1620
doi:10.1017/S1431927621005973
Microsc. Microanal. 27 (Suppl 1), 2021

Bayesian Approaches to Finding the Needles in the Microscopy Haystack

Joseph Simpson, Donovan Leonard and Chad Parish

ORNL, United States

One major difficulty in modern microscopy is the identification and quantification of distinct feature clusters in micrographs, where the term feature refers to precipitates, grains, pores, oxides, etc. Examples of clustering include determining whether a feedstock batch is contaminated, and if so, how many distinct contaminants are present, or the number of distinct precipitates in a novel alloy. One method for automating the discovery and quantification of such clusters is a two-step Bayesian approach of first applying a Dirichlet process to estimate the likely number of clusters, followed by an iterative Bayesian Gaussian Mixture Model to estimate the mean, variance, and weight of each cluster. A process flowchart with four-dimensional example Energy Dispersive Spectroscopy (EDS) inputs is provided in Figure 1 for reference. The Dirichlet process is a non-parametric method for estimating the probability of observation Yk+1 belonging to a new cluster given Yk existing observations, "i" existing clusters, and cluster parameter vector $\alpha = [\alpha 1, \ldots, \alpha i]$ [1]. A common analogy for the Dirichlet process is known as "stick breaking" [2], [3]. Presume that observed continuous measurements correspond to several multivariate Gaussian distributions and that the sum of the distribution weights is equal to unity. A stick of length one can be used to represent the weight of all distributions; the stick may be broken into an infinite number of pieces where each piece represents a unique Gaussian distribution and the length of the piece equal to the relative weight of the distribution. The Dirichlet process, therefore, technically describes an infinite number of Gaussian distributions. Because it is neither feasible nor efficacious to simulate an infinite number of distributions, it is common practice to truncate distributions with negligible contributions and instead provide distribution weight estimates over the range of (1, maximum plausible number of distributions) [4]. Selection of the maximum plausible number of distributions is subjective; however, there is a negligible computation penalty for oversizing the search range. For example, if an alloy family typically has 3-4 elementally unique phases, selecting a search range of (1, 10) would be reasonable. In the event that the Dirichlet process does not find a number of clusters that explains a significant percentage of observations, the most likely cause is either A) underestimating the maximum number of clusters or B) that cluster distributions are not Gaussian. Once the most likely numbers of distributions have been determined, users iteratively perform multivariate Gaussian regressions and incorporate subject matter expert opinion to determine the "true" number of distributions. Input data may be either raw elemental percentages or Principal Component Analysis' projections, though the latter adds a layer of obscurification. Alternative unsupervised machine learning methods can estimate the single most likely number of clusters and model a Gaussian mixture [5]; however, in the context of microscopy it is proposed to be advantageous to insert human supervision and review all likely numbers of clusters rather than outputting a single Gaussian mixture model. Non-elementally-Gaussian corrosion products or precipitates, low sampling rates for rare clusters, and data collection artifacts may result in less than ideal data. Therefore, the mathematically most likely number of clusters may not be accurate in an application-specific context. The advantages of such an approach are determining the number of clusters in a statistically informed manner rather than based on intuition, enabling users to shift time usage from manually identifying clusters to higher value activities such as analyzing correlations and investigating unexpected clusters, providing the covariance matrix for each cluster, and identifying which measurements have a high probability of being outliers. Furthermore, the approach is flexible by accepting any continuous variable, and may be implemented at either the pixel-level or feature-level. Implementation at the feature-level allows for incorporation of morphological data such as grain size, grain aspect ratio, and crystalline orientation from an Electron Backscatter Diffraction scan. The primary limitations of the proposed approach are the inability to incorporate categorical data such as crystalline structure and the assumption of clusters being

normally distributed. The approach is postulated to be particularly effective when using EDS measurements to determine the number of elementally distinct clusters in manufactured materials, as manufacturing variability and EDS measurement uncertainty can be reasonably modeled as Gaussian distributions and the calculated covariance matrices carry meaningful information on the intra-cluster composition variation. The authors present an application case study herein on a novel Oak Ridge National Laboratory developed FeNiCr alloy. Early results indicate that certain heat treatments may result in up to six elementally unique metallic and precipitate phases in the alloy, resulting in a rich exploratory dataset. A representative Backscatter Electron micrograph of the alloy is provided in Figure 2 for reference. The Bayesian machine learning approach outlined above is compared to both traditional multi-scale characterization and alternative machine learning methods including Principal Components Analysis with respect to determining the number of feature clusters.

https://doi.org/10.1017/S1431927621005973 Published online by Cambridge University Press

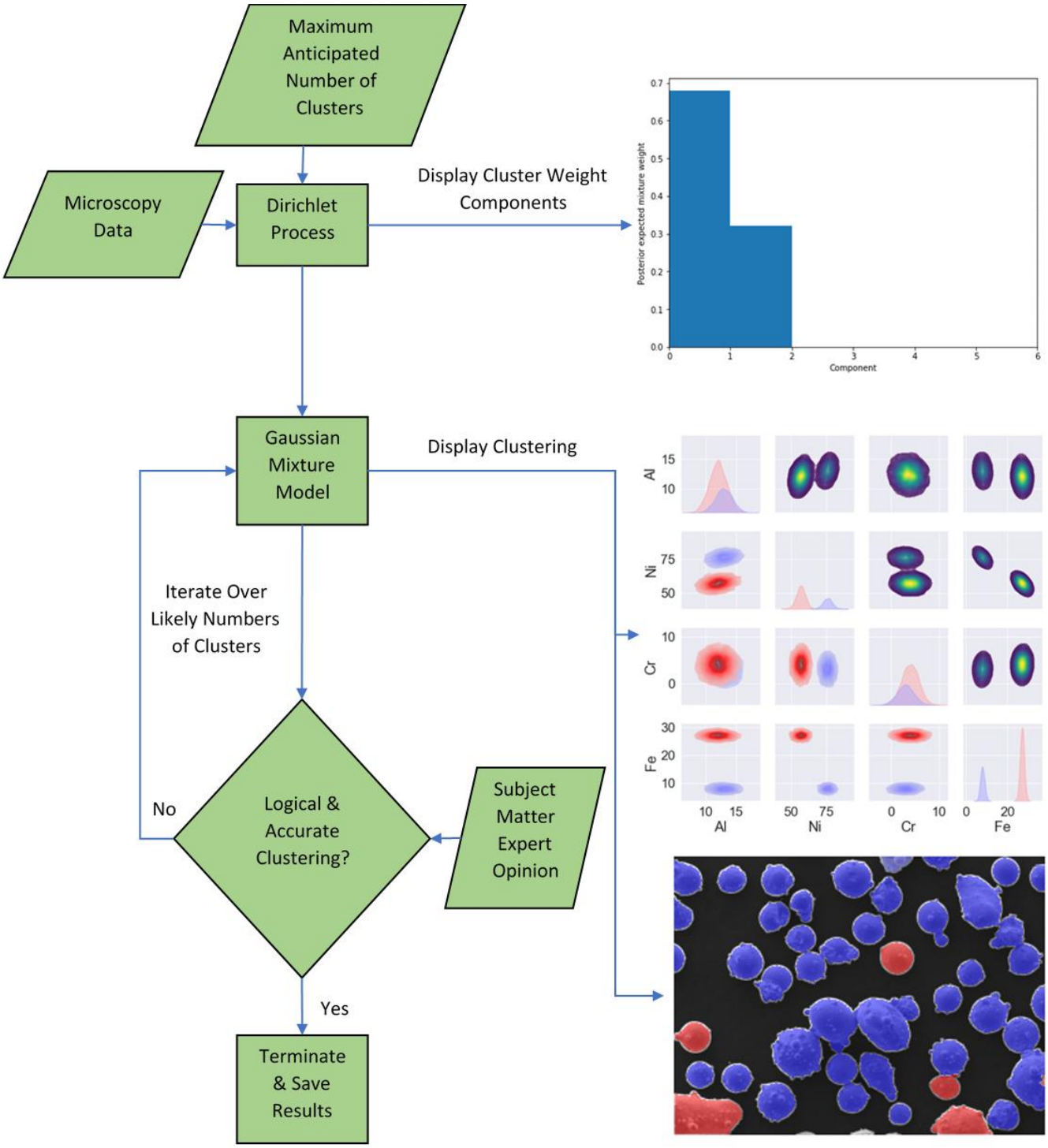

Figure 1. Process Flowchart for Two-Step Dirichlet-Gaussian Mixture Clustering Approach

https://doi.org/10.1017/S1431927621005973 Published online by Cambridge University Press

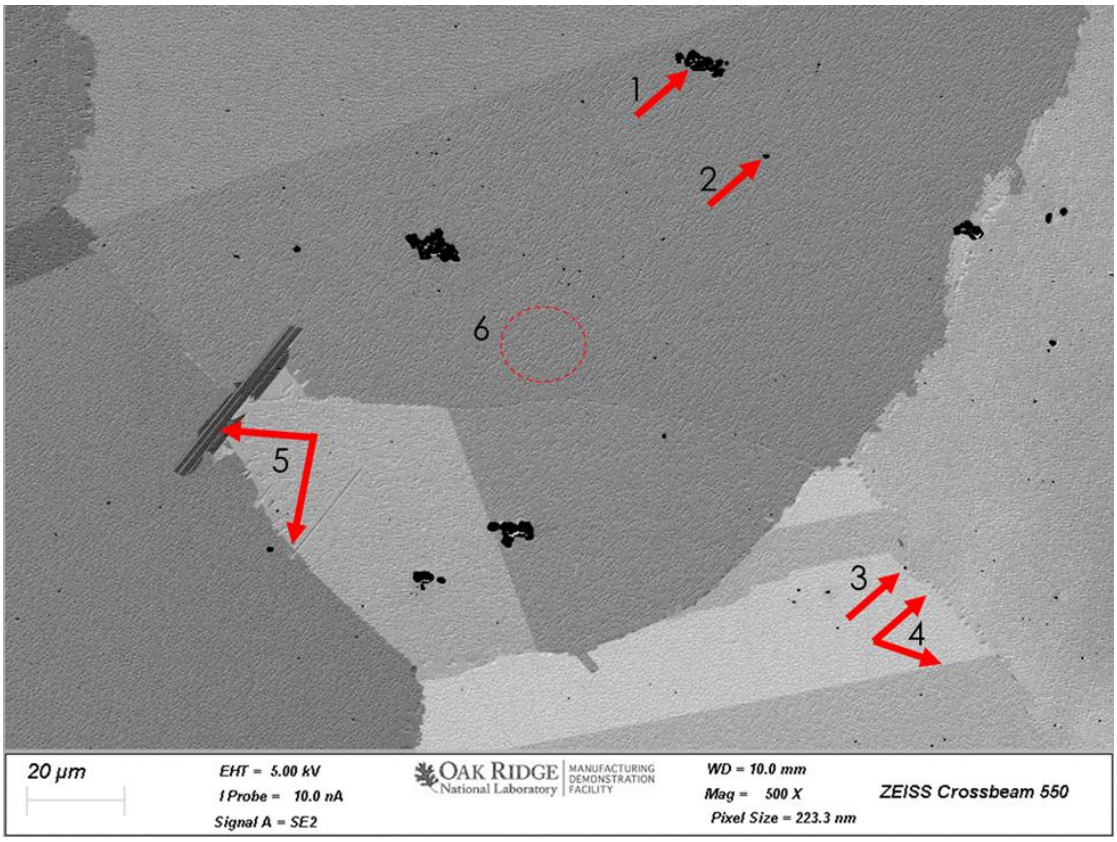

Figure 2. Sample Backscatter Diffraction Micrograph with Potential Unique Phases Indicated

References

[1] Y. W. Teh, "Dirichlet Process." 2010.
[2] A. Vlachos, Z. Ghahramani, and A. Korhonen, "Dirichlet Process Mixture Models for Verb Clustering," *Proc. ICML Work. Prior Knowl. Text Lang. Process. Helsinki, Finl.*, pp. 1–6, 2008, [Online]. Available: papers3://publication/uuid/612EDB87-A804-46C9-8254-BF5351273773.
[3] J. Sethuraman, "A CONSTRUCTIVE DEFINITION OF DIRICHLET PRIORS," *Inst. Stat. Sci.*, vol. 4, no. 2, pp. 639–650, 1994.
[4] K. Kurihara, M. Welling, and Y. W. Teh, "Collapsed variational dirichlet process mixture models," *IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 2796–2801, 2007.
[5] C. A. Bouman, "Cluster: An unsupervised algorithm for modeling Gaussian mixtures." 1997.

https://doi.org/10.1017/S1431927621005973 Published online by Cambridge University Press