Filling the gaps: Gaussian mixture models from noisy, truncated or incomplete samples

Peter Melchior\textsuperscript{a}, Andy D. Goulding\textsuperscript{a}

\textsuperscript{a}Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA

Abstract

Astronomical data often suffer from noise and incompleteness. We extend the common mixtures-of-Gaussians density estimation approach to account for situations with a known sample incompleteness by simultaneous imputation from the current model. The method, called GMMis, generalizes existing Expectation-Maximization techniques for truncated data to arbitrary truncation geometries and probabilistic rejection processes, as long as they can be specified and do not depend on the density itself. The method accounts for independent multivariate normal measurement errors for each of the observed samples and recovers an estimate of the error-free distribution from which both observed and unobserved samples are drawn. It can perform a separation of a mixtures-of-Gaussian signal from a specified background distribution whose amplitude may be unknown. We compare GMMis to the standard Gaussian mixture model for simple test cases with different types of incompleteness, and apply it to observational data from the NASA Chandra X-ray telescope. The python code is released as an open-source package at https://github.com/pmelchior/pyGMMis.

Keywords: density estimation; multivariate Gaussian mixture model; truncated data; missing at random

1. Introduction

The Gaussian mixture model (GMM) is an important tool for many data analysis tasks, usually employed to approximate a potentially complex density distribution for which there is no suitable parametric form, or to perform a clustering analysis. Its importance is reflected in the wide range of applications and the number of extensions it has received (e.g. McLachlan & Peel, 2000; Mengersen et al., 2011). In this work, we will employ the Expectation-Maximization (EM) algorithm to account for “incomplete” samples, i.e. a generalization of truncated samples where data from arbitrarily shaped subdomains of the feature space have a specified probability of not being reported. Also known as “sample selection bias”, it constitutes a long-standing problem for robust inference of population statistics in many scientific disciplines.

In observational astronomy, sample incompleteness occurs frequently, caused e.g. by gaps between sensors or by proximity to bright objects that render a portion of the observation useless. As a milder form of incomplete data, long-running surveys routinely encounter variations in how well the sky can be observed, for instance when the transparency of the atmosphere or the brightness of the moon changes. As a consequence, surveys exhibit complicated completeness functions for the density of observed samples and derived data products (e.g. Leistedt et al., 2016).

Several approaches for dealing with truncated data within the context of Gaussian mixtures have been presented. Given a simple truncation boundary, one can analytically integrate the Gaussian density distribution over the unobserved subdomains (e.g. Wolynetz, 1979; Lee & Scott, 2012, and references therein). McLachlan & Jones (1988) and Cadez et al. (2002) developed a method for binned data, which requires integration of the Gaussian distribution within bins, and realized that any truncated data can be thought to belong to one extra bin, for which the moments of the GMM are already fully specified by the values in the observed bins. Provided the binning is sufficiently fine, this approach allows for arbitrarily shaped truncation boundaries at the cost of introducing, and integrating over, bins for the entire observed domain.

Our approach instead follows a proposal by Dempster et al. (1977) that an estimate of the missing data be drawn from the current model and the EM be run on the combined (observed and estimated missing) data until convergence. This approach does not require binning and provides flexibility and computational efficiency. The method we develop here is similar to Multiple Imputation (Rubin, 1987) and Stochastic EM (Diebolt & Ip, 1996) schemes for samples with missing features, and we will draw from the terminology developed in that context to clarify what kind of incompleteness our density-estimation method can account for. While inspired by these earlier works, our contribution enables the treatment of missing samples, not just missing features, an extension that renders the GMM applicable to a wide range of observational cases. We also incorporate the “Extreme Deconvolution” approach of Bovy et al. (2011) to account for errors of the observed samples and clarify the interplay between noise and missingness.

The outline of the paper is as follows: In Section 2 we describe the EM algorithm for GMM optimization, show how it can accommodate noisy samples, and generalize it for incomplete and potentially noisy samples. We discuss several prac-
tical extensions of the algorithm in Section 3 and demonstrate the performance of the proposed algorithm on a variety of test cases in Section 4. We present an example application with data from the NASA Chandra X-ray telescope, exhibiting spatially varying completeness and positional errors, in Section 5 and conclude in Section 6.

The method is implemented in pure PYTHON, scalable to millions of samples and thousands of GMM components, and publicly released at https://github.com/pmelchior/pyGMMis.

2. GMM from noisy, incomplete samples

For a general mixture model over a \( d \)-dimensional feature space \( \mathbb{R}^d \), the probability density function (PDF) is

\[
p(x | \theta) = \sum_{k=1}^{K} \alpha_k p(x | \theta_k),
\]

with mixing weights that obey \( \sum_k \alpha_k = 1 \) and component density functions \( p(x | \theta_k) \) with parameters \( \theta_k \), the list of which \( \{ \theta_1, \theta_2, \ldots \} \) we simply write as \( \theta \). We will, for the sake of brevity, abbreviate \( p(x | \theta_k) = p_k(x) \). For a GMM, \( p_k \) is given by a multivariate normal distribution function,

\[
p_k(x) = \mathcal{N}(x | \mu_k, \Sigma_k) \equiv \frac{\exp\left[ -\frac{1}{2} (x - \mu_k)^\top \Sigma_k^{-1} (x - \mu_k) \right]}{(2\pi)^d \det(\Sigma_k)},
\]

the complete set of parameters per component is thus \( \alpha_k, \mu_k, \) and \( \Sigma_k \). The corresponding mixture log-likelihood for noise-free observations \( x_i \) with \( i \in \{1, \ldots, N\} \) given fixed parameters is

\[
\ln \mathcal{L} = \ln p(x | \alpha, \mu, \Sigma) = \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \alpha_k p_k(x_i).
\]

This form is also called uncategorized log-likelihood (Titterington et al., 1985) because there is no information about which component \( k \) generated sample \( i \). We introduce the discrete indicator \( S = (S_1, \ldots, S_N) \) with \( S_i = k \) if (and only if) \( x_i \) is generated by component \( k \). By rewriting \( \alpha_k \sum_k q_k p_k(x_i) = q_k p_k(x_i | \alpha, \mu, \Sigma) \) with the conditional probability \( q_k \) that \( x_i \) is generated by component \( k \), i.e., \( q_k \equiv \Pr(S_i = k | x_i, \alpha, \mu, \Sigma) \), we can form the complete-data log-likelihood

\[
p(x, S | \alpha, \mu, \Sigma) = p(x | \alpha, \mu, \Sigma) + \sum_{k=1}^{K} \sum_{i=1}^{N} I_{S_i = k} \ln \Pr(S_i = k | x_i, \alpha, \mu, \Sigma),
\]

the sampling distribution of the complete data \( (x, S) \) given the parameters of the model \( (\alpha, \mu, \Sigma) \). The second term of the RHS describes the additional information from knowing the allocation \( S \) of samples to components (Frühwirth-Schnatter, 2006).

2.1. Standard EM algorithm

As the indicator \( S \) is not known or directly observable, we need to estimate it by classifying each observation. With Bayes’ rule, and dropping the conditional dependence on the full set of model parameters \((\alpha, \mu, \Sigma)\) in all terms, the classification is given by

\[
\Pr(S_i = k | x_i) = \frac{\Pr(X = x_i | S_i = k) \Pr(S_i = k)}{\sum_{j} \Pr(X = x_i | S_i = j) \Pr(S_i = j)}.
\]

With it one can re-estimate the parameters of the model from the weighted moments of \( x \) given the estimate of \( S \). That is the central idea of the EM procedure, which first estimates the classification (the E-step) and then updates the model parameters (the M-step). Assuming flat prior distributions \( \Pr(S_i = k) \) and using the definition of \( q_{ik} \) from above, we get

\[
\text{E-step: } q_{ik} \leftarrow \frac{\alpha_k p_k(x_i)}{\sum_j \alpha_j p_j(x_i)}
\]

\[
\text{M-step: } \mu_k \leftarrow \frac{1}{q_k} \sum_i q_{ik} x_i,
\]

\[
\Sigma_k \leftarrow \frac{1}{q_k} \sum_i q_{ik} (\mu_k - x_i)(\mu_k - x_i)^\top.
\]

Dempster et al. (1977), later corrected by Wu (1983), showed that repeated iterations of these two steps monotonically converge to a local maximum of \( \mathcal{L} \).

2.2. EM algorithm with noisy samples

Observed data often exhibit measurement uncertainties, which we assume to be additive and Gaussian:

\[
y_i \equiv x_i + e_i,
\]

where \( e_i \sim \mathcal{N}(0, \Sigma) \). We want to emphasize that we do not assume that errors are identically distributed, only that they are independent, Gaussian, and that their covariance matrices are known. Bovy et al. (2011) derived an extended EM algorithm that maximizes the likelihood of the noise-free density \( p(x) \) from noisy samples \( y_i \). The key insight is that one can marginalize over the unknown values \( x_i \) and still obtain a GMM. With \( T_{ik} \equiv \Sigma_k + S_k \) the EM procedure amounts to

\[
\text{E-step: } q_{ik} \leftarrow \frac{\alpha_k p_k(y_i | \mu_k, T_{ik})}{\sum_j \alpha_j p_j(y_i | \mu_j, T_{ij})}
\]

\[
b_{ik} \leftarrow \mu_k + \Sigma_k T_{ik}^{-1}(y_i - \mu_k)
\]

\[
B_{ik} \leftarrow \Sigma_k - \Sigma_k T_{ik}^{-1} \Sigma_k
\]

\[
\text{M-step: } \mu_k \leftarrow \frac{1}{q_k} \sum_i q_{ik} b_{ik}
\]

\[
\Sigma_k \leftarrow \frac{1}{q_k} \sum_i q_{ik} (\mu_k - b_{ik})(\mu_k - b_{ik})^\top + B_{ik}.
\]

By including measurement uncertainties in the definition of \( q_{ik} \), the M-step is almost unchanged: the role of \( x_i \), which is not
directly observable, is played by \( b \), and the covariances get extra contributions \( B \). All of those modifications vanish when \( S \to 0 \). The resulting EM algorithm still converges monotonically to a local maximum of \( L \).

2.3. EM algorithm with incomplete samples

We will now add a stochastic element to the EM algorithm to account for sample incompleteness, by which we mean the following: the probability of a sample \( x \) being observed at all is determined by a completeness function, \( \Omega(x) : \mathbb{R}^d \to [0, 1] \), which means that in any subdomain \( D \subset \mathbb{R}^d \) the density of observed samples can systematically deviate from the true density in that region. Such a situation may arise if a measurement device is incapable of recording samples from \( D \), so that \( \Omega(x) = 0 \) \( \forall x \in D \). This is the typical situation of truncated samples. A softer version is given by \( \Omega(x) < 1 \), which occurs when observations suffer from under-reporting.\(^1\)

One can in principle correct an observed density \( p(x) \) by dividing with \( \Omega(x) \), but if the inverse of the completeness is not expressible as a GMM, the corrected distribution cannot itself be described as a GMM, thus losing its convenient analytic properties. Furthermore, if in some \( D \) the completeness is zero, such a correction would be ill-defined. Instead, we adapt the Stochastic EM (Diebolt & Ip, 1996) concept of augmenting missing data features by drawing imputing samples from the current model. In our case, entire samples can be unobserved, as opposed to some of their features being missing. We thus need to draw an estimate of the completely observed data, apply the selection process described by \( \Omega \), and keep the rejected samples, i.e. inverted rejection sampling with the current GMM as the proposal distribution. Complementing the deterministic EM of the observed data with its stochastic counterpart allows us correct for domains of reduced or zero completeness. To the best of our knowledge, this is the first time that a GMM approach has been made robust against not just truncation but the generalized form of incomplete sampling.

2.3.1. Missingness

Before we adapt the EM algorithm to incorporate \( \Omega \), it is instructive to clarify how our concept of completeness relates to the terminology of “missingness” introduced by Rubin (1976) for missing features. In Appendix A we argue that the situation we find here is equivalent to “missing at random” (MAR), for missing features. In Appendix A we argue that the situation we find here is equivalent to “missing at random” (MAR), which is irrelevant for density estimation as the resulting probability density function is simply a renormalized version of the true PDF. But since a density estimator necessarily obeys \( \int p(x) \, dx = 1 \), this normalization will automatically be removed. Thus, MAR data have the same expectation value of \( p(x) \) as completely observed data, the estimate is merely constrained by fewer samples. On the other hand, the data are not allowed to conform to the most general case called “missing not at random” (MNAR), which would arise from a relation between the missing data and the density itself, \( \Omega(x) \to \Omega(x, p(x)) \) (see e.g. Schafer & Graham, 2002), a generalization that goes beyond the scope of this work.

As the form of the likelihood function \( L \) under MAR is unchanged, the E-step remains unchanged as well, and all corrections appear in the M-step. Lee & Scott (2012) demonstrated that the effects of simple truncation can be addressed by computing the zeroth, first, and second moments of each Gaussian component when truncated in the same way as the data, which requires the analytic integration of \( p_k \) within the observed bounds. Essentially, the update equations for \( \mu_k \) and \( \Sigma_k \) get a correction term from the difference between the current iteration parameters and their truncated values. As we seek the ability to employ arbitrary completeness functions, with complex spatial shapes and probabilistic rejection, the analytical integration becomes cumbersome, so we prefer approaches that draw samples from \( \Omega \).

2.3.2. Stochasticity

One could use Monte-Carlo integration instead to compute the moments under any kind of \( \Omega \), but there is an even more direct approach: approximating a set of “missing” samples \( M \), which augments the observed samples \( O \) to yield an estimate of how the data would be distributed if \( \Omega = 1 \). Effectively, we draw \( oN \) samples, with an oversampling factor \( o \), from the current GMM, apply rejection sampling with an acceptance rate of \( 1 - \Omega \), and keep the retained samples as \( M \) with size \( N_M \). Because of the linearity in the data, we can split both E-step and M-step into an observed and a missing sample part, so that Equation 6 becomes

\[
\frac{\alpha_ip_k(x_i)}{\sum_j \alpha_j p_j(x_i)} \quad \forall i \in [O, M] \\
\frac{1}{N + oN_M} \left( \sum_{i \in O} q_{ik} + \frac{1}{o} \sum_{i \in M} q_{ik} \right) \equiv \frac{1}{N} q_k \\
\mu_k \left[ \frac{1}{L_k} \left( \sum_{i \in O} q_{ik} x_i + \frac{1}{o} \sum_{i \in M} q_{ik} x_i \right) \right] \\
\Sigma_k \left[ \frac{1}{L_k} \left( \sum_{i \in O} q_{ik} \left( (\mu_k - x_i)(\mu_k - x_i)^\top \right) + \frac{1}{o} \sum_{i \in M} q_{ik} \left( (\mu_k - x_i)(\mu_k - x_i)^\top \right) \right) \right].
\]

This approach, which we will call GMF1a, in which at each iteration we draw and augment the observed data with imputation samples, is summarized in Algorithm 1. It is guaranteed to maximize the complete-data log-likelihood and converges to a stationary distribution for the parameters \( \alpha, \mu, \Sigma \) (Diebolt & Ip, 1996; Nielsen, 2000). This method is efficient where \( p(x) (1 - \Omega(x)) \) is large; correction for weakly expressed components are more difficult to attain because \( M \) is drawn globally from the entire GMM, as opposed to being drawn from each

\(^1\)Without a loss in generality, we will assume that over-reporting has been properly corrected, so that \( \Omega(x) \leq 1 \), e.g. by \( \Omega(x) \to \Omega(x)/\max(x) \). We implicitly assume that there are not false positives in the data.
component individually. A minor limitation is that when Ω is large for a component, the values of the M-sums in the M-step equations are not well determined, however their impact on the result is also minor because that component is mostly observed.

In comparison with an analytic evaluation of the PDF in the unobserved domains, this approach is flexible and efficient, does not require numerical integrations of the component distributions, but, due to its stochastic nature, suffers from sample variance in the correction terms because M has finite size. As with all Stochastic EM approaches, it generates sequences of L, computed from O and M, that are not guaranteed to monotonically increase in every step. To reduce the stochastic contribution to L, we can increase the oversampling factor o when drawing M, so that the corrections become more precise and L-sequences closer to monotonic. For sparsely sampled data, larger values of o can become necessary to stabilize the algorithm, so that a general recommendation is to adopt o as large as possible. We found 1 ≤ o ≤ 4 to result in sample variances that are small compared to the increase of the observed likelihood until close to convergence, so that we simply terminate the EM sequence whenever L decreases beyond some tolerance. We leave possible optimization, e.g. increasing o individually for components whose imputation corrections suffer noticeably from the sample variance of O, for future work.

2.3.3. Incompleteness and Noise

Besides speed and flexibility, GMMiS retains the ability to deal with noisy samples. By adding noise to the imputation samples M, we can evaluate qik, bik, and Bjk for i ∈ M and modify Equation 8 analogously to Equation 9:

\[ q_{ik} \leftarrow \alpha_k p_i(y_i | \mu_k, T_{ik}) + \sum_{j \in O} \alpha_j p_j(y_i | \mu_j, T_{ij}) \]

\[ b_{ik} \leftarrow \mu_k + \Sigma_k T_{ik} y_i, \forall i \in O, M \]

\[ B_{ik} \leftarrow \Sigma_k - \Sigma_k T_{ik} \Sigma_k, \forall i \in O, M \]

\[ \alpha_k \leftarrow \frac{1}{N + \frac{1}{2} |M| \left( \sum_{i \in O} q_{ik} + \frac{1}{o} \sum_{i \in M} q_{ik} \right) \equiv \frac{1}{N} q_{ik} \]

\[ \mu_k \leftarrow \frac{1}{q_k} \left( \sum_{i \in O} q_{ik} b_{ik} + \frac{1}{o} \sum_{i \in M} q_{ik} b_{ik} \right) \]

\[ \Sigma_k \leftarrow \frac{1}{q_k} \left( \sum_{i \in O} q_{ik} (\mu_k - b_{ik})(\mu_k - b_{ik})^T + \right. \]

\[ \left. \frac{1}{o} \sum_{i \in M} q_{ik} (\mu_k - b_{ik})(\mu_k - b_{ik})^T \right) \]

To do so, we have to define \( T_{ik} = \Sigma_k + S_i \) for the missing samples, which do not have observed uncertainties. In some cases the positional uncertainties are known even in unobserved regions. If not, we need to make a guess of S(x) in the unobserved domain, e.g. from the mean or a smooth interpolator of the observed S_i.

We caution that not assuming errors for M would result in too large an effective weight because we evaluate its likelihood in the E-step without noise while the observed samples have a broadened likelihood due to the noise. As a result, the EM algorithm would become unduly dominated by missing samples and yield density estimates that are shifted towards regions of low Ω. On the other hand, we do not have to exactly match the errors of M to those of O as they only enter in the sums of the M-step. Matching the average errors of the samples associated with each component, instead of each individual sample, is therefore sufficient to estimate the parameters of all components.

We want to point out that artificially adding noise and applying Ω do not commute; in either order, GMMiS requires only that M can be created such that it completes the data with an estimate of the unobserved portion.

3. Practical considerations

3.1. Initialization

The EM algorithm only guarantees convergence to a local maximum of the likelihood. With the large number of free parameters of the GMM (1 + d + d(d + 1)/2 for each of K components), a suitable initialization is critical. Several initialization schemes have been proposed (e.g. Biernacki et al., 2003; Blömer & Bujana, 2013). For a completely observed data set, we adopt the simplest strategy, namely drawing the means at random from the data. In detail, given a user-defined length scale s, for each component k we draw i at random from \([1, \ldots, N] \) and \( \Delta x_i \sim N(0, s^2 I) \), and set

\[ \alpha_k = 1/K \]

\[ \mu_k = x_i + \Delta x_i \]

\[ \Sigma_k = s^2 I \]
which naturally follows the distribution of the data on scales larger than \( s \). To prevent strong initial localization, \( s \) should be chosen to exceed the typical clustering scale of the data, but small enough that multiple components do not strongly overlap.

In the case of incomplete samples, we initially make the assumption that it was complete, i.e. \( \Omega = 1 \), fit a GMM to the observed distribution, and use that to initialize the run with a specified \( \Omega \neq 1 \). While this approach will obviously fail if a component is entirely located in a region with \( \Omega = 0 \), we found that an initial guess based on the observed distribution much more quickly converges than the random initialization described above. To aid the exploration of the domains with \( \Omega < 1 \), we leave the components means unchanged but multiply the covariances by a factor > 1. If that factor is chosen too small, the EM algorithm will not be able to pick up the correct terms \( \sum_{i \in M} q_{ik} \) etc. in Equation 9 before re-converging to the previous, observed-sample location. In turn, if the factor is chosen too large, the convergence is slowed down. We therefore recommend a factor of 2 – 4, i.e. increasing the linear size of each component by 50 – 100%, as a compromise that works well in practice.

### 3.2. Split-and-merge operations

With a large number of free parameters, the EM algorithm can easily get trapped in local maxima of \( \mathcal{L} \). For clustered data, this behavior leads to GMM components being placed across several clusters or a single cluster being shared by multiple components. In the latter case, the weight of each component by 50 – 100%, as a compromise that works well in practice.

#### Ueda et al. (2000)

Ueda et al. (2000) proposed to merge the components \( k \) and \( l \) that maximize

\[
J_{\text{merge}}(k, l) = Q_k^T Q_l.
\]

(12)

with \( Q_k^T = (q_{ik}, \ldots, q_{iN}) \), i.e. the components whose posterior cluster assignment \( p(k|x) \) are most similar across the entire data set. This works well in practice as long as there are no “empty” components, which do unfortunately arise in multimodal situations if more components are locally available than are needed to explain the data. Because \( q_{ik} \to 0 \) when \( \alpha_k \to 0 \), the merge criterion above will not seek to merge such empty components even though they are obviously excellent choices to be merged. We therefore replace \( Q_k \to \frac{Q_k}{\alpha_k} \), which means that we seek to merge components whose \( p_k(x) \) is most similar for the entire data set.

For the split criterion, we found that the one suggested by Ueda et al. (2000), based on the Kullback-Leibler divergence, is unduly affected by outliers and often leads to split candidates that do not improve the likelihood. Instead, we took guidance from a proposal of Zhang et al. (2003) on how to best re-initialize the two new components that result from a split, namely to separate their means along the semi-major axis of the ellipsoid described by the pre-split \( \Sigma \). With this intuition it is natural to identify split candidates according to their largest eigenvalue \( \lambda_1 \) of \( \Sigma \). When searching for the component \( k = \arg \max_k (\lambda_{k1}) \), we assume it is strongly elongated because it seeks to describe two clusters at once. The main failure mode of that split criterion is again related to (almost) empty components. Their parameters, in particular \( \Sigma_k \), are only poorly determined. Some of them are erroneously large and would thus be identified as split candidates, while they constitute much better merge candidates. We thus propose to identify split candidates by selecting \( k \) as the one that maximizes

\[
J_{\text{split}}(k) = \alpha_k \lambda_{k1}.
\]

(13)

While the original split criterion of Ueda et al. (2000) will seek to eliminate any deviation of the local density from its approximation by a Gaussian-shaped component and is therefore generally applicable, our criterion appears to perform better for identifying a prominent failure mode for unconstrained GMMs: components that merge two clusters. As the split is only accepted if it leads to an overall increase in the likelihood, it does not lead to increased fragmentation from penalizing the largest components. Following Zhang et al. (2003), we then replace the means of two newly split components \( f / m \) as \( \mu_{lm} = \mu_k \pm \frac{1}{2} \text{EV}_1(\Sigma_k) \), i.e. along the primary eigenvector of the covariance matrix.

When dealing with incomplete data, we have not found split-and-merge operations to be problematic. They are most useful when enabled during the initialization run as described in Section 3.1, which will then not be plagued by strong failure modes of the EM algorithm. As a result, the imputation samples \( M \) will be more reliable, and the full algorithm will converge faster than without split-and-merge operations.

#### 3.3. Minimum covariance regularization

One problem of the objective function \( \mathcal{L} \) is that it becomes unbounded if \( \mu_k \to y_i \) for any \( k \) and \( i \) because \( \Sigma_k \to 0 \). Bovy et al. (2011) presented a regularization scheme to set a lower bound for every component as we determine the initial eigenvalue \( \lambda_1 \) of \( \Sigma_k \). We assume it is strongly elongated because it seeks to describe two clusters at once. The main failure mode of that split criterion is again related to (almost) empty components. Their parameters, in particular \( \Sigma_k \), are only poorly determined. Some of them are erroneously large and would thus be identified as split candidates, while they constitute much better merge candidates. We thus propose to identify split candidates by selecting \( k \) as the one that maximizes

\[
J_{\text{split}}(k) = \alpha_k \lambda_{k1}.
\]

(13)

While the original split criterion of Ueda et al. (2000) will seek to eliminate any deviation of the local density from its approximation by a Gaussian-shaped component and is therefore generally applicable, our criterion appears to perform better for identifying a prominent failure mode for unconstrained GMMs: components that merge two clusters. As the split is only accepted if it leads to an overall increase in the likelihood, it does not lead to increased fragmentation from penalizing the largest components. Following Zhang et al. (2003), we then replace the means of two newly split components \( f / m \) as \( \mu_{lm} = \mu_k \pm \frac{1}{2} \text{EV}_1(\Sigma_k) \), i.e. along the primary eigenvector of the covariance matrix.

When dealing with incomplete data, we have not found split-and-merge operations to be problematic. They are most useful when enabled during the initialization run as described in Section 3.1, which will then not be plagued by strong failure modes of the EM algorithm. As a result, the imputation samples \( M \) will be more reliable, and the full algorithm will converge faster than without split-and-merge operations.

One problem of the objective function \( \mathcal{L} \) is that it becomes unbounded if \( \mu_k \to y_i \) for any \( k \) and \( i \) because \( \Sigma_k \to 0 \). Bovy et al. (2011) presented a regularization scheme to set a lower bound for every component volume. In its simplest version it assumes the form of a \( d \)-dimensional sphere with variance \( \omega \), which modifies the last update equation in Equation 8 according to

\[
\Sigma_k \leftarrow \frac{1}{q_k + 1} \left[ \sum_i q_{ik} [\mu_k - \mu_{ik}] (\mu_k - \mu_{ik})^\top + \omega \right] + wI.
\]

(14)

In principle, the value of \( w \) is entirely arbitrary, but we choose to provide a more intuitive setting and associate it with a minimum scale \( \omega \) in feature space, below which the model does not possess explanatory power, i.e. we want \( \det(\Sigma_k) \geq \omega^{2d} \). For a constant regularization term, we need to adopt a typical value of \( q_k \), for which we use its mean over all components, \( \bar{q}_k = \frac{N}{K} \). We can thus set \( w = \omega^2 \left( \frac{N}{K} + 1 \right) \). This choice does not provide an exact lower bound for each component as we determine the regularization term from the average value of \( q_k \), which leads to
stronger regularization for components with small $\alpha_k$. We consider this an advantage in noisy situations, where the parameters of weakly expressed components can be hard to determine.

We note, however, that the form of the regularization in Equation 14 prefers features in the data to be of approximately equal size, otherwise the penalty term can dominate for small features while being ineffective for large features.

3.4. Fitting for a background distribution

In many situations, observed data comprise anomalous samples that appear unclustered, i.e. unrelated to the features of interest, and rather originate from a more uniform “background” distribution. One solution within the context of GMMs would be to add another component with very large variance and to fit for its amplitude only. We prefer to introduce a specific background distribution over the relevant region $\mathcal{D}$ of feature space, e.g. the most conventional form of a uniform background

$$p_{bg}(x) = \begin{cases} \int_\mathcal{D} dx = \text{const.} & x \in \mathcal{D} \\ 0 & x \not\in \mathcal{D}. \end{cases} \quad (15)$$

If the amplitude of the background component $\nu$ is unknown, one can introduce a two-level mixture model for the combined density distribution (e.g. Frühwirth-Schnatter, 2006, their section 7.2.4),

$$p(x | \alpha, \mu, \Sigma, \nu) = (1 - \nu) \sum_{k=1}^K \alpha_k p_k(x) + \nu p_{bg}(x). \quad (16)$$

but we prefer keeping the model strictly linear in the amplitudes and put the component and background amplitudes on equal footing:

$$p(x | \alpha, \mu, \Sigma, \nu) = \sum_{k=1}^K \alpha_k p_k(x) + \nu p_{bg}(x). \quad (17)$$

In this form, $\sum_k \alpha_k = 1 - \nu$. To determine the amplitude $\nu$ we require another indicator variable $q_{bg}$ to denote if a sample $x_i$ belongs to the background. Analogous to $q_{ik}$ in Equation 6 it is given by the posterior of $x_i$ under the background model, which leads to these E-step equations

$$q_{ik} \leftarrow \frac{\alpha_k p_k(x_i)}{\sum_k \alpha_k p_k(x_i) + \nu p_{bg}(x_i)}$$

$$q_{bg} \leftarrow \frac{\nu p_{bg}(x_i)}{\sum_k \alpha_k p_k(x_i) + \nu p_{bg}(x_i)}. \quad (18)$$

The M-step for the background amplitude is

$$\nu \leftarrow \frac{1}{N} \sum_i q_{bg}. \quad (19)$$

while the M-step of the GMM components remains unchanged.

If the samples are noisy, the previous equations in this section hold, but we need to marginalize over the positions of the unobserved noise-free samples (Bovy et al., 2011) and modify the background distribution as

$$p_{bg}(y_i | S_i) = \int dx \ p_{bg}(x) N(y_i | x, S_i), \quad (20)$$

which is equivalent to the change to the GMM $q_{ik}$ in Equation 8 compared to the noise-free case in Equation 6. For the uniform background distribution the marginalization amounts to the zero-th moment of the truncated multivariate normal distribution (e.g. Manjunath & Wilhelm, 2012).

One could think that sample incompleteness does not affect the inference of the background component because any information how $\Omega(x)$ acts on samples drawn from $p_{bg}$ is already entirely contained in $\Omega(x)$ itself. The problem with this notion is that we do not know the relative amplitudes of signal and background in the unobserved regions, which will vary because the signal does. For consistent results, we therefore create the imputation sample $\mathcal{M}$ by drawing from the GMM and the background model according to the current-iteration value of $\nu$, and proceed as in Section 2.3.

We caution that the introduction of a background component leads to additional uncertainties and a higher-dimensional parameter space. In particular, during the first iterations of the EM algorithm, the parameters of the GMM often only provide rather poor description of the data, so that many samples will have higher probability under the background model, resulting in few samples left to fit for the GMM parameters in the next iteration. This failure mode highlights the importance of a suitable initialization when working with a background model, especially when its intensity becomes dominant. To this end, we found in our tests that the $k$-means initialization from Blömer & Bujna (2013, their Algorithm 1) performed more robustly than the random initialization of Equation 11.

3.5. Averaging estimators

Even with well-chosen initial values and split-and-merge operations, any single GMM will get trapped in local maxima of the likelihood. We therefore advocate, for two reasons, to average several GMMs fit to the same data, a technique also known as ensemble learning.

First, ensemble estimators typically outperform even the best single estimator. In particular, Smyth & Wolpert (1999) built an improved estimator by employing the “stacking” method proposed by Wolpert (1992), which uses the cross-validation result to determine non-negative weights for each model (see also Breiman, 1996). In the context of GMMs, it is useful to realize that a mixture of mixture models is still a mixture model, therefore no conceptual changes have to be made when evaluating the estimator.

Second, stacked GMMs reduce the need to determine the optimal number of components $K$. The main concern when determining $K$ is that if set too high, the resulting GMM overfits

\footnote{The presence of a non-vanishing background intensity can therefore in principle be used to estimate $\Omega(x)$. However, for this work we require $\Omega$ to be known a priori.}
the data, following spurious features that increase the variance in the prediction, while if set too low, the model does not capture essential features of the data, leading to a large prediction bias. As this topic has been extensively discussed elsewhere, we will not address it here. We do note, however, that stacking exhibits better trade-offs between bias and variance than single-model selection or uniform averaging, rendering stacked GMM particularly robust against overfitting (Smyth & Wolpert, 1999). Even more so, combining GMMs with different $K$, or different covariance constraints, can provide a natural framework to describe data with a variety of spatial scales.

A decision if and how single-run GMMs are to be averaged will depend on the characteristics of the data at hand. In this work, we will only use single-run results to allow the reader to evaluate the performance of GMMs rather than that of the averaging scheme.

4. Experiments

4.1. A toy example

We first perform a simple test with no model misspecification, i.e., we draw the original sample from a GMM with $K = 3$. In Figure 1, we show that sample as red circles. We then add a noticeable amount of Gaussian noise (blue contours) and impose a purely geometric completeness $\Omega(x)$, whose boundaries are given by a box and a circle (dotted curves), resulting in the test sample (blue squares). With this test case we demonstrate how the EM algorithm reacts to noise and a non-trivial completeness. The standard EM algorithm (top-center panel) does what one needed to expect, namely to describe the observed sample as is. It clearly prefers the region inside of the boundaries and thereby misestimates amplitudes, locations, and covariances of the components affected by $\Omega$.

With the noise-deconvolution approach of Bovy et al. (2011), summarized in Equation 8, one can attempt to recover the noise-free distribution. However, in this case (shown in the top-right panel of Figure 1) the resulting shrinkage of the components is detrimental to the overall likelihood because the model is now even more confined to the observed region, reducing its ability to also describe samples beyond that. In essence, the model is biased as before but more confident in its correctness as the noise contribution has been removed. We generically find it to be true that noticeable incompleteness will need to be addressed first, and only then can one properly deconvolve from the noise, as we demonstrate with the progressive improvement in the next two tests.

When the sample incompleteness from $\Omega$ is correctly specified and considered via Equation 9, GMMs recovers component centers, orientations, and amplitudes much better (bottom-left panel), however the covariances remain inflated as the presence of noise is not yet corrected for. This level of model fidelity could have been achieved with analytical integration (Wolynetz, 1979; Lee & Scott, 2012) or sample binning (McLachlan & Jones, 1988; Cadez et al., 2002), but here none of those operations were necessary. In addition, neither of the aforementioned approaches can account for incompleteness and noise.

In the bottom-center panel, we use the full algorithm, i.e., the noise deconvolution of Equation 8 applied to both the $O$ and $M$ samples in Equation 9, resulting in the highest likelihood of the three variants, with just $A \ln L = -0.151$ compared to red sample drawn from the true underlying PDF.

By running the full algorithm (with $\Omega$ and noise treatment) ten times, we can investigate the spread associated with the given data set under the admittedly restrictive assumption of a GMM with $K = 3$. The bottom-right panel of Figure 1, whose contours have the same stretch as those of the previous panels, shows that there are considerable differences between runs, mostly associated with the $\Omega$-boundaries. Also, the components from different runs emphasize the importance of different apparent sub-clusters. This behavior is caused by the added noise, which obscures the presence of small-scale features. As we attempt to infer the noise-free PDF from noisy samples, the models will amplify small-scale density fluctuations. Because of such spurious sub-clusters or, more generally, local minima of the likelihood, we advocate the use of more sophisticated averaging schemes, as outlined in Section 3.5.

4.2. Sample size and noise level

In addition, we run a similar test setup as in Figure 1, using the same clusters but only a box-shaped boundary of $\Omega$, and vary $N$ and the noise level $\sigma$. The results are shown in the left panel of Figure 2. Unsurprisingly, increasing the sample size $N$ always improves fidelity of the inferred PDF estimate and can in most cases offset the effects of measurement noise. If the scatter grows large compared to the size of the clusters (whose effective width corresponds to 1.45, 1.15, and 0.57, respectively), even large sample sizes do not fully recover the underlying PDF, confirming the intuitive notion that information is actually lost in such a case.

This intuition is confirmed by the results shown in the right panel of Figure 2, where we vary the observed fraction at fixed $N = 1000$. We do this by shrinking the linear size of the box in Figure 1 so as to maintain a MAR situation. Only discarding samples at random would correspond to MCAR and therefore not strongly affect the recovery of the PDF: the GMM would only notice the sparsity of the resulting sample. Instead, in the chosen setup components eventually leave the observed area of some very small boxes. For instance, the top-left component is entirely unobservable around $N^2/N = 0.3$. While it becomes obvious that a reduced observed fraction has much stronger influence on $L$ than the noise level, at large $\sigma$ the recovery of the PDF does not reach the fidelity of lower noise data even when it is fully observed.

4.3. Limits of applicability

We want to investigate, in a more difficult, highly multi-modal situation, how strongly incomplete the sampling can be for the proposed method to still yield reasonable density estimates. We therefore set up a new test case in which we place $K = 50$ GMM components randomly in the $d = 3$ unit cube. Covariances are chosen such that the overlap between the components is not excessive; weights are drawn from a symmetric
Dirichlet distribution with concentration parameter of unity. We impose a probabilistic completeness function \( p(\Omega | x) = 1 - x_1 \), i.e. a linear ramp from fully observed to fully missing along one of the coordinate axes. The purpose of this setup is to allow all combinations of component weight \( \alpha_k \) and completeness \( \Omega \). We draw \( N=10,000 \) samples from the model, apply \( \Omega \), and record the size of samples \( N_k \) drawn from each component \( k \) as well as the mean \( \Omega_k \) experienced by those samples, so that we know the observed sample size \( N^O_k = N_k \Omega_k \). No noise is added to the samples. An example cube is shown in Figure 3.

We then fit the test data with another GMM with the same \( K = 50 \), resulting in best-fit parameters \( \hat{\theta} \), and repeat the process 10 times. As diagnostic, we directly compare the densities \( \rho(x) \) by splitting the cube into \( 50^3 \) cells and counting the input samples that fall into the cell covering \( x \). We do the same for the predicted density \( \hat{\rho}(x) \) by drawing \( N \) samples from the fit. As we know the expectation value of \( \Omega \) at the location of each cell, we can compute the fractional bias in the predicted density (\( \hat{\rho}(x) - \rho(x) \))/\( \rho(x) \) as a function of \( \Omega \) (Figure 4). The standard EM algorithm provides a reliable estimate of the observed density, in other words: the bias is equal to \( \Omega \). The reason for positive bias for \( \Omega > \frac{1}{2} \) lies in the normalization of the density estimator: as we drew the same number of samples \( N \) from the model as where given as input data, any underestimation of the density must be compensated elsewhere.

In the same test, GMMis shows a noticeably reduced sensitivity to \( \Omega \), but the resulting density estimate is still biased. This test demonstrates that there must be conditions under which the proposed algorithm fails. It is obvious that we cannot expect GMMis, or any other density estimator, to infer the properties or even existence of a cluster that is entirely unobserved, i.e. \( N^O_k = 0 \). Even for less extreme cases, Figure 4 shows that GMMis struggles with samples from regions with low \( \Omega \), over-compensating where \( \Omega \) is large.

To determine limits of applicability, we evaluate how well the input samples from each component are described by the fit. We compute the association fraction \( \eta_k \) of the input sample from component \( k \) that falls within “1 \( \sigma \)” of any component of the fit, which, for \( d > 1 \), is more precisely expressed as \( \{ i : (x_i - \hat{\mu}_k)^T \hat{\Sigma}_k (x_i - \hat{\mu}_k) \leq \chi^2_d(0.683) \text{ for any } k \} \), where \( \chi^2_d \)
Figure 2: Reconstruction fidelity for the three clusters in Figure 1 with a box-shaped boundary of $\Omega$ as a function of sample size $N$ (left) and observed sample fraction $N_O/N$ (right). Added noise is isotropic bivariate normal and should be compared to the effective cluster widths of 1.45, 1.15, and 0.57. The observed fraction in the left panel is $N_O/N = 0.6$, the sample size in the right panel is held constant at $N = 1000$, while the observed fraction is varied by changing the size of the box. Errorbars denote the standard deviation of 50 runs.

Figure 3: Example test data generated from $K = 50$ clusters in the $d = 3$ unit cube. The completeness $\Omega$ is probabilistic, decreasing linearly from unity (left side of the cube) to zero (right side). Observed samples are shown in blue, missing samples in red.

is the quantile function of the chi-squared distribution with $d$ degrees of freedom. For a component of the test data whose samples are perfectly fit (i.e. fit by the component from which they are generated), the association fraction should on average be $\eta_k = 68.3\%$. If $\eta_k$ is higher, either the associated component of the fit is too extended or the test samples are associated with multiple components. If $\eta_k$ is lower, the fit has effectively ignored some samples from input component $k$. In Figure 5 we plot $\eta_k$ as a function of the effective weight $N_k^O/N_k^\Omega$ of component $k$ given the observed data, while the size of the marker encodes the initial weight $\alpha_k = N_k/N$, and the color encodes $\Omega_k$. It is apparent that components of the test data with large $\alpha_k$ or $\Omega_k$ are generally well-described by the fit. For $\Omega_k \approx 0$, the fit will miss a large fraction of the original sample irrespective of $\alpha_k$, but there are also a few well-observed but low-$\alpha$ components that are largely being ignored by the fit. There is also an increased tendency of fitting multiple nearby input clusters with one larger component, which leads to artificially large $\eta_k$ at low $N_k^O/N_k^\Omega$. Without formal proof, we seek a criterion to identify components that are too poorly observed to yield a reasonable fit. By performing the test described above with different values of $N$, $K$, component volumes, and Dirichlet concentration.
parameter, we found
\[
N_k^O = \frac{N_k \Omega_k}{\sum_k N_k \Omega_k} \approx \frac{1}{K}
\] (21)
to perform well (shown as vertical dotted line in Figure 5). This
qualitatively agrees with findings of e.g. Naim & Gildea (2012)
that the best GMM fits with an EM algorithm are achieved when
all components have approximately equal weights, in particular
when they overlap. It also implies that \(GMM\) has to success-
fully fit the observed distribution first to be able to generate
meaningful imputation samples \(M\) that account and correct for
\(\Omega\). If the first step is flawed, the second one will be, too.

This two-step picture is confirmed when we remove from
the test data all components that would have violated Equa-
tion 21, in other words, the test data now only comprises clus-
ters that remain salient despite \(\Omega\). In this case, the algorithm
can detect all present clusters and correct for the sample incom-
pleteness. Accordingly, the \(GMM\) density estimates are now
largely unbiased (dash-dotted red line in Figure 4). We emphasi-
size that this shortcoming is not caused by an incorrect form
of the likelihood, rather by the economy of the EM algorithm
to converge to the nearest likelihood maximum, which favors
the most prominent clusters in the observed sample. If, on the
other hand, the parameters of all but one component were fixed
at their true values and there were only one component left to
fit, this component could experience incompleteness in excess
of Equation 21 and still be fit with a fidelity commensurate to
its number of samples \(N_k\).

5. Application to Chandra X-ray data

To demonstrate the capabilities of \(GMM\), we analyze the
distribution of X-ray photons of the nearby galaxy NGC 4636,
the observed data. Combining this imputation sample and the point-source masked data, we get a representation of the internal state of the model, in other words: its estimate how the data would look like if $\Omega = 1$. In each iteration these augmented data enter Equation 10 to determine the best-fit model of the extended emission of the galaxy without point sources (bottom-right panel).

We initialize the GMM with $K = 60$ components with means distributed according to a bivariate Gaussian, whose width $s$ was determined by fitting the galaxy with $K = 1$ first. Because the scene exhibits features of different scales, we set the component covariances to $\Sigma_k = 4^{-l}I_2$ with $l = 0, \ldots, 5$, and

Figure 6: Application to X-ray data. Top left: Histogram of the locations of photons in the energy range 0.5–2 keV for an observation of the galaxy NGC 4636 with the ACIS-I instrument of the Chandra X-ray telescope. Top right: Completeness $\Omega(x)$ for ACIS-I, determined by the location of its four CCDs and detector sensitivity variations; point source masks (red circles, sizes true to scale), where we reject data and explicitly set $\Omega = 0$, have been added for the purpose of this analysis. Bottom left: ACIS-I data augmented with the imputation sample drawn from the final models of the GMM and the background. Bottom right: Final GMM with $K = 60$ of the extended emission of the galaxy, after rejecting point sources, removing the background, and reconvolving to match the observation. The images are aligned such that North is up and East is left; all panels are shown in logarithmic stretch.
10 components at each level \( l \). This tailored initialization is necessary despite the GMM’s ability to adjust to arbitrary configurations because the rate of convergence is very slow for the weakly expressed features we are particularly interested in. The background amplitude \( ν \) was allowed to vary between 0.2 and 0.5, with \( ν = 0.396 \) being the best-fit value, in other words about 40% of the photons from the unobscured region shown in Figure 6 originate from the background.

The GMM is capable of faithfully describing large and small-scale features in the data: a bright, apparently bimodal, supermassive black hole; an extended halo of low-intensity emission from diffuse, \( 10^5 \) – \( 10^7 \) Kelvin hot gas with a noticeable skewness towards the lower right corner; and the location and intensity of several shock fronts caused by buoyantly rising pockets of plasma that were likely inflated during previous outbursts of the central supermassive black hole. However, because of the complexity of the model with \( K = 60 \), it is not guaranteed that particular features, e.g., shock fronts, are fit by a single component. This association could be made more clear if the analysis operated on a three-dimensional feature space of photon positions and energies, a study we leave for the future.

Because this is a demonstration of the capabilities of the method, we have not performed e.g., cross-validation tests to determine the optimal \( K \). The model is the result of a single run with a number of components that appeared visually reasonable to capture the key features of the data. Its \( \chi^2 \) per degree of freedom over an area where the galaxy dominates the emission is at least 1.34,\(^4\) indicating that a larger number of components may be necessary to capture all significant features.

6. Summary and conclusions

We describe a novel extension of the EM algorithm to perform density estimation with Gaussian mixture models in situations where the data exhibits a known incompleteness \( Ω \). The type of incompleteness may be described by a simple binary decision boundary, a case that is usually denoted as “truncated data”, or by an arbitrary probabilistic function \( Ω(χ) \), as long as the mechanism that causes the incompleteness is independent of the density (missing at random).

The key difference of the method described here to the situation that is often—and somewhat ambiguously—called “missing data” is that the latter refers to data, for which some features of any data sample may be absent. In contrast, we deal with entire samples being potentially absent, caused by the systematic limitation to observe the entire feature space or all of its relevant subdomains.

Our solution is based on drawing imputation samples from the current state of the GMM, and to recompute the model parameters in the M-step using both observed and missing samples. This technique is applicable to any generative model that is fit with an EM algorithm, as demonstrated with the signal-background model of Section 3.4. Its advantage is the flexibility to efficiently adjust to any incompleteness \( Ω(χ) \) because one avoids the analytic integration of the predicted density over the incomplete subdomains. Instead, we draw test samples from the current model, retain those that would not have been observed under \( Ω \), and obtain non-zero contributions to the moments of those components that extend into incompletely observed subdomains. If the model is accurate, adding those moments to the ones obtained from the observed samples results in parameter estimates consistent with those for completely observed data.

We detail practical adjustments of the algorithm, regarding its initialization, split-and-merge operations, and simultaneously fitting for a uniform background. We also recommend averaging GMMs from independent runs to smooth out the dependence of the EM algorithm on the starting position to estimate the uncertainty of the estimators.

In simplified tests we demonstrate that the algorithm
1. recovers an estimate of the underlying density in presence of complex incompleteness functions,
2. correctly accounts for incompleteness and additive noise if the imputation samples describe the noisy unobserved data,
3. can only partially correct for incompleteness if the too many of the original samples have not been observed.

The last finding is central to the applicability of the method. Constraining complex GMMs with finite amounts of samples becomes even more challenging with non-negligible incompleteness. Correcting for the latter requires being able to perform the former reasonably well, otherwise the imputation samples do not describe the true unobserved samples. In absence of an analytical estimate of the fidelity of GMMs from the EM algorithm, we provide a best-effort estimate for the limits of the algorithm in Equation 21.

We demonstrate the usefulness of the algorithm with example data from the NASA X-ray telescope Chandra, where the incompleteness stems from gaps between the chips of the ACIS instrument. We directly estimate the extended emission of the galaxy NGC 4636 from the location of photon hits, while accounting for the window-frame configuration of the detector, the spatially varying point-spread function, and a uniform X-ray background.

The presented method provides an efficient and flexible tool to estimate the density of a process that is affected by moderate levels of incompleteness of the MAR type. As such situations may arise in many areas of the physical and social sciences, we believe this contribution to be of general use and have therefore made our python implementation available at https://github.com/pmelchior/pyGMMis.

Appendix A. Missingness for density estimation

To understand what form the likelihood assumes when some samples are not observed, we need to determine how observed and missing samples are related. For this purpose Rubin (1976)\(^4\) states the following.

\(^4\)For the estimate of the degrees of freedom we assumed that no parameters are learned from the data, yielding an upper bound on the dof. If all parameters were linearly independent, which is not an appropriate assumption for GMMs, the dof would be reduced by 3%.
introduced the missingness mechanism $R$ that determines which part of the complete data $Y$ is observed ($Y_o$) or missing ($Y_m$). The relationship between $R$ and $Y_o$ or $Y_m$ gives rise to three distinct types of missingness:

$$ p(R) $$ missing completely at random (MCAR)

$$ p(R|Y_o) $$ missing at random (MAR)

$$ p(R|Y_o, Y_m) $$ missing not at random (MNAR)

Under MCAR, the missingness mechanism does not depend on the data at all; in turn, the data do not reveal properties of $R$. The key distinction between MAR and MNAR is whether $R$ depends only on observed or also on missing features.

For the problem of density estimation with incomplete samples $X$, it is not immediately obvious what $Y_o$ and $Y_m$ correspond to. We take guidance from the close relation between density estimation and function approximation by positive linear operators (e.g., Ciesielski, 1991). We therefore take $X$ to be the entire feature space $\mathbb{R}^d$ and $Y = \mathbb{R}$ the co-domain of a scalar function $p : X \rightarrow Y$, namely the PDF $p(x)$. The data then span $(X, Y) = \mathbb{R}^{d+1}$, and $R$ determines at what locations $x$ a value $y \in Y$ is recorded.

In Section 2.3, we introduced the completeness function $\Omega()$, which is equivalent to $p(R|\cdot)$. Only a spatially uniform completeness function fulfills the MCAR condition, but is irrelevant for density estimation as the resulting probability density function is simply a renormalized version of the true PDF. If $\Omega$ only depends on $x$, the MAR condition applies because $X$ is by construction completely observed (only values of $Y$ may be missing). Technically, MAR still holds if $\Omega$ depends on $x$ and the density $p(x')$ at some other observed position $x' \neq x$, for instance when the observed samples “shadow” some others. For the sake of brevity, we have not listed this case in Section 2.3. Only if samples are missing because of their own value of $y = p(x)$, or because of their relation to other missing values $p(x')$, we have MNAR. This can happen e.g. when an experimental device is locally invalidated by saturation or if samples are only recorded if their local abundance exceeds a certain threshold. An investigation under which conditions the proposed approach can account for MNAR cases is beyond the scope of this paper.

References

Biernacki C., Celeux G., Govaert G., 2003, Comput. Stat. Data Anal., 41, 561
Blommer J., Bujna K., 2013, preprint, (arXiv:1312.5946)
Bovy J., Hogg D. W., Roweis S. T., 2011, Annals of Applied Statistics, 5, 1657
Breiman L., 1996, Machine Learning, 24, 49
Cadez I. V., Smyth P., McLachlan G. J., McLaren C. E., 2002, Mach. Learn., 47, 7
Ciesielski Z., 1991, Probability And Mathematic Statistics, 12, 1
Dempoeter A. P., Laird N. M., Rubin D. B., 1977, Journal of the Royal Statistical Society. Series B (Methodological), 39, 1
Diebolt J., Ip E., 1996, in W.R. Gilks S. Richardson D. S., ed., , Markov Chain Monte Carlo in Practice. Chapman & Hall, London
Frühwirth-Schnatter S., 2006, Finite Mixture and Markov Switching Models. Springer Series in Statistics, Springer New York, http://www.springer.com/us/book/9783540729093
Goulding A. D., et al., 2016, Apl, 826, 167
Lee G., Scott C., 2012, Comput. Stat. Data Anal., 56, 2816
Leisteb M., et al., 2016, ApJS, 226, 24
Manjunath B. G., Wilhelm S., 2012, preprint, (arXiv:1206.5387)
McLachlan G. J., Jones P. N., 1988, Biometrics, 44, 571
McLachlan G., Peel D., 2000, Finite Mixture Models. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, https://books.google.com/books?id=YXgf1vEACAAJ
Mengersen K. L., Robert C., Titterington M., 2011, Mixtures: estimation and applications. Vol. 896, John Wiley & Sons
Naim I., Gildea D., 2012, preprint, (arXiv:1206.6427)
Nielsen S. F., 2000, Bernoulli, 6, 457
Rubin D. B., 1976, Biometrika, 63, 581
Rubin D., 1987, Multiple Imputation for Nonresponse in Surveys, Wiley Series in Probability and Statistics. Wiley Series in Probability and Statistics, Wiley, https://books.google.com/books?id=0Kr4AAAMAAJ
Schafner J. L., Graham J. W., 2002, Psychological Methods, 7, 147
Smyth P., Wolpert D., 1999, Machine Learning, 36, 59
Titterington D., Smith A., Makov U., 1985, Statistical analysis of finite mixture distributions. Wiley series in probability and mathematical statistics: Applied probability and statistics, Wiley, https://books.google.com/books?id=b2QQAQAAIAAJ
Ueda N., Nakano R., Graharamazi Z., Hinton G. E., 2000, Journal of VLSI signal processing systems for signal, image and video technology, 26, 133
Wolpert D. H., 1992, Neural Netw., 5, 241
Wolynetz M. S., 1979, Journal of the Royal Statistical Society. Series C (Applied Statistics), 28, 195
Wu C. F. J., 1983, The Annals of Statistics, 11, 95
Zhang Z., Chen C., Sun J., Chan K. L., 2003, Pattern Recognition, 36, 1973