01, the conditional distribution given the mean $\bar{X}_n$ is the same as for i.i.d., therefore uncorrelated, Gaussian random variables. But the mean does not hold information about correlations (or rather, any information about correlations is confounded with the information about the true means), meaning that model M1 cannot be distinguished from model M0 based on the data alone.

**Lemma 1** $\mathbf{X}_n = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$, $\mathbf{Y}_n = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$. *Assume*

$$\mathcal{L}(\mathbf{X}_n) = \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \ \boldsymbol{\mu} = \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, \ \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \ldots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \ldots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \ldots & \sigma^2 \end{bmatrix}.$$

*Then, for*

$$\mathcal{L}(\mathbf{Y}_n) = \mathcal{N}_n(\boldsymbol{\mu}, (1-\rho)\sigma^2 \mathbf{I}_n) : \ \mathcal{L}(\mathbf{X}_n \mid \bar{X}_n) = \mathcal{L}(\mathbf{Y}_n \mid \bar{Y}_n),$$

*which does not depend on $\mu$, where $\mathbf{I}_n$ is the n-dimensional unit matrix.*

Thus, conditionally on the mean, $\mathbf{X}_n$ will look like i.i.d. Gaussians with variance $(1-\rho)\sigma^2$; for $\rho > 0$ there is less variation of the $X_i$ given their mean than their unconditional variance. On the other hand, $\bar{X}_n$ has a larger variance than under independence ($\rho = 0$).

In fact, the model can equivalently be written as a model with a single realisation of a random effect $Z$, $i = 1, \ldots, n$:

$$X_i = \mu + Z + E_i, \ Z \sim \mathcal{N}(0, \tau_1^2), \ E_i \sim \mathcal{N}(0, \tau_2^2),$$
$$\sigma^2 = \tau_1^2 + \tau_2^2, \ \rho = \frac{\tau_1^2}{\tau_1^2 + \tau_2^2}, \tag{1}$$

which suggests that observed data look like i.i.d. $\mathcal{N}(\mu^*, \tau_2^2)$ with $\mu^* = \mu + Z$, and $Z$ is unknown and unobservable. The definitions and results in Sect. 3 aim at making precise a general sense in which observations give no information about $\rho$ and limited information about $\mu$.

## 3 Empirical identifiability and distinguishability

Let $X_1, X_2, \ldots, X_n, \ldots$ be random variables on a space $\mathcal{X}$, for $n \in \mathbb{N}$ : $\mathcal{L}(X_1, \ldots, X_n) = P_{n;\theta}$ with parameter $\theta \in \Theta$. $P_{\infty;\theta}$ denotes the distribution of the whole sequence. The spaces $\mathcal{X}$ and $\Theta$ can be very general, but assume that $\Theta$ is a metric space with metric $d_\Theta$. The focus may be on the parameter $\theta$ in full, or it may be on $g(\theta)$, where $g : \Theta \mapsto \Lambda$, $\Lambda$ being a metric space with metric $d_\Lambda$. No further conditions on $\Theta$ and $\Lambda$ are required for the general definitions. It is generally assumed that the underlying $\sigma$-algebras are rich enough so that the sets required in the arguments are measurable. In the specific cases discussed here this is always fulfilled using standard (Borel) $\sigma$-algebras and parameter spaces. Sometimes but not always $g(\theta) = \theta$ and $\Lambda = \Theta$ are considered. Other examples for $g$ are a projection on a lower dimensional space, or an indicator function for a parameter subset (hypothesis) of interest.

**Definition 1** $g(\theta)$ is called *empirically identifiable* if it is possible to find a consistent sequence of estimators $(T_n)_{n \in \mathbb{N}}$, i.e., with $T_n = \mathcal{X}^n \mapsto \Lambda$, $\forall \theta \in \Theta$ : $T_n(X_1, \ldots, X_n) \to g(\theta)$ in probability.

Consistency is always meant with respect to $P_{\infty;\theta}$. Traditionally, statistical identifiability of a parametric model $(P_\theta)_{\theta \in \Theta}$ means that $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$; for parameter parts, $g(\theta_1) \neq g(\theta_2) \Rightarrow P_{\theta_1} \neq P_{\theta_2}$ is often referred to as partial identifiability (Prakasa Rao 1992; Ho and Rosen 2017). If parameters are not (partially) identifiable, they can obviously not be empirically identifiable, because no consistent estimator can tell equal distributions apart:

**Corollary 1** *Parameters and parameter parts that are empirically identifiable are also identifiable.*

Here, data generating mechanisms are treated that do not allow to empirically identify parameters that are in fact identifiable in the traditional sense. Model M01 is an example. Obviously, distributions with different correlation parameters $\rho_1 \neq \rho_2$ are different from each other, and $\rho$ can be estimated consistently if the whole sequence of $n$ observations is repeated *independently*. In this case, assuming equal correlation

between any two components, the sequence of length $n$ of observations becomes an $n$-variate Gaussian, and the sample correlation between any two of the $n$ components will estimate $\rho$ consistently, although a better estimator will of course use information from all components. The data generating mechanism modelled in Sect. 2 does not allow for independent repetition; all available observations are dependent on all other observations, and this makes consistent estimation of $\rho$ impossible:

**Theorem 1** *Using the notation of Lemma* 1, *if, for $n \in \mathbb{N}$: $\mathcal{L}(\mathbf{X}_n) = \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then, with $\theta = (\mu, \sigma^2, \rho)$, $g(\theta) = \rho$ is not empirically identifiable in model M01.*

Not only is the correlation $\rho$ not empirically identifiable, the same holds for $\mu$, meaning that in practice using an estimator different from $\bar{X}_n$ does not help dealing with the potential existence of $\rho > 0$.

**Theorem 2** *Using the notation of Lemma* 1 *and Theorem* 1, *if for $n \in \mathbb{N}$: $\mathcal{L}(\mathbf{X}_n) = \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $g(\theta) = \mu$ is not empirically identifiable.*

The proof of Theorem 2 relies on the random effects formulation (1) with only a single realisation of the random effect. A similar case can be made for a standard random effects model assuming that the number of realised values of the random effect is bounded even if the number of observations goes to infinity, as expressed in the following model M2:

$$X_{ij} = \mu + Z_i + E_j, \ Z_i \sim \mathcal{N}(0, \tau_1^2), \ E_j \sim \mathcal{N}(0, \tau_2^2),$$

$i = 1, \ldots, m$ (group), $j = 1, \ldots, n_i$ (within group observation), $n = \sum_{i=1}^m n_i$, $\theta = (\mu, \tau_1, \tau_2)$. Let $m$ be fixed, whereas $n$ is allowed to grow. Let $\mathbf{X}_n$ be the vector collecting all $X_{ij}$.

Such a model could make sense for a random effects meta analysis with a low number $m$ of studies, each of which is potentially large, but it does not allow to empirically identify the random effects' variance, and neither the overall mean, unless $m \to \infty$. Consequently, common advice in the meta analysis literature is to not use a random effects model if the number of studies is low [see, e.g., Kulinskaya et al. (2008)].

**Lemma 2** *In model M2, $g_1(\theta) = \mu$ and $g_2(\theta) = \tau_1^2$ are not empirically identifiable, whereas $g_3(\theta) = \tau_2^2$ is empirically identifiable.*

In model M01, there is a difference between trying to estimate $\rho$ on one hand and $\mu$ on the other hand. While $\mu$ cannot be estimated consistently, in case that $\rho$ is small, the data can give fairly precise information about its location, whereas there is no information in the data about $\rho$ at all. The following definition aims at formalising this difference.

**Definition 2** For $n \in \mathbb{N}$, $\alpha \leq \beta \in (0, 1]$, an observable set $A$, i.e., any measurable set expressing an observable event, is an $(\alpha, \beta, n)$-*distinguishing set* for $\theta_1 \neq \theta_2 \in \Theta$ if

$$P_{n;\theta_1}(A) \leq \alpha, \ P_{n;\theta_2}(A) > \beta. \tag{2}$$

**Definition 3** Two values $\lambda_1 \neq \lambda_2 \in \Lambda$ are called *empirically distinguishable* if $\exists n, \alpha \in (0, 1]$, and $\forall \theta_1, \theta_2 \in \Theta$ with $g(\theta_1) = \lambda_1$, $g(\theta_2) = \lambda_2$ there is an $(\alpha, \alpha, n)$-distinguishing set $A$.

Obviously, this definition is symmetric in $\lambda_1, \lambda_2$. Before returning to the problem of constant correlation between Gaussian observations, empirical distinguishability is discussed in some more generality.

For $\epsilon > 0$ and $\eta_0$ in some metric space $H$ with metric $d_H$, define $B_\epsilon(\eta_0) = \{\eta : d_H(\eta, \eta_0) \leq \epsilon\}$. If $g(\theta) = \theta$, empirical distinguishability follows from empirical identifiability, because there is a consistent estimator $T_n$ of $\theta$, and $A = \{T_n \in B_\epsilon(\theta_2)\}$ will distinguish $\lambda_1 = \theta_1 \neq \lambda_2 = \theta_2$ for large enough $n$ if $\epsilon$ is chosen small enough that $\theta_1 \notin B_\epsilon(\theta_2)$.

In general, empirical identifiability does not imply empirical distinguishability. If $g(\theta)$ specifies only a part of the information in $\theta$, it may happen that no set $A$ can distinguish $g(\theta_1) = \lambda_1$ from $g(\theta_2) = \lambda_2$ uniformly over the information in $\theta$ that is not in $g(\theta)$, even if $g(\theta)$ is empirically identifiable.

***Example 2*** Let $X_i$, $i \in \mathbb{N}$ be independently distributed according to $P_\theta$, $\theta = (p, m)$, $p \in [0, 1]$, $m \in \mathbb{N}$, defined as follows: For $i \geq m$, $\mathcal{L}(X_i) =$Bernoulli$(p)$. For $i < m$, $\mathcal{L}(X_i) =$Bernoulli$(q)$, where $q$ is randomly drawn from $\mathcal{U}(0, 1)$. $g(\theta) = p$ is empirically identifiable, because $\bar{X}_n$ is consistent for it. But any $p_1 \neq p_2$ are not empirically distinguishable, because for any $n < m$, $X_1, \ldots, X_n$ do not contain any information about $p$. In this situation, the data may carry information about whether $n > m$ (namely where it can be observed that at some point in the past $q$ may likely have changed to $p$), in which case it also carries information about $p$, but for $m = 1$ this can never happen, and for very small $m$ this can hardly ever be diagnosed with any reliability.

If $\lambda = g(\theta)$, in order to make $\lambda_1$ and $\lambda_2$ empirically distinguishable from having a consistent estimator $(T_n)_{n \in \mathbb{N}}$ of $\theta$, in general $g$ needs to be uniformly continuous, and $(T_n)_{n \in \mathbb{N}}$ needs to be uniformly consistent on $C = g^{-1}(\lambda_1) \cup g^{-1}(\lambda_2)$, i.e.,

$$\forall \epsilon > 0, \alpha > 0 \; \exists n_0 \forall n \geq n_0, \theta \in C : \; P_{n;\theta}\{T_n \in B_\epsilon(\theta)\} > 1 - \alpha.$$

The latter holds automatically in case that $\theta$ is empirically identifiable if $|C|$ is finite, and in particular if $g$ is bijective.

**Lemma 3** *If $\theta \in \Theta$ is empirically identifiable, g is uniformly continuous on an open superset of C, requiring that such a set exists, and there is an estimator $T_n$ of $\theta$ that is uniformly consistent on C, then $\lambda_1 \neq \lambda_2 \in \Lambda$ are empirically distinguishable.*

Only assuming $\lambda = g(\theta)$ but not $\theta$ to be empirically identifiable, consistency of $(T_n)_{n \in \mathbb{N}}$ as estimator of $g(\theta)$ needs to be uniform on $C$ for making $\lambda_1 \neq \lambda_2$ empirically distinguishable.

Consistent estimators are uniformly consistent in many situations. For example, the behaviour of affine equivariant multiple linear regression estimators is uniform over the whole parameter space. Therefore, if they are consistent for the full parameter vector, they are uniformly consistent, and will according to Lemma 3 empirically distinguish subvectors and single coefficients of the regression parameter.

**Remark 1** There are possible variants of Definition 3 that may be taken as different "grades" of empirical distinguishability.

In the situation in Example 2, events may be observed for large enough $n$ that can distinguish $p_1$ and $p_2$, even though this is not guaranteed to happen. A concept that allows $p_1$ and $p_2$ to be seen as distinguishable in some sense (at least if too low $m$ is excluded; $m$ needs to be large enough that a "change point" after $m$ observations can be diagnosed with nonzero probability regardless of $p$) is "potential distinguishability", see Definition 4.

Furthermore, it would make a difference to not allow the probabilities $P_{n;\theta_1}(A)$, $P_{n;\theta_2}(A)$ to be arbitrarily close. Consider a simple i.i.d. $\mathcal{N}(\theta, 1)$ model. With the given definition, and, for given fixed $\theta_0$, $g(\theta) = 1(\theta = \theta_0)$, 0 and 1 can be empirically distinguished (the rejection region of the standard two-sided test will do), i.e., it can be distinguished whether $\theta = \theta_0$ or $\theta \neq \theta_0$. Using a definition that requires a "gap" of some $\beta > 0$ between $P_{n;\theta_1}(A)$ and $P_{n;\theta_2}(A)$, 0 cannot be distinguished from 1, because for $\theta \to \theta_0$, $P_\theta(A)$ for any $A$ gets arbitrarily close to $P_{\theta_0}(A)$. Both of these definitions can be seen as appropriate, from different points of view. It could be argued that $\bar{X}$ contains some, if not necessarily conclusive, information about whether $\theta = \theta_0$, and it should therefore count as distinguishable from $\theta \neq \theta_0$, as achieved by Definition 3,

But even with arbitrarily large $n$, $\bar{X}$ will not be exactly zero, and will therefore be at least as compatible with some $\theta \neq \theta_0$ as with $\theta_0$, which could be used to argue that the two should not be defined as distinguishable. Choosing $g$ as the identity, any fixed $\theta \neq \theta_0$ could still be distinguished from $\theta_0$; in this case there is a positive distance between $\theta$ and $\theta_0$, which makes $\bar{X}_n$ a better fit for the closer parameter. This can be achieved by the concept of "empirical gap distinguishability", see Definition 4.

**Definition 4** (a) $\lambda_1 \in \Lambda$ is called *potentially empirically distinguishable* from $\lambda_2 \in \Lambda$ if $\exists \alpha \in (0, 1)$, $n \in \mathbb{N}$, a set $D \subseteq g^{-1}(\lambda_1) \times g^{-1}(\lambda_2)$ so that

$$\forall \theta_1 \in g^{-1}(\lambda_1) \,\exists \theta_2 \in g^{-1}(\lambda_2) : \; (\theta_1, \theta_2) \in D,$$

and a set $A$ that $(\alpha, \alpha, n)$-distinguishes $\theta_1$ from $\theta_2$ for all $(\theta_1, \theta_2) \in D$.
(b) Two values $\lambda_1 \neq \lambda_2 \in \Lambda$ are called *empirically gap distinguishable* if $\exists n, \alpha < \beta \in (0, 1]$, and $A$ that is an $(\alpha, \beta, n)$-distinguishing set $\forall \theta_1, \theta_2 \in \Theta$ with $g(\theta_1) = \lambda_1$, $g(\theta_2) = \lambda_2$.

Potential distinguishability in particular implies $\forall \theta_1 \in g^{-1}(\lambda_1) : \; P_{n;\theta_1}(A) \leq \alpha$, so that $\lambda_1$ can be "rejected" by the indicator of $A$ (keeping in mind that $\alpha$ cannot necessarily be chosen small), even though this test may be biased against some $\theta_2 \in g^{-1}(\lambda_2)$. Potential distinguishability is not symmetric in $\lambda_1$ and $\lambda_2$, but in Example 2, in fact $p_1$ is potentially distinguishable from $p_2$ ($A$ can be chosen as intersection of a set that rejects the null hypothesis of no change point in the binary sequence, see Worsley (1983), and $|\bar{X}_n^* - p_1|$ being large, where $\bar{X}_n^*$ is the mean after the estimated change point), and $p_2$ is potentially distinguishable from $p_1$ in the same way. The reason why there is symmetric potential distinguishability here but not standard distinguishability is that different distinguishing sets are required for the two directions, and that no set

works uniformly over all $\theta \in g^{-1}(p_1) \cup g^{-1}(p_2)$. See Example 5 for a genuinely asymmetric instance of potential distinguishability.

Empirical gap distinguishability is stronger than empirical distinguishability, whereas potential distinguishability is weaker. Still, lack of empirical gap distinguishability means that in terms of the parameterised probabilities of observable sets, $\lambda_1 \neq \lambda_2$ appear infinitesimally close, regardless of $d_\Lambda(\lambda_1, \lambda_2)$.

Theorem 3 treats empirical distinguishability in model M01.

**Theorem 3** *In model M01, any two $\rho_1 \neq \rho_2 \geq 0$, $\rho_1, \rho_2 < 1$, are indistinguishable and not even potentially distinguishable in any direction, whereas any two $\mu_1 \neq \mu_2$ are empirically distinguishable.*

**Remark 2** Apart from partial identifiability, there are further weaker versions of the classical identifiability concept. A parameter value is locally identifiable if in an open neighbourhood in the parameter space there is no other parameter that parameterises the same distribution (Rothenberg 1971). Set identifiability (Ho and Rosen 2017) means that sets of equivalent parameter values can be identified and potentially be estimated as opposed to a precise parameter value. In other words, parameters from non-equivalent sets could be (empirically) distinguished, whereas equivalent parameters could not be distinguished. Empirical versions of these definitions are possible, but all the situations treated here that are not empirically identifiable would not be empirically local or set identifiable either. All proofs of empirical non-identifiability in the Appendix rule out the existence of consistent estimators that can tell any two parameter values apart, so neither parameter sets nor parameter values in any neighbourhood of each other can be consistently told apart.

## 4 Examples from the literature

The concepts of empirical identifiability and distinguishability provide a framework that fits various existing results on the limitations of empirically identifying parameters or hypotheses that are identifiable according to the classical definition.

**Example 3** The so-called "incidental parameter problem" was introduced by Neyman and Scott (1948), see Lancaster (2000) for a review. It refers to a situation in which there are observed units the number of which is allowed to go to infinity, and for these observed units there is a bounded finite number of observations, say $x_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, t$, where $n \to \infty$ but $t$ fixed. The model for the distribution of the corresponding random variables $X_{ij}$ involves some parameters $\alpha_i$, $i = 1, \ldots, n$. For the estimation of each of these there are only $t$ observations available, and the $\alpha_i$ will not be estimated consistently if $n \to \infty$, so they are not empirically identifiable, although they may well be empirically distinguishable, depending on the specific model. The problem can be avoided in many situations by modeling the $\alpha_i$ as random effects, so that they are governed by only one or few parameters, but in some situations, e.g., panel data in economics (Lancaster 2000), researchers may be interested in inference about specific $\alpha_i$, and also standard distributional assumptions for the random effects distribution may not seem realistic.

**Example 4** In a famous paper, Bahadur and Savage (1956) show the non-existence of valid statistical inference for the problem of finding out about the true mean in a sufficiently large class of distributions $\mathcal{F}$ with existing mean (essentially requiring that for every $P$ and $Q$ also any mixture of them is in $\mathcal{F}$, and that $\forall \mu \in \mathbb{R} \; \exists P \in \mathcal{F}$ : $E_P(X) = \mu$).

Applying the terminology of the present paper, their Theorem 1 shows that any two means $\mu_1$ and $\mu_2$ are not empirically gap distinguishable. The reason is that with $E_P(X) = \mu_1$, $E_Q(X)$ can be chosen so that for arbitrarily small $\epsilon > 0$ and $R = (1 - \epsilon)P + \epsilon Q$, $E_R(X)$ can take any value $\mu_2 \neq \mu_1$. For arbitrarily large $n$ and small enough $\epsilon$, $P^n(A)$ and $R^n(A)$ are arbitrarily close.

**Example 5** Expanding the work of Bahadur and Savage (1956), Donoho (1988) considered functionals $J$ of distributions, including the number of modes of the density, the Fisher information, any $L_p$-norm of any derivative of the density, the number of mixture components, and the negentropy. He showed that for a sufficiently rich nonparametric family $\mathcal{P}$ of distributions, the graph (i.e., the set of pairs $(P, J(P))$ with distribution $P$ and functional value $J(P)$) is dense in the epigraph (the set of pairs $(P, j)$ where $j \geq J(P)$ using a "testing topology" induced by $d(P, Q) = \sup_{0 \leq \psi \leq 1} |\int \psi \, dP - \int \psi \, dQ|$. As $\psi$ can be the indicator of a distinguishing set, two classes of distributions $\mathcal{P}$ and $\mathcal{Q}$ are not empirically gap distinguishable if $\inf_{P \in \mathcal{P}, Q \in \mathcal{Q}} d(P, Q) = 0$. On the positive side, Donoho proves the existence of confidence sets for lower bounds of these functionals (except the negentropy); enough data can make it possible to identify that $J(P) \geq k$ fixed and given, provided that this is indeed the case for $P$. $J(P) = k_1$ can be potentially distinguished from $J(P) = k_2 < k_1$, but not from $J(P) = k_3 > k_1$.

Whereas distinguishability results are mostly negative, Donoho (1988) constructs a consistent estimator of the number of modes (Corollary of Theorem 3.4) and shows therefore that the number of modes is identifiable from data, but consistency is not uniform, and for fixed $n$ only a lower bound for $J(P)$ can be given.

The key ingredient for Donoho's results [as well as the result of Bahadur and Savage (1956)] is the richness of $\mathcal{P}$. If $\mathcal{P}$ is suitably constrained by assumptions, the functionals can be empirically identified; however it cannot be empirically identified whether such assumptions hold.

**Example 6** Spirtes et al. (1993), Robins et al. (2003) deal with the possibility to infer the presence or absence of causal arrows in a graphical model. Chapter 4 of Spirtes et al. (1993) is about "statistical indistinguishability". They define several indistinguishability concepts of different strengths for the problem of identifying the causal graph. As classical identifiability, these concepts regard the model, not making reference to observable data, but empirical identifiability is also of key interest regarding the issue how much can be inferred about the causal relationship between observed variables in the presence of an unobserved confounder.

Robins et al. (2003) show that in several situations "uniformly consistent tests" do not exist, whereas "pointwise consistent tests" exist; the latter however do not allow to distinguish presence or absence of causal arrows for any fixed $n$ uniformly over the possible parameters. "Consistent tests" are procedures that can have outcomes 0 ("accept absence of arrow"), 1 ("reject absence or arrow"), or 2 ("inconclusive").

Their Example 2 is very simple and most instructive. Assume observed binary random variables $X$ and $Y$ and a categorical unobserved confounder $Z$. The existence of a causal arrow between $X$ and $Y$ is operationalised by a parameter $\theta^*$ encoding the strength of the causal effect, where $\theta^* = 0$ means that there is no causal arrow. There are eight different possible causal graphs encoding the possible conditional independence structures [Fig. 3 in Robins et al. (2003)]. The authors assume that the distribution is faithful to the graph, meaning that there are not more independence relationships in the distribution than encoded in the graph.

Adapting the terminology of the present work, the problem is to distinguish two classes of graphs. The first class $C_1$ consists of those graphs that encode marginal independence between $X$ and $Y$, which implies $\theta^* = 0$ under faithfulness. The second class $C_2$ are the graphs that imply marginal dependence between $X$ and $Y$, in which case a consistent test should give an inconclusive result, because there is a possible graph in which both $X$ and $Y$ are influenced by $Z$, which causes marginal dependence, despite $X$ and $Y$ being independent given $Z$, therefore $\theta^* = 0$. The results in Robins et al. (2003) imply that within the second class, the existence of a causal arrow between $X$ and $Y$ is not empirically identifiable, as only dependence or independence between $X$ and $Y$ can be observed. There is a pointwise consistent test (i.e., a consistent estimator of the indicator variable, therefore empirical identifiability) that can tell apart the first and the second class. The two classes are not empirically gap distinguishable but this is not very surprising as $\theta$ can be arbitrarily close to 0 if a causal arrow exists. What is more remarkable is that the proof of their Theorem 1 (which states that no uniformly consistent test exists) implies that $C_1$ cannot even be empirically gap distinguished from $C_2 \cap \{\theta^* = \theta_0^*\}$, where $\theta_0^* \neq 0$ is a fixed parameter value. This is because a graph in the second class that has causal arrows between each pair of $X$, $Y$, and $Z$ encodes a model that allows for dependence between $Z$ and $X$, and also between $Z$ and $Y$ in such a way that $X$ and $Y$ can be arbitrarily close to marginally independent despite the existing causal effect $\theta_0^*$.

**Example 7** Regarding models for missing values, a key distinction is between missing at random (MAR) and missing not at random (MNAR) mechanisms. MAR holds if the distribution of the missingness indicator only depends on the complete data (including missing values) through the non-missing observations. This is a very convenient assumption for dealing with missing values, because it means that the non-missing data provide enough information to allow for unbiased inference. As acknowledged by the missing values literature, it is doubtful that this assumption is realistic, though, and it is also doubtful whether the data allow to check this assumption, as key information for this is hidden in the missing values.

Molenberghs et al. (2008) made this concern precise by showing that for every MAR model there exists an MNAR model that reproduces the same observed likelihood function, meaning in particular that the densities of the observed data and therefore the probabilities for every observable set are equal between these models. This translates into empirical indistinguishability (not even potential distinguishability is possible) of an indicator of MAR vs. MNAR, once more assuming a sufficiently rich class of models. The authors state that MAR and MNAR may be distinguishable under certain parametric assumptions that restrict the flexibility of the MNAR models to emulate

the likelihoods for certain MAR models; but then it is not empirically distinguishable whether such a model holds or not.

**Example 8** *p*-Dimensional ordinal data is often modelled as generated by discretisation of latent Euclidean continuous variables. There is much work about dimension reduction, but assume here that there are $p$ continuous latent variables, every one of which corresponds to an observed ordinal variable, i.e., if $Y_i$ is the $i$th latent variable, and the observed ordinal variable $X_i$ takes the ordered categories $j = 1, \ldots, k_i$ with probabilities $\pi_1, \ldots, \pi_{k_i}$ with $\pi_0 = 0$, $X_i = j$ if $Y_i$ is between the $\sum_{l=0}^{j-1} \pi_l$- and $\sum_{l=0}^{j} \pi_l$-quantiles of $\mathcal{L}(Y_i)$.

The most popular approach is to assume the latent variables as multivariate Gaussian, see Muthén (1984). A mis-specification of the distribution of the latent distribution can have consequences in practice, see Foldnes and Grønneberg (2020), who discuss tests of this assumption. In fact, all such tests only test the dependence structure of $(Y_1, \ldots, Y_p)$ rather than the shape of the marginal distributions of $\mathcal{L}(Y_i)$, which makes sense as the assignment of categories of $X_i$ according to quantiles of $\mathcal{L}(Y_i)$ obviously works for any continuous $\mathcal{L}(Y_i)$. Following Sklar's famous theorem (Sklar 1959), the joint distribution of $(Y_1, \ldots, Y_p)$ has a cumulative distribution function (CDF) $H$ that can be written as $H(y_1, \ldots, y_p) = C(F_1(y_1), \ldots, F_p(y_p))$ where $F_1, \ldots, F_p$ are the marginal CDFs and $C$ is a copula. This means that any dependence structure observed in ordinal data is compatible with any choice of the marginals. Indeed, for the latent variable model for such data, Proposition 2 of Almeida and Mouchart (2014) implies that $F_1, \ldots, F_p$ are not empirically identifiable, and that any two vectors of marginal CDFs are not even empirically distinguishable (be it potentially). The authors formulate this as classical identifiability statement, but involving the $(Y_1, \ldots, Y_p)$ in the model, even though not observable, means that the model is identified but not empirically, i.e., not from what is observable.

# 5 Distinguishing independence and dependence

The problem of identifying constant correlation between Gaussian observations is an instance of the more general problem to detect dependence between the observations in a sample, particularly if they are meant to be analysed by methods that assume independent data. Here the focus is on i.i.d. data. There will not be sophisticated results in this Section, the focus is on general ideas.

Existing tests such as the runs test (Wald and Wolfowitz 1940) require additional information about the kind of dependence to be detected. Most of them test for dependence governed by the observation order, which is sensible if it can be suspected that closeness in observation order can give rise to dependence. This is often the case if observations are a time series, but also other meaningful orderings are conceivable, and also dependence governed by "closeness" on external variables such as spatial location. Alternatively, there may be a known grouping of observations and possible within-group dependence, as modelled for example by random effects models.

In practice, the observation order is not always meaningful, an originally existing meaningful observation order may be unavailable to the data analyst, or dependence

structures can be suspected that cannot be detected by examining relations between observations that are in some sense "close". Constant correlation between any two Gaussian observations as in model M1 is one example of such a structure.

The question of interest here is whether independence and dependence can be distinguished in case that the observation order carries no relevant information, and neither is there secondary information from additionally observed variables. This amounts to observing the empirical distribution of the data only.

For real-valued data $X_1, \ldots, X_n$, $\hat{F}_n$ denotes the empirical distribution function. Assume that only $\hat{F}_n$ is observed. The concept of empirical identifiability can be applied to binary "parameters", and particularly to a parameter that indicates, within an underlying model, whether $X_1, \ldots, X_n$ are i.i.d. or not. Empirical identifiability involves asymptotics. For $n \to \infty$ here it needs to be assumed that a new sequence $X_1, \ldots, X_n$ is generated for each $n$, because observing a sequence $\hat{F}_1, \ldots, \hat{F}_n, \hat{F}_{n+1}, \ldots$ based on the same sequence of observations will re-introduce observation order information.

At first sight the task of identifying dependence from $\hat{F}_n$ may seem hopeless, given that $\hat{F}_n$ is perfectly compatible with i.i.d. data generated from distributions with a "close" true CDF or even $\hat{F}_n$ itself. As in other identifiability problems, information can come from restrictive assumptions.

**Example 9** Consider model M2 and assume for the number of groups $m \geq 2$, but only $\hat{F}_n$ is observed, meaning that it cannot be observed to which group $i$ an observation $X_{ij}$ belongs. Independence amounts to $\tau_1^2 = 0$, and the interest here is in $g(\theta) = 1(\tau_1^2 = 0)$ denoting the indicator function for $\{\tau_1^2 = 0\}$. In case that $\tau_1^2 = 0$, the underlying distribution of $X_{ij}$ is i.i.d. Gaussian. In case that $\tau_1^2 > 0$, the underlying distribution of $X_{ij}$ partitions the data into different Gaussians for different $i$. Assuming that $\frac{n_i}{n} \to \pi_i > 0$, $\hat{F}_n$ will for large enough $n$ look like a Gaussian mixture, which can be told apart from a single Gaussian. The order of a Gaussian mixture can be consistently estimated (James et al. 2001), and therefore $1(\tau_1^2 = 0)$, which is equal to the indicator of a single mixture component, is empirically identifiable. Note that the cited result is for i.i.d. data from a Gaussian mixture, which in model M2 would require the group memberships to be modelled i.i.d. multinomial$(1, \pi_1, \ldots, \pi_m)$, and then conditioning on the unknown values of $Z_1, \ldots, Z_m$.

The example illustrates that certain empirical distributions can indeed indicate dependence, if corresponding distributional shapes (here a Gaussian mixture with $m \geq 2$ components) are assumed as impossible under independence but can occur under dependence.

Even if general marginal distributions are allowed, there are specific dependence structures that can be identified from the empirical distribution alone. In order to simplify matters, from now on consider binary data $X_1, \ldots, X_n$, for which observing the empirical distribution is equivalent to observing the number of ones or $\bar{X}_n = 1 - \hat{F}_n(0)$, and all marginal distributions are Bernoulli($p$). Call the i.i.d. model M3. A problem of interest is whether there are models for dependence for which all marginal distributions are identical that can be told apart based on $\hat{F}_n$ from M3.

**Example 10** Here is an example for a dependence structure that can be distinguished from $\hat{F}_n$. This relies on $\bar{X}_n$ concentrating on a pre-specified value (here $\frac{1}{2}$) with larger

probability than under M3 even if $p$ equals this value. Consider a model M4 where $\mathcal{L}(X_1) = \text{Bernoulli}(p)$. With $q = p1\left(p \leq \frac{1}{2}\right) + (1-p)1\left(p > \frac{1}{2}\right)$, $r = 2\frac{q}{q+1}$, let $Y$ be an unobserved Bernoulli$(r)$-random variable. If $Y = 0$, $X_2, \ldots, X_n, \ldots$ i.i.d. Bernoulli$(p/2)$. If $Y = 1$, $X_3, X_5, X_7, \ldots$ i.i.d. Bernoulli$(1/2)$, and for all even $n : X_n = 1 - X_{n-1}$, so that $\bar{X}_n = \frac{1}{2}$. $r$ is chosen so that all marginal distributions are Bernoulli$(p)$. With $A_n = \{\bar{X}_n = \frac{1}{2}\}$, $P(A_n) \to 0$ under M3, whereas for even $n$, $P(A_n) = r$ under M4. Therefore $A_n$ gap distinguishes M3 from M4.

Such examples rely on the definition of a subset $A_n$ of possible values of $\hat{F}_n$, which under the dependence model for large enough $n$ has a probability either higher than the maximum over $p$ under M3, or smaller than the corresponding minimum. For checking dependence in practice, $A_n$ needs to be specified in advance, meaning that the user needs to know a priori which values of $\hat{F}_n$ can be suspected to indicate dependence even given the possibility under M3 that $p = \bar{X}_n$. Such information is rarely available in practice.

Summarising, dependence can be diagnosed from the data only if

- It is governed by the known order of observations or known external variables,
- Or it favours (or avoids) specific events regarding the observed empirical distribution compared to the independence model of interest,

both in ways that the user has to specify in advance. It can be suspected that many existing dependence structures are not of this kind, meaning that only very limited aspects of independence between observations, regarding a sequence of such observations, can be checked. It is therefore very important to think through all background information about the data generating process to become aware of further potential issues with independence.

## 6 Cluster membership in $k$-means clustering

$k$-means clustering is probably the most popular cluster analysis method (Jain 2010). It can be connected to a "fixed classification model" (Bock 1996): Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$, $\mathbf{X}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, be independently distributed with

$$\mathcal{L}(\mathbf{X}_i) = \mathcal{N}_p\left(\boldsymbol{\mu}_{\gamma_i}, \sigma^2 \mathbf{I}_p\right), \ \gamma_i \in \{1, \ldots, k\}, \ k > 1, \ \sigma^2 \geq 0. \tag{3}$$

This model can be interpreted as generating $k$ different Gaussian distributed clusters characterised by cluster means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^p$, all with the same spherical covariance matrix, and $\gamma_i$ indicates the true cluster membership of $\mathbf{X}_i$. The $\gamma_i$ take discrete values, and their number converges to $\infty$ with $n$, so these are nonstandard parameters, but in many applications they are of practical interest.

The maximum likelihood (ML)-estimator for $\theta = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \gamma_i, \ldots, \gamma_n)$ in this model is given by $k$-means clustering (the ML-estimator for $\sigma^2$ is easily derived, but this is not relevant here) of data $\tilde{\mathbf{X}}_n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$:

$$T_n(\tilde{\mathbf{X}}_n) = (\mathbf{m}_{1n}, \ldots, \mathbf{m}_{kn}, g_{in}, \ldots, g_{nn})$$

$$= \underset{\mathbf{m}_1,\ldots,\mathbf{m}_k,g_1,\ldots,g_k}{\arg\min} W(\mathbf{m}_{1n},\ldots,\mathbf{m}_{kn},g_{in},\ldots,g_{nn}),$$

$$W(\mathbf{m}_{1n},\ldots,\mathbf{m}_{kn},g_{in},\ldots,g_{nn}) = \sum_{i=1}^{n} \|\mathbf{X}_{in} - \mathbf{m}_{g_{in}}\|^2$$

with ties in the arg min broken in an arbitrary way. For given $\mathbf{m}_1,\ldots,\mathbf{m}_k$, the $g_1,\ldots,g_n$ minimising $W$ are given by

$$g_i = \underset{j\in\{1,\ldots,k\}}{\arg\min} \|\mathbf{X}_i - \mathbf{m}_j\|, \ i = 1,\ldots,n,$$

and with these write $W(\mathbf{m}_{1n},\ldots,\mathbf{m}_{kn}) = W(\mathbf{m}_{1n},\ldots,\mathbf{m}_{kn},g_{in},\ldots,g_{nn})$. For given $g_1,\ldots,g_n$, the cluster-wise mean vectors minimise $W$. As there are only finitely many values of $g_1,\ldots,g_n$, the ML-estimator does always exist, if not necessarily uniquely. Two issues with identifiability here are (i) that the numbering of the clusters is arbitrary and (ii) that $\boldsymbol{\mu}_q = \boldsymbol{\mu}_r$ for $q \neq r$ means that it is not possible to distinguish between $\gamma_i = q$ and $\gamma_i = r$ for $i \in \mathbb{N}$. Therefore assume that the $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k$ are pairwise different and lexicographically ordered (i.e., with obvious notation, $\mu_{11} \leq \ldots \leq \mu_{k1}$ with ties broken by the second variable, or, if there's still equality, by the third and so on, same for $\mathbf{m}_{1n},\ldots,\mathbf{m}_{kn}$). This makes the model identifiable according to the traditional definition.

However, due to the nonstandard nature of the model parameters, the ML-estimator is known to be inconsistent, even if only the estimation of the $k$ mean vectors alone is of interest (Bryant 1991).

The cluster membership parameters are another example for parameters that are identifiable according to the classical definition (because $\gamma_i$ uniquely defines the distribution of $\mathbf{X}_i$), but cannot be empirically identified.

**Theorem 4** *The parameters $\gamma_i$, $i \in \mathbb{N}$ in the fixed classification model defined in (3) are not empirically identifiable.*

It may be suspected that this is a consequence of the fact that for $i = 1,\ldots,n$, only $\mathbf{X}_i$ holds information about the parameter $\gamma_i$, and the number of these parameters goes to $\infty$ with $n \to \infty$. But this is not quite true. More observations add information about the clusters that can in turn be used to classify individual observations better. The problem here is rather the Gaussian distribution assumption, which implies that the marginal density of $\mathbf{X}_i$ is everywhere nonzero, so that the single observation made of $\mathbf{X}_i$ is not enough to determine with probability 1 to what cluster the observation belongs (the setup by which the classical identifiability could be used to estimate this parameter would be to have a potentially infinite amount of replicates of $\mathbf{X}_i$), even if there is an infinite amount of information about the clusters. In fact, there is a different model setup in which the $\gamma_i$ are empirically identifiable, which requires that, where densities exist, the marginal density $f_{\theta^*,n}(\mathbf{X}_i = \mathbf{x})$ is zero wherever $f_{\theta,n}(\mathbf{X}_i = \mathbf{x}) > 0$, where $\theta^*$ equals $\theta$ in all components except $\gamma_i$.

Defining

$$W(P) = (\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_k^*) = \underset{(\mathbf{m}_1,\ldots,\mathbf{m}_k)\in(\mathbb{R}^p)^k}{\arg\min} \int \underset{\mathbf{m}\in\{\mathbf{m}_1,\ldots,\mathbf{m}_k\}}{\min} \|\mathbf{x} - \mathbf{m}\|^2 dP(\mathbf{x}),$$

Pollard (1981) showed that for a distribution $P$ satisfying

$$E_P\|\mathbf{X}\|^2 < \infty, \ W(P) \text{ is unique up to the numbering of the means}, \qquad (4)$$

the $k$-means estimator $\left(T_n^m\right)_{n\in\mathbb{N}}$, where $T_n^m(\tilde{\mathbf{X}}_n) = (\mathbf{m}_{1n}, \ldots, \mathbf{m}_{kn})$, is strongly consistent for $W(P)$. Assume further that

$$\forall j \neq l \in \{1, \ldots, k\}: \ P\{\|\mathbf{X} - \boldsymbol{\mu}_j^*\|^2 = \|\mathbf{X} - \boldsymbol{\mu}_l^*\|^2\} = 0. \qquad (5)$$

For $\mathcal{L}(\mathbf{X}) = P, \ j = 1, \ldots, k$, define

$$A_j = \left\{\mathbf{X}: \ j = \arg\min_l \|\mathbf{X} - \boldsymbol{\mu}_l^*\|^2\right\}, \ P_j = \mathcal{L}(\mathbf{X} \mid \mathbf{X} \in A_j), \ \pi_j = P(A_j).$$

$P_j$ is $P$ constrained to the set $A_j$ of points that are closest to the mean $\boldsymbol{\mu}_j$ ($A_1, \ldots, A_k$ form a Voronoi tesselation of $\mathbb{R}^p$), and

$$P = \sum_{j=1}^k \pi_j P_j \qquad (6)$$

(every distribution can be written as a mixture in this form; as a side remark, $P$ might be a Gaussian mixture, but in this case the $P_j$ are not its Gaussian components). Mixture models of this form can be derived from a model for outcomes $(G, \mathbf{X})$ with $G \in \{1, \ldots, k\}$ distributed according to a categorical distribution with probabilities $(\pi_1, \ldots, \pi_k)$ and $\mathcal{L}(\mathbf{X} \mid G = j) = P_j$. Then $\mathcal{L}(\mathbf{X}) = P$ (McLachlan and Peel 2000). For an i.i.d. sequence $\mathbf{Y} = ((G_1, \mathbf{X}_1), (G_2, \mathbf{X}_2), \ldots)$ let $\mathcal{L}(\mathbf{Y}) = \tilde{P}$, $\mathbf{G} = (G_1, G_2, \ldots)$.

Now consider $\mathcal{L}(\tilde{\mathbf{X}}_n) = P^*$ so that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independently distributed with

$$\mathcal{L}(\mathbf{X}_i) = P_{\gamma_i} \ \gamma_i \in \{1, \ldots, k\}, \ k > 1, \ i = 1, \ldots, n. \qquad (7)$$

This defines a fixed classification model associated to the mixture $P$. Let $Q_P$ be an infinite i.i.d. product of categorical distributions on $\{1, \ldots, k\}$ with probabilities $(\pi_1, \ldots, \pi_k)$. Assume for given $P$ that $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots)$ fulfill

$$\tilde{P}\left\{\lim_{n\to\infty} T_n^m(\tilde{\mathbf{X}}_n) = W(P) \mid \mathbf{G} = \boldsymbol{\gamma}\right\} = 1. \qquad (8)$$

Because of the strong consistency of $T_n^m$, (8) holds with probability 1 under $Q_P$, but note that under (7), $\boldsymbol{\gamma}$ is a fixed parameter and not a random variable, and the fact that (8) holds for $Q_P$-almost all $\boldsymbol{\gamma}$ just means that (8) is not more restrictive than assuming a mixture with fixed proportions, although it will not allow for fully general $\boldsymbol{\gamma}$.

**Theorem 5** *Assuming* (4)*,* (5)*, and* (8)*, the parameters* $\gamma_i$*,* $i \in \mathbb{N}$*, in the fixed classification model defined by* (7) *are empirically identifiable.*

Already from Pollard (1981) it is clear that $k$-means does not actually estimate the centres of the spherical Gaussians in (3), but rather the Voronoi tesselation resulting from $P$, and the resulting clusters are not necessarily spherical. Added here is the observation that one can define meaningful cluster indicators in this setup, and that these can be consistently estimated, even though there is one such indicator for every observation. This is not possible in (3). Furthermore (6) interprets $P$ as a mixture, and thus shows that there is a mixture that $k$-means estimates consistently.

The reader may wonder about empirical distinguishability of the parameters $\gamma_i$, $i \in \mathbb{N}$, i.e., about whether the given values $j_1$ and $j_2$ of $\gamma_i$ for given $i$ could be distinguished. In the situation of Theorem 5, this follows from Lemma 3 (projections of discrete parameters are by definition uniformly continuous). The situation of Theorem 4 is less obvious. It is however clear that the data contain some information about $\gamma_i$ through $\mathbf{X}_i$. If it were possible to empirically identify $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$, $A = \{\|\mathbf{X}_i - \boldsymbol{\mu}_{j_1}\| > \|\mathbf{X}_i - \boldsymbol{\mu}_{j_2}\|\}$ could distinguish $j_1$ and $j_2$. A conjecture is that finding consistent estimators for $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ requires additional conditions on the sequence $(\gamma_i)_{i \in \mathbb{N}}$ that allow to use a consistent estimator from an i.i.d. mixture, see Redner and Walker (1984), Bryant (1991).

# 7 Conclusion

There are various potential issues with the identification of parameters, and the four definitions given here (empirical identifiability, empirical distinguishability, empirical gap distinguishability, potential empirical distinguishability) may not cover all of them; even using the definition of distinguishing sets, further definitions are possible, for example empirical set identifiability, but what is already present allows to deal with many examples.

Apart from the precise definitions, there are also different sources for identifiability and distinguishability problems. In some situations (Examples 7, 8) the problem is that modelled information is not observed, either because it regards missing values, or latent variables. In some situations (Examples 4, 5), the issue is that the class of distributions to consider, even for a single parameter value of interest, is too rich, allowing for so much flexibility that the probability of any observable set cannot be sufficiently constrained. Identifiability and empirical distinguishability can be the result of model assumptions constraining this flexibility, see Example 9; in Example 1, constraining the variance $\sigma^2$ would in turn allow the data to be informative about $\rho$. These model assumptions cannot be justified from the data alone though. In some further situations, the data carry information about the parameter of interest, but this information, or more precisely the growth of the information over $n$, is limited [Lemma 2, Example 3, Theorem 4; Example 1 is also of this kind, using (1)]. Example 2 is constructed so that the distinguishing information may not occur at any finite $n$, and the parameterisation in Example 6 is arbitrarily close to a situation of classical non-identifiability, which is only avoided by the faithfulness assumption.

Some examples such as model M01 and the problem in Sect. 6 are characterised by not allowing for i.i.d. repetition; the corresponding parameters can be identified if the whole sequence of observations is repeated i.i.d., and the lack of empirical identifiability is due to the assumed impossibility to do this. It may be wondered whether such models have a valid frequentist interpretation, which seems to rely on i.i.d. replicability at least in principle. Frequentism needs to be interpreted in a rather "idealist" way to accommodate such situations, appealing to replication of an infinite sequence as a thought experiment, although this issue can arguably be made regarding time series and other models as well; for more on this, see Hennig (2020).

The most relevant and unsettling implication of the work for practice regards the lack of possibility to check certain model assumptions, particularly independence; flexible enough models allowing for non-identical marginals can be impossible to detect as well, although this is not shown here.

The considerations regarding requirements for detecting dependence in Sect. 5 do not only hold for data for which only the empirical distribution is observed; they hold in the same way also for situations in which the observation order, even if known and potentially meaningful, is not informative for the dependence structure, and no external variables exist either that hold such information. This is probably a very common situation. The only way to justify independence then is knowledge about the subject matter and the data generation. Bayesians may think that the lack of information in the data about parameters such as $\rho$ in M01 could be compensated by a prior distribution, but the lack of empirical identifiability and distinguishability of $\rho$ raises the question where quantitative information should come from to set up the prior. A prior could only be obtained from existing qualitative information about the data generating process, and as there is no information in the data, the prior will determine the impact of $\rho$ without the possibility of being "corrected" by the data.

## Appendix A: Proofs

### Proof of Lemma 1

Recall that, in general, if $\mathcal{L}\begin{pmatrix}\mathbf{Z}_1 \\ \mathbf{Z}_2\end{pmatrix} = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_2\end{bmatrix}\right)$, then

$$\mathcal{L}(\mathbf{Z}_1 \mid \mathbf{Z}_2) = \mathcal{N}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}(\mathbf{Z}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_{12}\right). \qquad \text{(A1)}$$

Now consider

$$
\mathcal{L}\begin{pmatrix}X_1 \\ \vdots \\ X_n \\ \bar{X}_n\end{pmatrix} = \mathcal{N}_{n+1}\left(\begin{pmatrix}\mu \\ \vdots \\ \mu\end{pmatrix}, \begin{bmatrix}\sigma^2 & \rho\sigma^2 & \ldots & \rho\sigma^2 & \tilde{\sigma}^2 \\ \rho\sigma^2 & \sigma^2 & \ldots & \rho\sigma^2 & \tilde{\sigma}^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \ldots & \sigma^2 & \tilde{\sigma}^2 \\ \tilde{\sigma}^2 & \tilde{\sigma}^2 & \ldots & \tilde{\sigma}^2 & \tilde{\sigma}^2\end{bmatrix}\right).
$$

Let $\tilde{\sigma}^2 = \frac{\sigma^2(1+(n-1)\rho)}{n} > 0$. Using (A1), $\mathcal{L}(\mathbf{X}_n \mid \bar{X}_n) = \mathcal{N}_n(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, where

$$
\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} + \begin{pmatrix}\tilde{\sigma}^2 \\ \vdots \\ \tilde{\sigma}^2\end{pmatrix} \frac{1}{\tilde{\sigma}^2}(\bar{X}_n - \mu) = \begin{pmatrix}\bar{X}_n \\ \vdots \\ \bar{X}_n\end{pmatrix},
$$

$$
\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \begin{pmatrix}\tilde{\sigma}^2 \\ \vdots \\ \tilde{\sigma}^2\end{pmatrix} \frac{1}{\tilde{\sigma}^2}(\tilde{\sigma}^2, \ldots, \tilde{\sigma}^2)
$$

$$
= \boldsymbol{\Sigma} - \begin{bmatrix}\tilde{\sigma}^2 & \ldots & \tilde{\sigma}^2 \\ \vdots & \ddots & \vdots \\ \tilde{\sigma}^2 & \ldots & \tilde{\sigma}^2\end{bmatrix} = \begin{bmatrix}\left(1-\frac{1}{n}\right)\sigma_*^2 & -\frac{1}{n}\sigma_*^2 & \cdots & -\frac{1}{n}\sigma_*^2 \\ -\frac{1}{n}\sigma_*^2 & \left(1-\frac{1}{n}\right)\sigma_*^2 & \cdots & -\frac{1}{n}\sigma_*^2 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n}\sigma_*^2 & -\frac{1}{n}\sigma_*^2 & \cdots & \left(1-\frac{1}{n}\right)\sigma_*^2\end{bmatrix}
$$

with $\sigma_*^2 = (1-\rho)\sigma^2$. Applying (A1) once more, $\mathcal{L}(\mathbf{Y}_n \mid \bar{Y}_n) = \mathcal{N}_n(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$, where

$$
\boldsymbol{\mu}_* = \boldsymbol{\mu} + \begin{pmatrix}\frac{\sigma_*^2}{n} \\ \vdots \\ \frac{\sigma_*^2}{n}\end{pmatrix} \frac{n}{\sigma_*^2}(\bar{Y}_n - \mu) = \begin{pmatrix}\bar{Y}_n \\ \vdots \\ \bar{Y}_n\end{pmatrix},
$$

$$
\boldsymbol{\Sigma}_* = \sigma_*^2\mathbf{I}_n - \begin{pmatrix}\frac{\sigma_*^2}{n} \\ \vdots \\ \frac{\sigma_*^2}{n}\end{pmatrix} \frac{n}{\sigma_*^2}\left(\frac{\sigma_*^2}{n}, \ldots, \frac{\sigma_*^2}{n}\right)
$$

$$
= \begin{bmatrix}\left(1-\frac{1}{n}\right)\sigma_*^2 & -\frac{1}{n}\sigma_*^2 & \cdots & -\frac{1}{n}\sigma_*^2 \\ -\frac{1}{n}\sigma_*^2 & \left(1-\frac{1}{n}\right)\sigma_*^2 & \cdots & -\frac{1}{n}\sigma_*^2 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n}\sigma_*^2 & -\frac{1}{n}\sigma_*^2 & \cdots & \left(1-\frac{1}{n}\right)\sigma_*^2\end{bmatrix}.
$$

Therefore,     $\mathcal{L}(\mathbf{X}_n \mid \bar{X}_n) = \mathcal{L}(\mathbf{Y}_n \mid \bar{Y}_n)$,

and these do not depend on $\mu$.

## Proof of Theorem 1

Let $(R_n)_{n \in \mathbb{N}}$ with $R_n : \mathcal{X}^n \mapsto [0, 1]$ be a consistent estimator of $\rho$ ($\rho = 1$ and $R_n = 1$ are assumed possible, although the following contradiction to consistency relies on $\rho < 1$; $R_n$ could be allowed to take negative values even though $\rho \geq 0$ must be assumed because otherwise the matrices $\boldsymbol{\Sigma}_j$ below would become negative definite for large enough $n$).

For $j = 1, 2$, consider $\mathcal{L}(\mathbf{X}_n^{(j)}) = \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}_j)$, where $\boldsymbol{\Sigma}_1$ defined by $0 \leq \rho = \rho_1 < 1$, $\sigma^2 = \sigma_1^2 > 0$ and $\boldsymbol{\Sigma}_2$ defined by $\rho_1 < \rho = \rho_2 < 1$, $\sigma^2 = \sigma_2^2 = \frac{1-\rho_1}{1-\rho_2}\sigma_1^2$. With $\mathcal{L}(\mathbf{Y}_n) = \mathcal{N}_n(\boldsymbol{\mu}, (1-\rho_1)\sigma_1^2 \mathbf{I}_n)$, Lemma 1 yields

$$\mathcal{L}(\mathbf{X}_n^{(1)} \mid \bar{X}_n^{(1)}) = \mathcal{L}(\mathbf{Y}_n \mid \bar{Y}_n) = \mathcal{L}(\mathbf{X}_n^{(2)} \mid \bar{X}_n^{(2)}). \tag{A2}$$

Let $\epsilon < (\rho_2 - \rho_1)/2$, and $A = [\rho_1 - \epsilon, \rho_1 + \epsilon]$. Consistency implies

$$\lim_{n \to \infty} P^{\mathbf{X}_n^{(1)}}\{R_n \in A\} = 1, \text{ but } \lim_{n \to \infty} P^{\mathbf{X}_n^{(2)}}\{R_n \in A\} = 0. \tag{A3}$$

This implies that there must be a set $B \subseteq \mathbb{R}$ so that

$$\lim_{n \to \infty} P^{\bar{X}_n^{(1)}}(B) = 1, \ \forall x \in B : \lim_{n \to \infty} P^{\mathbf{X}_n^{(1)}}\{R_n \in A \mid \bar{X}_n^{(1)} = x\} = 1.$$

The conditional distribution $\mathcal{L}(\mathbf{X}_n^{(1)} \mid \bar{X}_n^{(1)} = 1)$ is determined up to a null set of values of $\bar{X}_n^{(1)}$ according to the disintegration theorem. This means that the required $B$ exists for any valid choice of $\mathcal{L}(\mathbf{X}_n^{(1)} \mid \bar{X}_n^{(1)} = x)$ for all $x$. The same argument holds later when conditioning on sets with probability zero.

By (A2) and (A3):

$$\forall x \in B : \lim_{n \to \infty} P^{\mathbf{X}_n^{(2)}}\{R_n \in A \mid \bar{X}_n^{(2)} = x\} = 1 \Rightarrow \lim_{n \to \infty} P^{\bar{X}_n^{(2)}}(B) = 0.$$

Now

$$\mathcal{L}(\bar{X}_n^{(1)}) = \mathcal{N}\left(\mu, \frac{(1-\rho_1)\sigma_1^2}{n} + \rho_1\sigma_1^2\right) \to_{n \to \infty} \mathcal{N}\left(\mu, \rho_1\sigma_1^2\right),$$

$$\mathcal{L}(\bar{X}_n^{(2)}) = \mathcal{N}\left(\mu, \frac{(1-\rho_2)\sigma_2^2}{n} + \rho_2\sigma_2^2\right) \to_{n \to \infty} \mathcal{N}\left(\mu, \frac{\rho_2(1-\rho_1)}{1-\rho_2}\sigma_1^2\right),$$

both with non-vanishing variance, and therefore $B$ with

$$\lim_{n \to \infty} P^{\bar{X}_n^{(1)}}(B) = 1, \ \lim_{n \to \infty} P^{\bar{X}_n^{(2)}}(B) = 0$$

cannot exist, with contradiction, so $R_n$ cannot be consistent.

## Proof of Theorem 2

Let $(T_n)_{n \in \mathbb{N}}$ be a consistent estimator of $\mu$. For $j = 1, 2$, consider $\mathcal{L}(\mathbf{X}_n^{(j)}) = \mathcal{N}_n(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$. For $\mu_1 \neq \mu_2$, let $\epsilon < (\mu_2 - \mu_1)/2$, and $A = [\mu_1 - \epsilon, \mu_1 + \epsilon]$. Consistency implies

$$\lim_{n \to \infty} P^{\mathbf{X}_n^{(1)}} \{T_n \in A\} = 1, \text{ but } \lim_{n \to \infty} P^{\mathbf{X}_n^{(2)}} \{T_n \in A\} = 0. \tag{A4}$$

Now consider the equivalent model formulation (1), for $j = 1, 2$, denoting the $Z$ as $Z^{(j)}$. There must be a set $B \subseteq \mathbb{R}$ so that

$$\lim_{n \to \infty} P^{Z^{(1)}}(B) = 1, \ \forall z \in B : \ \lim_{n \to \infty} P^{\mathbf{X}_n^{(1)}} \{T_n \in A \mid Z^{(1)} = z\} = 1.$$

By (1), with $\mathbf{z} = (z, \dots, z)'$ and $\boldsymbol{\mu}_1 = (\mu_1, \dots, \mu_1)'$,

$$\mathcal{L}(\mathbf{X}_n^{(1)} \mid Z^{(1)} = z) = \mathcal{N}(\boldsymbol{\mu}_1 + \mathbf{z}, \boldsymbol{\Sigma}) = \mathcal{L}(\mathbf{X}_n^{(2)} \mid Z^{(2)} = z + \mu_1 - \mu_2).$$

Therefore, from (A4)

$$\forall z \in B : \ \lim_{n \to \infty} P^{\mathbf{X}_n^{(2)}} \{T_n \in A \mid Z^{(2)} = z + \mu_1 - \mu_2\} = 1 \Rightarrow$$

$$\lim_{n \to \infty} P^{\mathbf{X}_n^{(2)}} \{Z^{(2)} - \mu_1 + \mu_2 \in B\} = 0.$$

But this is not possible, because $\mathcal{L}(Z^{(2)})$ is just a mean shifted version of $\mathcal{L}(Z^{(1)}) = \mathcal{N}(0, \tau_1^2)$ regardless of $n$, and will therefore assign nonzero probability to everything that has nonzero probability under $\mathcal{L}(Z^{(1)})$.

## A.1 Proof of Lemma 2

Regarding the empirical non-identifiability of $g_1(\theta) = \mu$, the proof is basically the same as the proof of Theorem 2, with the following changes. For $k = 1, 2$, let $\mathcal{L}(\mathbf{X}_n^{(k)})$ be defined as in model M2 with $\mu = \mu_k$. Denote the corresponding $\mathbf{Z} = (Z_1, \dots, Z_m)$ as $\mathbf{Z}^{(k)}$. For a consistent $(T_n)_{n \in \mathbb{N}}$, there must be a set $B \subseteq \mathbb{R}^m$ so that

$$\lim_{n \to \infty} P^{\mathbf{Z}^{(1)}}(B) = 1, \ \forall \mathbf{z} \in B : \ \lim_{n \to \infty} P^{\mathbf{X}_n^{(1)}} \{T_n \in A \mid \mathbf{Z}^{(1)} = \mathbf{z}\} = 1,$$

with $A$ as in the proof of Theorem 2. With $\mathbf{z} = (z_1, \dots, z_m)'$ and $\boldsymbol{\mu}_k = (\mu_k, \dots, \mu_k)'$, $k = 1, 2$, obtain for $i = 1, \dots, m$, $j = 1, \dots, n_i$:

$$\mathcal{L}(X_{ij}^{(1)} \mid \mathbf{Z}^{(1)} = \mathbf{z}) = \mathcal{N}(\mu_1 + z_j, \tau_2^2) = \mathcal{L}(X_{ij}^{(2)} \mid \mathbf{Z}^{(2)} = \mathbf{z} + \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

The same argument as in the proof of Theorem 2 then establishes that $(T_n)_{n \in \mathbb{N}}$ cannot be consistent.

Obviously no consistent estimator for $g_2(\theta) = \tau_1$ can exist because only $m < \infty$ instances of $Z_i$ are generated.

Regarding $g_3(\theta) = \tau_2^2$, for $n \to \infty$ there must be $i_0 \in \{1, \ldots, m\}$ with $n_{i_0} \to \infty$. For $j = 1, \ldots, n_{i_0}, \mathcal{L}(X_{i_0 j} \mid Z_{i_0} = z) = \mathcal{N}(\mu + z, \tau_2^2)$ i.i.d., and a standard consistent estimator for the Gaussian variance on these observations will estimate $\tau_2^2$ consistently.

## Proof of Lemma 3

Let $\epsilon = d_\Lambda(\lambda_1, \lambda_2)/3$. Uniform continuity of $g$ implies that there is $\delta > 0$ so that for $i = 1, 2$:

$$\theta \in B_\delta(\theta_i) \Rightarrow g(\theta) \in B_\epsilon(g(\theta_i)).$$

Consistency of $T_n$ means that for large enough $n$ and any $\theta_i$ with $g(\theta_i) = \lambda_i$, $P_{\theta_i}(T_n \in B_\delta(\theta_i))$ is arbitrarily large, as is, because of uniform consistency, $P_{\theta_i}(g(T_n) \in B_\epsilon(\lambda_i))$. As this also implies that $P_{\theta_1}(g(T_n) \in B_\epsilon(\lambda_2))$ is arbitrarily low, $A = \{g(T_n) \in B_\epsilon(\lambda_2)\}$ can distinguish $\lambda_1$ and $\lambda_2$.

## Proof of Theorem 3

First consider $\rho_1 \neq \rho_2$. Denote $\mathcal{L}(X_1, \ldots, X_n)$ in model M01 by $P_{n;\mu,\sigma^2,\rho}$. Consider parameters $\mu_1$, $\sigma_1^2$, and any set $A$ so that for given $n$ : $P_{n;\mu_1,\sigma_1^2,\rho_1}(A) < P_{n;\mu_1,\sigma_1^2,\rho_2}(A)$ and $P_{n;\mu_1,\sigma_1^2,\rho_2}(A) > \alpha$. Show that there exist $\mu_2, \sigma_2^2$ so that $P_{n;\mu_2,\sigma_1^2,\rho_1}(A) \geq P_{n;\mu_2,\sigma_2^2,\rho_2}(A)$, which means that $A$ cannot distinguish $\rho_1$ and $\rho_2$, and it cannot even potentially distinguish $\rho_1$ from $\rho_2$.

Choose $\sigma_2^2 = \frac{1-\rho_1}{1-\rho_2}\sigma_1^2$ so that, according to Lemma 1, for all $\mu$ : $P_{n;\mu,\sigma_1^2,\rho_1}(\bullet \mid \bar{X}_n) = P_{n;\mu,\sigma_2^2,\rho_2}(\bullet \mid \bar{X}_n)$. Then, for any $\mu$:

$$P_{n;\mu,\sigma_1^2,\rho_1}(A) = \int P_{n;\mu,\sigma_1^2,\rho_1}(A \mid \bar{X}_n = x) dP_{n;\mu,\sigma_1^2,\rho_1}^{\bar{X}_n}(x)$$

$$= \int P_{n;\mu,\sigma_2^2,\rho_2}(A \mid \bar{X}_n = x) dP_{n;\mu,\sigma_1^2,\rho_1}^{\bar{X}_n}(x),$$

$$P_{n;\mu,\sigma_2^2,\rho_2}(A) = \int P_{n;\mu,\sigma_2^2,\rho_2}(A \mid \bar{X}_n = x) dP_{n;\mu,\sigma_2^2,\rho_2}^{\bar{X}_n}(x).$$

Recall $P_{n;\mu,\sigma^2,\rho}^{\bar{X}_n} = \mathcal{N}\left(\mu, \frac{(1-\rho)\sigma^2}{n} + \rho\sigma^2\right)$. According to Lemma 1, $P_{n;\mu,\sigma^2,\rho}(\bullet \mid \bar{X}_n = x)$ does not depend on $\mu$. With

$$h(x) = P_{n;\mu,\sigma_2^2,\rho_2}(A \mid \bar{X}_n = x), \ \tau_1^2 = \frac{(1-\rho_1)\sigma_1^2}{n} + \rho_1\sigma_1^2,$$

$$\tau_2^2 = \frac{(1-\rho_2)\sigma_2^2}{n} + \rho_2\sigma_2^2 : P_{n;\mu,\sigma_1^2,\rho_1}(A) = \int h(x) d\mathcal{N}(\mu, \tau_1^2)(x),$$

$$P_{n;\mu,\sigma_2^2,\rho_2}(A) = \int h(x)d\mathcal{N}(\mu,\tau_2^2)(x).$$

The Gaussian density $\varphi_{\mu,\sigma^2}(x)$ with mean $\mu$ and variance $\sigma^2$ is symmetric in $\mu$ and $x$, so that for $x \in \mathbb{R}$ :

$$\int \varphi_{\mu,\sigma^2}(x)d\mu = \int \varphi_{\mu,\sigma^2}(x)dx = 1, \text{ and,}$$

$$\text{for } \tau_1, \ \tau_2 : \int \left(\varphi_{\mu,\tau_1^2}(x) - \varphi_{\mu,\tau_2^2}(x)\right)d\mu = 0.$$

Now,

$$P_{n;\mu,\sigma_1^2,\rho_1}(A) - P_{n;\mu,\sigma_2^2,\rho_2}(A) = \int h(x)\left(\varphi_{\mu,\tau_1^2}(x) - \varphi_{\mu,\tau_2^2}(x)\right)dx.$$

As $0 \le h(x) \le 1$, this is integrable over $\mu$, so

$$\int \int h(x)\left(\varphi_{\mu,\tau_1^2}(x) - \varphi_{\mu,\tau_2^2}(x)\right)dxd\mu =$$
$$\int h(x) \int \left(\varphi_{\mu,\tau_1^2}(x) - \varphi_{\mu,\tau_2^2}(x)\right)d\mu dx = 0. \tag{A5}$$

But this means that if $P_{n;\mu_1,\sigma_1^2,\rho_1}(A) < P_{n;\mu_1,\sigma_2^2,\rho_2}(A)$, then there must be a $\mu_2$ so that $P_{n;\mu_2,\sigma_1^2,\rho_1}(A) > P_{n;\mu_2,\sigma_2^2,\rho_2}(A)$, and $A$ cannot distinguish $\rho_1$ and $\rho_2$.

$\mu_1 \ne \mu_2$ can be distinguished by the set $A = \{|\bar{X}_n - \mu_1| > |\bar{X}_n - \mu_2|\}$, as

$$\forall n, \ \psi = (\sigma^2, \rho) : \ P_{n;(\mu_1,\psi)}(A) < \frac{1}{2}, \ P_{n;(\mu_2,\psi)}(A) > \frac{1}{2}.$$

**Proof of Theorem 4**

Consider for $n \to \infty$ a true parameter vector

$$\theta = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \sigma^2, \gamma_1, \gamma_2, \ldots), \ \sigma^2 > 0.$$

Denote $\mathcal{L}(\tilde{\mathbf{X}}_n) = P_{\theta,n}$. W.l.o.g., consider estimation of $\gamma_1 = q \in \{1, \ldots, k\}$. Let $(G_n)_{n\in\mathbb{N}}$ be a consistent estimator of $\gamma_1$, meaning (because of discreteness) that

$$\lim_{n\to\infty} P_{\theta,n}\{G_n(\tilde{\mathbf{X}}_n) = q\} = 1. \tag{A6}$$

Consider now a parameter vector $\theta^*$ for which $\gamma_1 = r \ne q$, but which is otherwise equal to $\theta$. Consistency implies

$$\lim_{n\to\infty} P_{\theta^*,n}\{G_n(\tilde{\mathbf{X}}_n) = r\} = 1. \tag{A7}$$

As $\gamma_1$ determines the distribution of $\mathbf{X}_1$ but nothing else, for almost all $x \in \mathbb{R}$ : $P_{\theta,n}(\tilde{\mathbf{X}}_n \mid \mathbf{X}_1 = \mathbf{x}) = P_{\theta^*,n}(\tilde{\mathbf{X}}_n \mid \mathbf{X}_1 = \mathbf{x})$. But then (A6) requires the existence of a set $B \subseteq \mathbb{R}$ :

$$\lim_{n \to \infty} P_{\theta,n}\{\mathbf{X}_1 \in B\} = 1, \ \forall \mathbf{x} \in B : \ \lim_{n \to \infty} P_{\theta,n}\{G_n(\tilde{\mathbf{X}}_n) = q \mid \mathbf{X}_1 = \mathbf{x}\} = 1,$$

and (A7) requires

$$\lim_{n \to \infty} P_{\theta^*,n}\{\mathbf{X}_1 \in B\} = 0.$$

Such a set $B$ does not exist, because the distribution of $\mathbf{X}_1$ under both $\theta$ and $\theta^*$ are fixed Gaussian distributions with nonzero density everywhere. Therefore $G_n$ cannot be consistent.

### Proof of Theorem 5

W.l.o.g., consider estimation of $\gamma_1$. For $n \in \mathbb{N}$ let

$$G_n(\tilde{\mathbf{X}}_n) = \underset{j \in \{1,\dots,k\}}{\arg\min} \|\mathbf{X}_1 - \mathbf{m}_{nj}\|^2$$

be the $k$-means estimator of the cluster membership of $\mathbf{X}_1$ (in case of non-uniqueness of the arg min, any tie breaking rule can be used). Because of (8) and the Theorem in Pollard (1981), which applies because of (4),

$$P^* \left\{ \lim_{n \to \infty} T_n^m(\tilde{\mathbf{X}}_n) = (\boldsymbol{\mu}_1^*, \dots, \boldsymbol{\mu}_k^*) \right\} = 1.$$

$\mathbf{X}_1$ is fixed once it is observed and does not change for $n \to \infty$. Unless $\exists j \neq l \in \{1, \dots, k\} : \|\mathbf{X}_1 - \boldsymbol{\mu}_j^*\|^2 = \|\mathbf{X}_1 - \boldsymbol{\mu}_l^*\|^2$ [which happens with probability 0 due to (5)],

$$\gamma_1 = \underset{j \in \{1,\dots,k\}}{\arg\min} \|\mathbf{X}_1 - \boldsymbol{\mu}_j^*\|^2, \text{ and } \epsilon = \min_{j \neq \gamma_1}(\|\mathbf{X}_1 - \boldsymbol{\mu}_j^*\|^2 - \|\mathbf{X}_1 - \boldsymbol{\mu}_{\gamma_1}^*\|^2) > 0.$$

For $n$ large enough, with probability 1, therefore $\|\mathbf{X}_1 - \mathbf{m}_{\gamma_1 n}\|^2 = \min_{j \in \{1,\dots,k\}} \|\mathbf{X}_1 - \mathbf{m}_{jn}\|^2$, and $G_n(\tilde{\mathbf{X}}_n) = \gamma_1$. Therefore $G_n$ is consistent for $\gamma_1$, and $\gamma_1$ as well as all $\gamma_i$ are empirically identifiable.

## References

Almeida C, Mouchart M (2014) Testing normality of latent variables in the polychoric correlation. Statistica 74:3–25. https://doi.org/10.6092/issn.1973-2201/4594

Bahadur R, Savage L (1956) The nonexistence of certain statistical procedures in nonparametric problems. Ann Math Stat 27:1115–1122

Bock HH (1996) Probabilistic models in cluster analysis. Comput Stat Data Anal 23:5–28. https://doi.org/10.1016/0167-9473(96)88919-5

Bryant PG (1991) Large-sample results for optimization-based clustering methods. J Classif 8:31–44

Bunke H, Bunke O (1974) Identifiability and estimability. Mathematische Operationsforschung und Statistik 5(3):223–233

Donoho D (1988) One-sided inference about functionals of a density. Ann Stat 16:1390–1420

Foldnes N, Grønneberg S (2020) Pernicious polychorics: the impact and detection of underlying non-normality. Struct Equ Model Multidiscip J 27(4):525–543. https://doi.org/10.1080/10705511.2019.1673168

Hennig C (2020) Frequentism-as-model (arxiv:2007.05748). eprint2007.05748

Ho K, Rosen AM (2017) Partial identification in applied research: benefits and challenges. In: HonoréB, Pakes A, Piazzesi M et al (eds) Advances in economics and econometrics: eleventh world congress, econometric society monographs, vol 2. Cambridge University Press, pp 307–359. https://doi.org/10.1017/9781108227223.010

Hsiao C (1983) Chapter 4: identification. In: Handbook of econometrics, vol 1. Elsevier, pp 223–283. https://doi.org/10.1016/S1573-4412(83)01008-9

Jacquez JA, Greif P (1985) Numerical parameter identifiability and estimability: integrating identifiability, estimability, and optimal sampling design. Math Biosci 77:201–227

Jain AK (2010) Data clustering: 50 years beyond k-means. Pattern Recogn Lett 31:651–666. https://doi.org/10.1016/j.patrec.2009.09.011

James LF, Marchette DJ, Priebe CE (2001) Consistent estimation of mixture complexity. Ann Stat 29(5):1281–1296. https://doi.org/10.1214/aos/1013203454

Kulinskaya E, Morgenthaler S, Staudte RG (2008) Meta analysis: a guide to calibrating and combining statistical evidence. Wiley, New York

Lancaster T (2000) The incidental parameter problem since 1948. J Econom 95(2):391–413. https://doi.org/10.1016/S0304-4076(99)00044-5

Maclaren OJ, Nicholson R (2020) What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems (arxiv:1904.02826). eprint1904.02826

McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York

Molenberghs G, Beunckens C, Sotto C et al (2008) Every missingness not at random model has a missingness at random counterpart with equal fit. J R Stat Soc Ser B (Stat Methodol) 70(2):371–388. https://doi.org/10.1111/j.1467-9868.2007.00640.x

Muthén B (1984) A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika 49:115–132. https://doi.org/10.1007/BF02294210

Neyman J, Scott EL (1948) Consistent estimation from partially consistent observations. Econometrika 16:1–32

Pollard D (1981) Strong consistency of $k$-means clustering. Ann Stat 9:135–140. https://doi.org/10.1214/aos/1176345339

Prakasa Rao BLS (1992) Identifiability in stochastic models. Academic Press, Boston

Redner RA, Walker HF (1984) Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev 26:195–239

Robins JM, Scheines R, Spirtes P et al (2003) Uniform consistency in causal inference. Biometrika 90(3):491–515

Rothenberg TJ (1971) Identification in parametric models. Econometrica 39(3):577–591

Sklar A (1959) Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de statistique de l'Universitéde Paris 8:229–231

Spirtes P, Glymour C, Scheines R (1993) Causation, prediction, and search. Springer, New York

Wald A, Wolfowitz J (1940) On a test whether two samples are from the same population. Ann Math Stat 11:147–162

Worsley KJ (1983) The power of likelihood ratio and cumulative sum tests for a change in a binomial probability. Biometrika 70(2):455–464

Yakowitz SJ, Spragins JD (1968) On the identifiability of finite mixtures. Ann Math Stat 39(1):209–214. https://doi.org/10.1214/aoms/1177698520